Institution: Cardiff University

Unit of Assessment: Biological Sciences (5)

Title of case study: Improved detection, monitoring and treatment of HIV, Influenza and SARS-CoV-2 through genomics and bioinformatics

Period when the underpinning research was undertaken: 2013 – 2019

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Thomas Connor	Professor	01/11/2012 – present
Matthew Bull	Research Associate	18/08/2014 – 31/02/2016
Anna Price	Research Software Engineer	13/11/2017 – present

Period when the claimed impact occurred: 2017 – 2020

Is this case study continued from a case study submitted in 2014? No

1. Summary of the impact (indicative maximum 100 words)

Next-generation sequencing (NGS) facilitates rapid genome analysis of pathogens, but its use in public health surveillance and routine clinical practice has been limited by a lack of robust and customisable data analysis pipelines. Building on research establishing a new UK-wide Cloud Infrastructure for Microbial Bioinformatics (CLIMB), Cardiff researchers implemented a bespoke £5M NGS bioinformatics infrastructure for Public Health Wales' new Pathogen Genomics Unit. This new clinical service resulted in: (i) more efficient and bespoke treatment for all HIV patients in Wales; (ii) improved, rapid surveillance of influenza and SARS-CoV-2 infection informing global vaccine development; and (iii) provision of critical data on the spread of SARS-CoV-2 informing Welsh and UK Government pandemic responses.

2. Underpinning research (indicative maximum 500 words)

Next-generation sequencing (NGS) enables rapid, cost-effective sequencing of complete genomes used widely in research. A lack of robust, reproducible, and customisable data analysis pipelines, however, prevented large-scale expansion of NGS for routine clinical practice. The MRC-funded Cloud Infrastructure for Microbial Bioinformatics (CLIMB) **[G3.1]**, a collaboration between Cardiff, Warwick, Birmingham, Swansea, Bath and Leicester Universities and the Quadram Institute, is a high-performance computing facility customised for microbial data banking and analysis, and collaborative sharing of NGS genomic data. Cardiff researcher Connor led the design and development of the CLIMB IT infrastructure using his expertise in biocomputing for bioinformatics **[3.1]**. CLIMB is recognised by the Welsh Government's Genomics for Precision Medicine Strategy as a key area of national research excellence. Over the REF period, Cardiff researchers used CLIMB to integrate NGS and high-end biocomputing analysis tools for investigation of pathogen evolution and transmission in local and global outbreaks. These studies included:

2.1 Reliable mapping of viral genomes

The Cardiff team, with Public Health Wales (PHW), benchmarked existing software programmes for mapping viral genomes, and found that some failed to map zoonotic (animal origin) viruses, due to use of single reference sequences. During sequencing, 'reads' (short fragments of DNA generated from microbial samples) are mapped against reference genomes but viral RNA-based genomes (including influenza, HIV, SARS-CoV-2) are small and can be diverse. These issues cause key sequence material to be discarded (where it does not match reference genomes), creating a significant risk of incorrect results (e.g., when testing whether two cases form part of a transmission chain). To address this issue, the Cardiff team designed a novel tool (VAPOR) allowing selection of multiple reference sequences; this improved the proportion of mapped reads by 13%. VAPOR was also able to classify 6.6 million reads in a mean time of 3.7 minutes, a significant improvement from standard approaches (e.g., BLAST, which takes over 20 hours to classify just 2 million reads) **[3.2]**.

2.2 Identifying human DNA in microbial datasets

Microbial infection samples are often contaminated with human genetic material. Using CLIMB, the Cardiff team evaluated the effectiveness of bioinformatic approaches for cleaning

up viral and bacterial genomic datasets. They identified that a novel combination of two sequence filtering approaches (Bowtie, followed by SNAP **[3.3]**) was highly effective in cleaning up shorter (150 base pair) bacterial infection sequence datasets. This approach was validated by re-examining over 11,000 published bacterial datasets, revealing that 6% of the datasets were contaminated by unidentified human DNA. Implementation of Cardiff's sequence clean-up approaches significantly reduced the risk of false results or unusable samples **[3.3],** as vital prerequisites for clinical accreditation of a diagnostic service.

2.3 Tracking the spread of SARS-CoV-2

The bioinformatics approaches developed by the Cardiff team, and embedded in the MRC CLIMB platform, meant that, upon the emergence of the novel coronavirus pandemic, critical infrastructure to enable a UK-wide genomics response was readily available. CLIMB supported rapid establishment of the COVID-19 Genomics UK Consortium (COG-UK), which includes the Cardiff team. COG-UK used the CLIMB platform to track the geographical and temporal spread of SARS-CoV-2 in the UK and Europe **[3.4]**, and to identify new virus variants. In October 2020, the MRC awarded £600K **[G3.2]** to the Cardiff part of the consortium to provide these capabilities internationally **[3.4]**.

