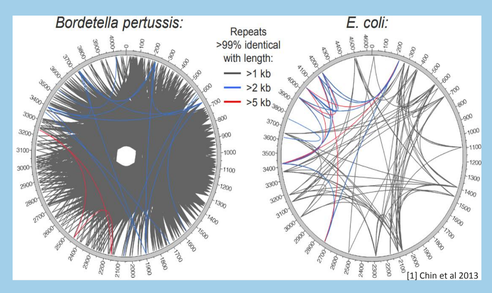
# Resolving the complex *B. pertussis* genome with barcoded nanopore sequencing

Natalie Ring, Jonathan Abrahams, Andrew Preston & Stefan Bagby
Department of Biology and Biochemistry, University of Bath

## Why study *Bordetella pertussis?*

- *Bordetella pertussis* causes whooping cough
- Whooping cough has **resurged** in recent years, despite **no decrease** in vaccine uptake and **few genotypic changes** [2]




- The *B. pertussis* genome is **complex**, with **high GC content** and many **long repetitive Insertion Sequence elements**
- Short-read sequencing has been **unable to resolve** the genome
- Nanopore sequencing may **produce single-contig *B. pertussis* assemblies** using reads longer than the repetitive sections
- **Structural resolution** may reveal previously **uncharacterised genomic differences** and **explain phenotypic changes** [3]

## What were our aims?

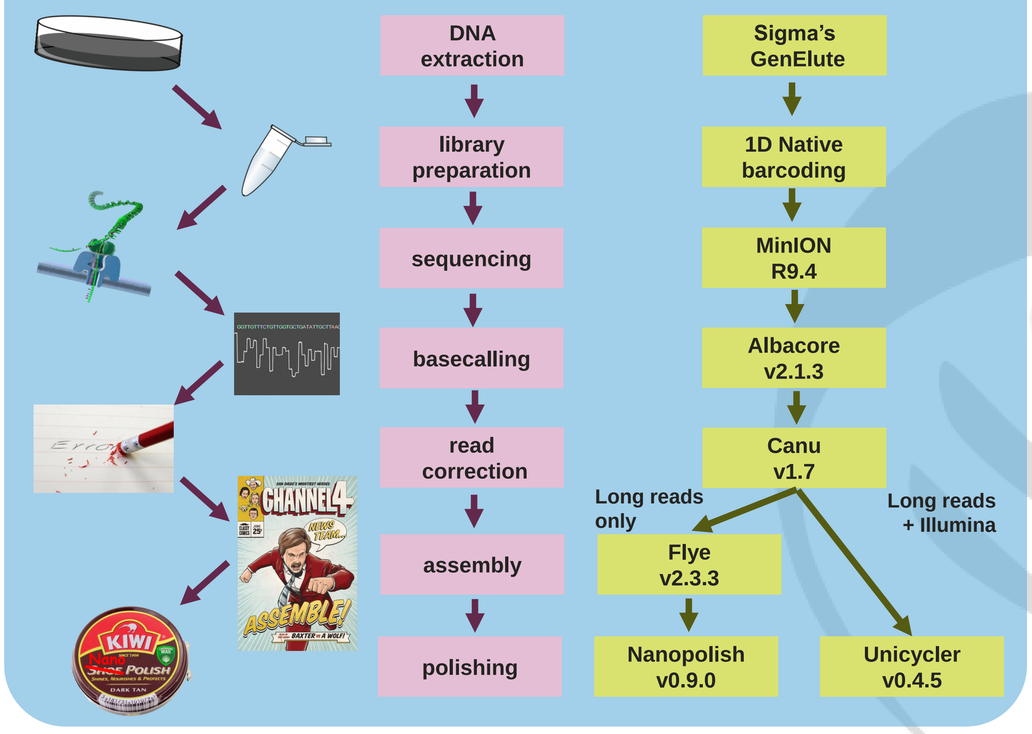
- **Determine** the optimal sequencing pipeline for *B. pertussis*
- **Visualise** genome differences between *B. pertussis* strains

## How did we choose our pipeline?

We tested exhaustive combinations of:

- Two successive **sequencing chemistries** (R7 & R9)
- Five **library preparation** kits
- Two **basecalling** tools
- Seven **genome assembly** tools
  - **Hybrid** vs **long-read-only** assembly
    - Three **assembly polishing** tools

