COMPARATIVE METAGENOMIC
Strain-level microbial epidemiology and population genomics from shotgun metagenomics

Matthias Scholz, Doyle V Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L Morrow & Nicola Segata
WHAT ARE WE GOING TO TALK ABOUT

• Identifying microbial strains

• The challenge

• The solution - PanPhlAn
IDENTIFYING MICROBIAL STRAINS

- Pathogen discovery
- Epidemiology and population genomics
The Challenge

- Obtain robust sequence signatures and accurate profiles of microbial function across samples at strain-level resolution

- Currently achieved by sequencing the genomes of organisms of interest isolated from the environment and grown *in vitro*.

- Although important differences between microbes occur between strains, it has been difficult to achieve accurate metagenomic profiling beyond the species level.
PANPHLAN
PANGENOME-BASED PHYLOGENOMIC ANALYSIS

• A tool that uses metagenomic data to achieve strain-level microbial profiling resolution

• Produced the largest strain-level population genomic study of human-associated bacteria

• Profiled the transcriptional activity of strains in complex communities

• Significantly advance our ability to interrogate the activity of strains in vivo.
PANPHLAN

- Identifies which genes are present or absent within different strains of a species, based on the entire gene set of the species’ pangenome

- When available reference genomes are included in the analysis, PanPhlAn also permits the identification and comparison of both known and new strains

- Provides transcription rates for the identified genes, thus capturing *in vivo* transcriptional activity for the strain of interest
The outbreak was caused by an enteroaggregative (EAEC) strain O104:H4.

PanPhlAn gene family profiles binned the outbreak positive samples in a cluster of seven *E. coli* strains distinct from known *E. coli* subclades.

Detected a panel of outbreak-enriched gene families and pathways comprising virulence-related secretion systems and heavy metal tolerance.

PanPhlAn provides an accurate tool for identifying and characterizing outbreak strains in a metagenomic, cultivation-free setting.
**PANPHLAN**

**ESCHERICHIA COLI OUTBREAK IN GERMANY**

*Figure 1* | PanPhIAn validation and comparison with assembly. Plots report the $F_1$ score (a balanced measure of precision and recall) of detected gene families at different coverage levels of strains from three different species. Solid and dashed lines correspond to target strain presence or absence in the PanPhIAn reference database (ref. DB), respectively. The panels report the results of several species and validation settings (Online Methods): (a) Complex synthetic samples containing *E. coli* outbreak strain O104:H4 (ref. 19) merged with synthetic reads from 100 other microbes. (b) Synthetic reads from the *E. coli* outbreak strain merged with real human gut samples from three different subjects. (c) Three *E. coli* strains not present in the PanPhIAn reference database. (d) Synthetic reads from *S. aureus* strains merged with real skin samples. (e) *S. epidermidis* strains merged with samples from anterior nares. (f) Comparison with metagenomic assembly that has been aided by excluding simulated reads from non-target genomes.
Applied PanPhlAn to 1,316 gut metagenomes from publicly available large-scale investigations

PanPhlAn detected and profiled *E. coli* strains in 114 samples

The *E. coli* species diversity fell into six functionally distinct clades, but only four major groups represent *E. coli* phylogroups A, B\textsubscript{1}, B\textsubscript{2} and D and are the only groups present in the metagenomic samples we analyzed (the other two groups are composed of reference genomes only)

The German outbreak samples were located in phylogroup B\textsubscript{1}, in which they formed a subcluster
Network analysis suggested similar conclusions and identified a strain closely related to the German outbreak in a data set of Chinese individuals.

The Chinese sample appears to contain an O104:H4 strain encoding genes associated with the enterohemorrhagic plasmids, but it differs from the German outbreak strain in that it lacks the Shiga toxin-encoding region.

This observation was confirmed in an independent, ad hoc coverage analysis and assessed the impact on profiling accuracy in the case of multiple strains with similar relative abundance.
PANPHLAN

- Applied PanPhlAn to other species in an expanded set of 1,830 gut metagenomes from eight large-scale studies

- Focused on two prevalent gut species with uncharacterized population genomics, *Eubacterium rectale* and *Akkermansia muciniphila*
PanPhlAn distinguished *E. rectale* strains into three distinct clusters:
- Chinese cohorts clustered, European and North American cohorts clusters suggesting a geographically associated adaptation of strains from this species.

- The strain clustering, however, was not completely *exclusive*, as a few ‘European’ strains were present in the Chinese-dominated cluster, and some ‘Chinese’ strains were present in the other two clusters, suggesting potentially intriguing epidemiological events.
PanPhlAn resolved six discrete clades present in all populations.

These results could not be achieved by metagenomic assembly or by other available tools. Given the availability of only two reference genomes for this species.
PanPhIAn compared to other approaches, such as metagenomic assembly, allowed us to expand our population genomic study to other organisms and environments:

- Evidence of adaptation to geographically separated populations and strain-level temporal stability
- Able to profile several organisms from skin and oral samples (including cutaneous staphylococci and *Neisseria meningitides*)
- Profiled environmental communities by running PanPhIAn on >1,200 marine metagenomes

Overall, these analyses demonstrated that PanPhIAn is a convenient and powerful tool for population genomics, even for species where few reference genomes exist.