As part of COG-UK, the Cardiff team led on genomic sequence analysis of the virus's spike protein and a mutation within it (D614G). The D614G mutation is one of multiple spike protein mutations in the highly transmissible SARS-CoV-2 B117 strain. Via the world-leading sequencing of 40,000 SARS-CoV-2 genomes in under 6 months, the Cardiff team and COG-UK collaborators found that the D614G variant did not cause increased disease severity, but was linked to increased transmissibility and higher viral loads in younger people **[3.5, G3.3]**.

3. References to the research (indicative maximum of six references)

[3.1] Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, et al. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microb Genom.* 2016;2(9). https://www.climb.ac.uk/
[3.2] Southgate JA, **Bull MJ**, Brown CM, Watkins J, Corden S, Southgate B, Moore C, **Connor TR.** Influenza classification from short reads with VAPOR facilitates robust mapping pipelines and zoonotic strain detection for routine surveillance applications. *Bioinformatics* 2020;36(6):1681-1688. doi:10.1093/bioinformatics/btz814
[3.3] Bush SJ, **Connor TR,** Peto TEA, Crook DW, Walker AS. Evaluation of methods for detecting human reads in microbial sequencing datasets. *Microb Genom.* 2020;10.1099/mgen.0.000393. doi:10.1099/mgen.0.000393
[3.4] Alm E, Broberg EK, **Connor TR**, Hodcroft EB, Komissarov AB, Maurer-Stroh S, sMelidou A, Neher RA, O'Toole Á, Pereyaslov D. The WHO European Region sequencing laboratories and GISAID EpiCoV group. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Euro Surveill.* 2020;25(32):pii=2001410. doi:10.2807/1560-7917.ES.2020.25.32.2001410
[3.5] Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, Southgate J, Johnson R, Jackson B, Nascimento FF, Rey SM, Nicholls SM, Colquhoun RM, da Silva Filipe A, Shepherd J, Pascall DJ, Shah R, Jesudaso, N, Li K, Jarrett R, **Connor, TR.** Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell* 2021:184(1), 64–75.e11. doi:10.1016/j.cell.2020.11.020 Published online November 18th 2020

Selected grants:
[G3.1] Prof T Connor, £8.4M Medical Research Council (MRC) Consortium for Medical Microbial Bioinformatics (CLIMB); 01/04/2014-31/8/2020. Cardiff award value £1,156,410
[G3.2] Prof T Connor, CLIMB Global Pathogen genomics analysis platform, MRC World Class Laboratory Award; 23/10/2020-31/03/2021 £600,000
[G3.3] Prof T Connor, COVID-19 Genomics UK (COG-UK) Consortium Severe Acute Respiratory Syndrome virus (SARS-COV-2), MRC. 01/04/2020-31/3/2022 £980,200

4. Details of the impact (indicative maximum 750 words)

Via the MRC-funded CLIMB infrastructure, Cardiff researchers developed key methodological innovations delivering sensitive and rapid analysis of large-scale pathogen datasets. Working

with Public Health Wales (PHW), the research led to the establishment of a bespoke NGS bioinformatics infrastructure for PHW's Pathogen Genomics Unit (total investment over £5M), a key component in delivering the Welsh Government's Genomics in Precision Medicine Strategy. This enhanced clinical and policy outcomes through, for example, improvements to HIV and influenza diagnostic services in Wales, and significant contributions to global surveillance of influenza and SARS-CoV-2.

4.1 A new NGS bioinformatics service for PHW Pathogen Genomics Unit

In response to the Welsh Government's Genomics for Precision Medicine Strategy, PHW recognised the need to provide a more advanced microbiology genomics service. The Head of PHW's Pathogen Genomics Unit, Dr Sally Corden, identified Connor as a key expert to provide the required bioinformatics expertise to implement NGS as a routine diagnostic and surveillance tool. Corden noted that Connor was *"somebody who understood how to write the code to translate information from the sequencer and put it back together into something that's a readable format for a clinician"* **[5.1]**. Corden initiated Connor's secondment to PHW from Cardiff University in 2017 (0.2 FTE, rising to 0.8 FTE in 2019-20). Dr Quentin Sandifer, Director of Public Health Services at PHW confirmed that Connor's appointment resulted in "*key impacts on our service, including the development of the Pathogen Genomics Unit"* **[5.2]**.