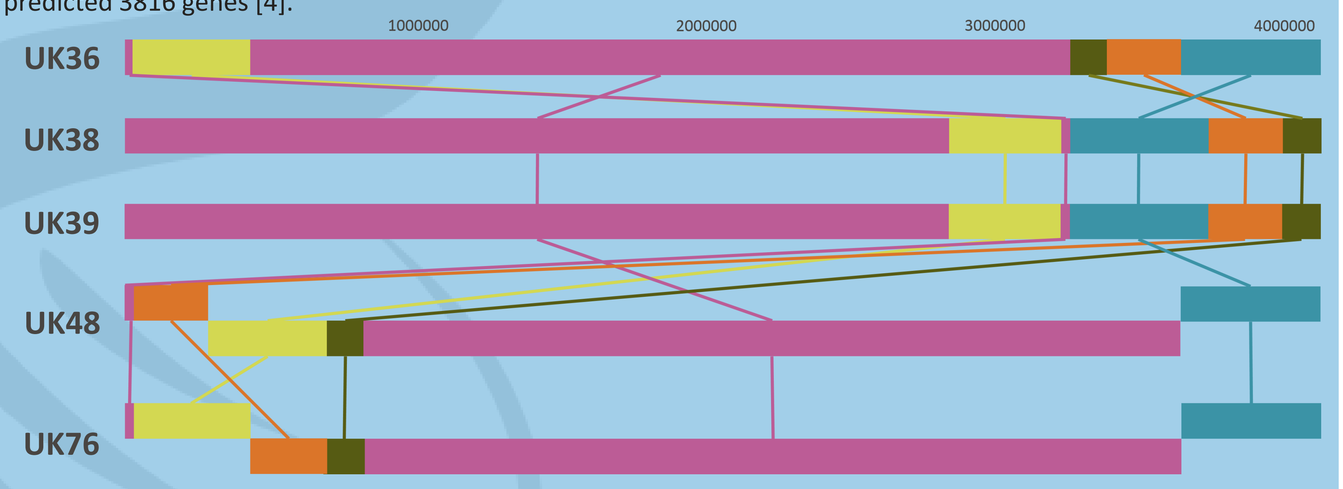
## Our pipeline!




## What did we find?

| Strain | # contigs | Size Mb | GC content % | Identity % | # genes predicted |
|---|---|---|---|---|---|
| UK36 | 1 | 4.108 | 67.7 | 99.57 | 3980 |
| UK38 | 1 | 4.108 | 67.71 | 99.33 | 3974 |
| UK39 | 1 | 4.108 | 67.71 | 99.39 | 3974 |
| UK48 | 2 | 4.112 | 67.70 | 99.32 | 3977 |
| UK76 | 1 | 4.113 | 67.69 | 98.93 | 3980 |

**Assembly of five strains using our barcoded hybrid pipeline consistently produces resolved genomes.** Five UK *B. pertussis* strains were sequenced in a single barcoded MinION run (mean read length 5kb), followed hybrid assembly with Unicycler (using Illumina short reads from the NCBI's SRA). % identity was estimated by comparing our assemblies to an Illumina-only assembly for each, produced by ABySS v2.0.3, and number of genes was predicted using annotation with Prokka v1.13. The original annotation of the reference genome, Tohama I, predicted 3816 genes [4].



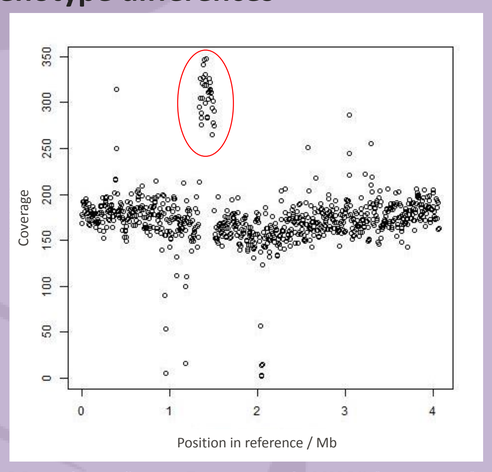

**Alignment with progessiveMauve reveals extensive inter-strain genome rearrangement.** Using the same barcoded data as above, we also assembled the five strains using our nanopore-only pipeline. The resulting assemblies were rearranged manually to set the first gene in the *B. pertussis* genome, gidA, at the start of each assembly, and one (UK48) was reverse-complemented using a homemade script. The final assemblies were aligned using progressiveMauve, which revealed at least minor differences in the arrangements of all strains except UK38 and UK39.

