Ultra-long reads and ultra-long duplications:
What nanopore sequencing is revealing about *Bordetella pertussis*

Natalie Ring[1], Jonathan Abrahams[1], Joshua Quick[2], Nick Loman[2], Andrew Preston[1] & Stefan Bagby[1]

[1]Department of Biology and Biochemistry, University of Bath, UK
[2]School of Biosciences, University of Birmingham, UK

Background

The *B. pertussis* genome is **repetitive**. The average *B. pertussis* genome contains **280 copies of >1,000 bp insertion sequence (IS) elements**, representing **~7% of the 4.1 Mb total genome** length. The many IS copies mean that **closed genome assemblies cannot be produced using short-read sequencing** (e.g. Illumina), because each IS element is longer than the short reads

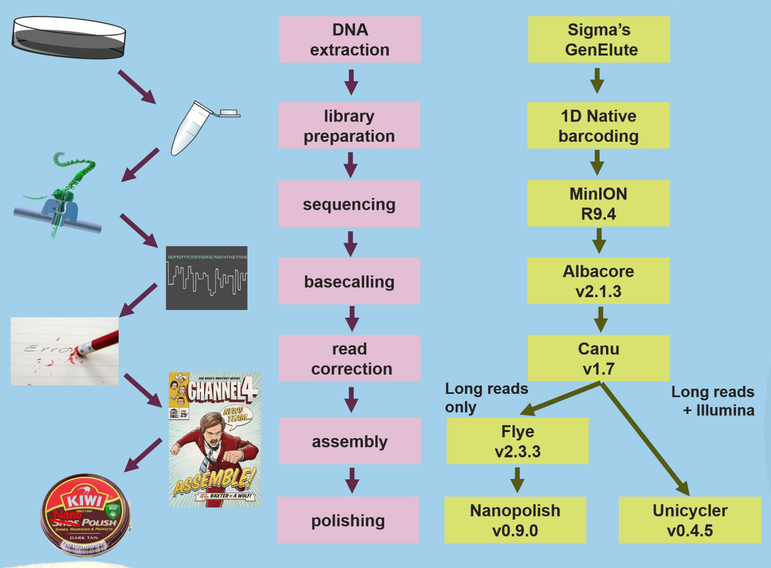
B. pertussis is traditionally described as a **monomorphic** species: very **few base-level differences** exist between different strains. The presence of so many mobile IS elements in the genome, however, means that genome-level differences, such as **rearrangements, deletions and duplications**, are possible

We are using **long-read sequencing** to identity **genome-level differences** between otherwise highly similar *B. pertussis* strains

Our nanopore sequencing pipeline*

Through extensive testing and optimisation, we defined a **sequencing and data processing pipeline**, using Oxford Nanopore Technologies' MinION sequencing, to produce **reads longer than 1,000 bp** which can be used to **assemble closed *B. pertussis* genomes**

Using barcodes, up to **12 genomes** can be sequenced **per flow cell**



N.B. new basecalling tools, which are more accurate than Albacore, now exist. A new pore, R10, was also released in March 2019

*For more information, see Ring et al. 2018 [1]

Our long read genomes

Strain	Contigs	Size Mb	IS *481* copies
UK36	1	4.108	258
UK38	1	4.108	258
UK39	1	4.108	258
UK48	2	4.112	262
UK76	1	4.113	262

We used our nanopore sequencing pipeline to **sequence five *B. pertussis* strains**, isolated during the 2012 whooping cough outbreak in the UK [1,2]. The genomes of **all but one** of these strains could be assembled into a **closed contig** using this pipeline, which produced reads with a **mean length >6,000 bp**

Two genomes, **UK48 and UK76**, were **longer than the others**, and also **had more copies of IS *481***. On closer inspection, the genomes of these two strains appear to contain **the same ultra-long duplicated region**, ~1.3 Mb into the reference genome

We then used an **ultra-long gDNA extraction** method [3] to produce **100x coverage** of the UK48 genome in **reads longer than 100,000 bp**, and a **maximum read length of 645,000 bp**. However, we produced only 30x coverage in reads longer than the duplicated region (~180,000 bp), which was **insufficient to produce a closed genome sequence**

Our "ultra-long" UK48 reads:

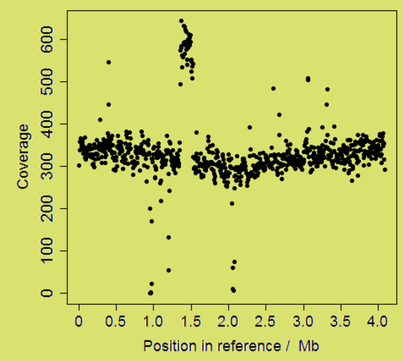
Read length / bp	Number of reads
1-100,000	35,968
100,001-200,000	2300
200,001-300,000	277
300,001-400,000	45
400,001-500,000	8
500,000+	2