Connor used his research expertise to benchmark and support the development of laboratory processes for a new sample processing pipeline, while also designing a new computational infrastructure to meet requirements for large-scale data analysis, reproducibility and archiving, based on elements of the CLIMB infrastructure **[3.1, 3.2]**. Connor also built an end-to-end bioinformatics analysis and reporting platform to meet clinician requirements around diagnoses. Sandifer explained that PHW "*benefited enormously from Dr Connor's expertise in the area of bioinformatics and pathogen genomics which has underpinned the development of the bioinformatics pipelines that produce the reports that are then communicated to clinicians and other healthcare staff*" **[5.2]**.

PHW's novel microbiology genomics system resulted in new accredited services with automated reporting of pathogen sequence data directly to international repositories, including CLIMB and GISAID (influenza). Based on the success of Connor's work, PHW invested in four specialist bioinformaticians as part of a new PHW Pathogen Genomics Unit bioinformatics team (led by Connor) **[5.2]**. Connor also assisted the Unit in writing a successful multi-million pound business case to the Welsh Government for further funding to expand the NGS clinical service; this was awarded in 2019-2022 **[5.2]**. The Chief Scientific Advisor for Health in Wales, Dr Rob Orford, confirmed the link between Connor's research and the investment made in the Unit: *"The confidence the Welsh Government has in Dr Connor's work is reflected in the overall level of investment we have made/committed to the Pathogen Genomics work within Public Health Wales – of which he and his team are a core and leading part – which stands at over £5M to date"* **[5.3]**.

Orford explained the significance of Cardiff's research for the delivery of the Welsh Government's Genomics in Precision Medicine Strategy, noting that it has "*improved the ability of NHS Wales to diagnose, monitor and treat patients across a range of conditions"* and that the systems created by Connor *"provide situational updates to Welsh Government to support policy making"* **[5.3]**.

4.2 Advancing clinical services for HIV, influenza and SARS-CoV-2

Application of Cardiff's research **[3.2, 3.3],** alongside the new computational architecture and analysis pipelines Connor established at the PHW's Pathogen Genomic Unit **[3.1]**, enabled PHW's pathogen genomics service to deliver the following improvements to clinical services:

a. A new ISO accredited HIV service

New capabilities within the Pathogen Genomics Unit allowed PHW to launch a national HIV screening and monitoring service in July 2018. Prior to the service, patients' RNA samples were sent to Birmingham, creating a 6 week wait for results. Further, the central facility in Birmingham was not able to analyse samples with a low viral load **[5.2]**. PHW's Pathogen Genomic Unit's new HIV service, using Cardiff research pipelines and methods **[3.1, 3.2, 3.3]**,

conducts HIV sequencing in-house, reducing sample turnaround to 7 days, and is able to provide more detailed and sensitive sequencing information than was previously available **[5.2]**. A further benefit of the new service is enhanced engagement with clinicians, allowing emergencies to be handled quickly, in a way that was not previously possible. Corden cites the example of, *"an HIV positive lady who was pregnant, and was late booking into the service. As the service is local, we were able to work during a weekend to turn around the report for the clinician within six days, as the treatment of the unborn baby was critical"* **[5.1]**.

Cardiff's testing methods also enhanced clinical services for patients with low viral load: "*We are able to detect resistance mutations below the 20% standard: down to 10%, and this allows us to detect resistance to anti-viral drugs at an earlier point, and allow clinicians to make changes to a patient's treatment*" **[5.1]**. This innovation delivers important public health benefits: once patients are receiving appropriate, tailored treatment, they are no longer at risk of transmitting HIV. The new service implemented by Connor and colleagues is also considerably more cost-effective for the NHS: analysis of each sample now costs £150, compared to the previous £250 per sample **[5.2]**. A PHW Surveillance Report into HIV and STI trends in Wales (2018) estimated there were 1,585 patients being treated for HIV in Wales; this reduced cost therefore represents a saving of approximately £159K per year to NHS Wales. In July 2019, the service received formal accreditation (ISO 15189) **[5.4]**, making it the first ISO-accredited NGS-based HIV typing service in the UK **[5.2]**. Sandifer confirmed that the service "*…has been built thanks to Dr Connor's research and expertise, which have allowed us to offer integrase inhibitor sensitivity screening as standard, at improved turnaround time, and lower cost*" **[5.2]**.

b. Improved and rapid influenza surveillance

The whole genome sequencing approach of PHW's Pathogen Genomics Unit also enabled more advanced and rapid sequencing of seasonal influenza, both of which are vital for successful global vaccine development. In 2017-2018, the Wales Specialist Virology Centre was awarded £90K by the Welsh Government to pilot the use of whole genome sequencing for influenza surveillance, "*based on Dr Connor's pioneering next generation sequencing and analysis approach*" **[5.5]**. Dr Catherine Moore, Consultant Clinical Scientist, Public Health Wales and Wales National Virology Lead for Respiratory Virus Surveillance confirmed "*as a result of the success of the pilot, we implemented changes in our influenza surveillance methods to make use of whole genome sequencing in time for the 2018-19 influenza season*" **[5.5]**. Wales is now able to provide extra information by sequencing all eight segments of the influenza virus genome (compared to the two via Sanger sequencing, as had been used previously). This enhances monitoring for resistant mutations within segments PB1, PB2, and PA, which have the potential to affect new classes of influenza antivirals **[5.5]**.