## What can we conclude?

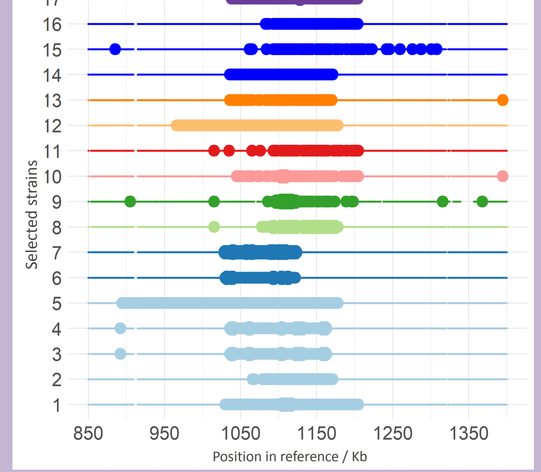
- **Nanopore-only assembly strategies produce *B. pertussis* genomes with 99.20% accuracy on average, whilst hybrid strategies yielded average 99.51% accuracy** (both compared to an Illumina short-read-only assembly for the same strains)
- **Barcoded nanopore sequencing enabled the assembly of single-contig *B. pertussis* genomes for at least five strains per flow cell**, allowing visualisation of **inter-strain genomic rearrangement**

## What's next?

- Initial comparison of resolved genomes has revealed **extensive rearrangements between strains**. We will **investigate** whether these rearrangements correlate with **phenotypic differences**
- Mapping of raw reads to the Tohama I *B. pertussis* reference genome indicates sections of enriched coverage in some strains, which could correspond to **large duplication events** which are **not resolved by the current sequencing pipeline**
- We will trial an **ultra-long read sequencing strategy** to resolve these ultra-long duplications, followed by **investigation of correlation between large duplications and phenotype differences**




**Mapping of raw *B. pertussis* UK48 long reads to a reference genome suggests a large duplication event in UK48**




**Mapping the short read data of all archived strains to the reference suggests a duplication is present at the same locus in multiple strains**

Scan the QR code to view full methodology, results and data repository

## References

[1] Chin et al 2013. Automated, Non-Hybrid De Novo Genome Assemblies and Epigenomes of Bacterial Pathogens. https://bit.ly/2JGkAld
[2] Bart, M. et al. (2010). Comparative genomics of prevaccination and modern *Bordetella pertussis* strains. *BMC Genomics*, 11, p627
[3] Belcher, T. & Preston, A. (2015). *Bordetella pertussis* evolution in the (functional) genomics era. *FEMS Pathogens and Disease*. 73(8)
[4] Parkhill *et al* 2003. Comparative analysis of the genome sequences of *Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica*. Nature Genetics, 35, pp32-40

## Tools

**ABySS**: https://github.com/bcgsc/abyss
**Albacore**: https://community.nanoporetech.com/downloads
**Canu**: https://github.com/marbl/canu
**Flye**: https://github.com/fenderglass/Flye
**Nanopolish**: https://github.com/jts/nanopolish
**Prokka:** https://github.com/tseemann/prokka
**Unicycler**: https://github.com/rrwick/Unicycler

## Acknowledgements

Thanks to Oxford Nanopore Technologies for part-funding my PhD, and providing lab-space for the R9 sequencing.

Additional thanks to the Nanopore Group at UC Santa Cruz for their help and lab-space for the R7 sequencing, and for their ongoing advice.

## About the author

I am a 2nd year PhD student at the University of Bath, researching microbial genomics with an emphasis on sequencing and bioinformatics.
I previously worked for 4 years in the Data Coordination Centre for the International Mouse Phenotyping Consortium at MRC Harwell
n.a.ring@bath.ac.uk
@NatalieAnneRing