Some *B. pertussis* genomes contain long duplicated regions

Even using nanopore sequencing reads longer than IS *481*, some strains' genomes **cannot be assembled into closed contigs**

Mapping the raw reads from these strains to the genome of the reference strain, Tohama I, reveals **regions of enriched coverage**

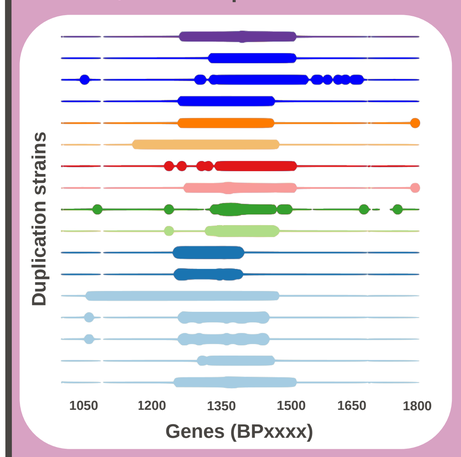
In UK48 (below), we see a region of ~0.2 Mb, with almost exactly **twice as much coverage** as the rest of the genome. This suggests that this region of the genome is **present twice** in UK48



The "motility" duplication

Over **15 strains** sequenced globally appear to have **a duplication of the same region** which is duplicated in UK48 and UK76 [4]

Although the exact length and genes included in the duplicated region varies slightly from strain to strain, the duplication tends to be **centred around the same set of "core" genes**: BP1334 to BP1492

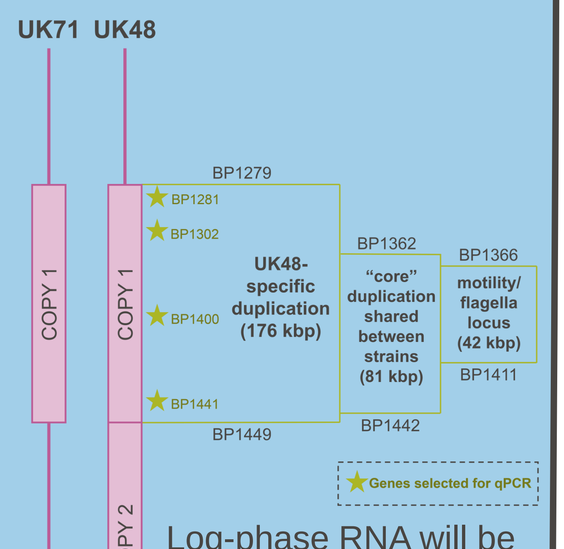


42 of the 77 genes in the "core" duplication are related to **motility and/or flagella synthesis**

We are now aiming to characterise this duplication further

For more about copy number variation in *Bordetella pertussis*, visit the poster of Jonathan Abrahams!

Does the duplication affect gene expression?



We are using **qPCR** to compare **gene expression** in **UK48** with gene expression in **UK71**, a strain which is highly similar to UK48 but whose genome **does not contain the duplication**

Log-phase RNA will be extracted from both strains in **Bvg(+) and Bvg(-)** conditions, and expression of **four genes** from the duplicated region will be measured

Future questions: Does the duplication affect **growth rate**?
Does the duplication affect **motility**?

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

References

[1] Ring et al. (2018). Resolving the complex *Bordetella pertussis* genome using barcoded nanopore sequencing. Microbial Genomics **4**(11)
[2] Sealey et al. (2015). Genomic analysis of isolates from the United Kingdom 2012 pertussis outbreak reveals that vaccine antigen genes are unusually fast evolving. Journal of Infectious Diseases **212**(2)
[3] Quick (2018). Ultra-long read sequencing protocol for Rad004. https://www.protocols.io/view/ultra-long-read-sequencing-protocol-for-rad004-mrxc57n
[4] Abrahams et al. (In preparation). Duplications drive diversity in the monomorphic pathogen *Bordetella pertussis* on an underestimated scale.

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 3rd year PhD student at the University of Bath, researching *Bordetella pertussis* genomics with an emphasis on sequencing and bioinformatics.
I previously worked for 4 years as a bioinformatician at MRC Harwell, and have a PGDip in Science Communication

n.a.ring@bath.ac.uk
@NatalieAnneRing