The Unit's new bioinformatics analysis methods also enabled more effective geographic mapping of outbreak transmissions **[5.5]**. Moore explains: "*Our new system, developed in partnership with Dr Connor rapidly demonstrated its effectiveness, enabling us to sustainably sequence and share Influenza genome sequences with international surveillance databases..."* **[5.5]**.

As a result, Welsh influenza samples can be typed and submitted to international surveillance systems GISAID and ECDC within 7 days **[5.5]**. This improved analysis pipeline resulted in PHW sequencing 147.4 influenza genomes per million people in Wales, compared to only 44.6 and 2.2 influenza genome sequences per million people in England and Scotland, respectively (over the same period of time, the 2018-19 and 2019-20 influenza seasons) **[5.5]**. Moore confirmed that "*the sequences we submitted that have been generated using Dr Connor's methods will have been used to inform the design of the 2018, 2019, and 2020 influenza vaccine*", contributing to global efforts to protect those most vulnerable to influenza **[5.5]**.

c. Rapid sequencing of all Welsh SARS-CoV-2 samples

Sequencing of SARS-CoV-2 is vital for effective tracking of virus spread, identification of mutations, and design of successful vaccines. During the current coronavirus pandemic, the PHW Pathogen Genomics Unit is responsible for sequencing all Welsh SARS-CoV-2 samples,

feeding these data into the COVID-19 Genomics UK Consortium (COG-UK). This work was facilitated by Connor and his team establishing a new dedicated CLIMB-COVID platform, which directly enables COG-UK to analyse and integrate UK-wide SARS-CoV-2 data.

At the start of the first UK-wide lockdown (between March 22-31, 2020) PHW's Pathogen Genomic Unit had sequenced the highest number of SARS-CoV-2 genomes in the UK **[5.6a]**. This work ensured that the UK led the world in terms of genetic knowledge of coronavirus, having sequenced 806 samples by March 31, 2020 (compared to 744 in the USA and 296 in China) **[5.6a]**. By December 2020, PHW continued to lead global sequencing efforts (20,736 sequences), more than any other country in the world, except for England (130,325) **[5.6a]** and the USA (51,212) **[5.6b]**. [Text redacted]

d. Informing government decision-making to minimise SARS-CoV-2 transmission

Based on his significant pathogen and bioinformatics expertise, Connor is a member of Wales' Technical Advisory Cell (TAC), which provides scientific advice to the Welsh Government on coronavirus **[5.8a]**. In July 2020, TAC advised the Welsh Government to include genomics in its list of 'Early Warning Indicators' for SARS-CoV-2 in Wales **[5.8b]**. An October 2020 report by Connor, identifying the entry and westward spread of new SARS-CoV-2 strains across Wales **[5.9]**, was also used by the First Minister Mark Drakeford to justify travel restrictions into Wales from October 16, 2020 **[5.9]**. Orford confirmed that the research *"directly influenced Welsh government policy at the highest level of Government, demonstrating the spread of infection geographically and supporting decision making around restrictions such as the '5 mile rule' and travel from high-prevalence areas in order to protect the Welsh population"* **[5.3].**

[Text redacted]

5. Sources to corroborate the impact (indicative maximum of 10 references)

[5.1] Testimonial: Dr Sally Corden, Pathogen Genomics Unit, PHW
[5.2] Testimonial: Dr Quentin Sandifer, Director of Public Health Services, PHW
[5.3] Testimonial: Dr Rob Orford, Chief Scientific Advisor for Health in Wales
[5.4] UKAS ISO 15189 accreditation certificate and schedule
[5.5] Testimonial: Dr Catherine Moore, Consultant Clinical Scientist, PHW, Wales National Virology Lead for Respiratory Virus Surveillance
[5.6] a. Reports confirming numbers of cases sequenced: COG-UK April and December 2020 reports, **b.** *Washington Post* article with data showing share of coronavirus outbreak sequenced by countries with at least 100 reported cases, 23 December 2020
[5.7] [Text redacted]
[5.8] a. Welsh Government Technical Advisory Cell (TAC) membership, **b.** TAC Circuit Breakers / Early Warning Indicators report July 2020
[5.9] Mark Drakeford letter to Boris Johnson, 13 October 2020 referencing SARS-CoV-2 Genomic Insights report October 2020