

WELLCOME GENOME CAMPUS ADVANCED COURSES

Genomics and Clinical Microbiology

22 January – 27 January 2017

**Held at
Wellcome Genome Campus Advanced Courses
Laboratory
Wellcome Genome Campus
Cambridge, UK**

**(c) Wellcome Genome Campus Advanced Courses and Scientific Conferences
2017**

STAFF

Advanced Courses Manager **Dr Rebecca Twells** (*email: advancedcourses@wellcomegenomecampus.org*)
The Wellcome Trust, Hinxton, Cambridge

Course Instructors **Professor Martin Maiden** (e mail: martin.maiden@zoo.ox.ac.uk)
Department of Zoology
University of Oxford

Professor Stephen Gillespie (e mail: shg3@st-andrews.ac.uk)
University of St Andrews

Dr Cath Arnold (e mail: catherine.arnold@hpa.org.uk)
Public Health England

Dr Claire Jenkins (e mail: Claire.Jenkins1@phe.gov.uk)
Public Health England

Course Assistants **Dr Carina Brehony**
University Hospital Galway

Dr Tim Dallman
Public Health England

Dr Keith Jolley
University of Oxford

Dr Sandra Nicoletti
Public Health England

Dr Katarina Oravcova
University of St Andrews

Dr Wilber Sabiiti
University of St Andrews

Dr Ulf Schaefer
Public Health England

Dr Danny Sewell
Public Health England

Advanced Courses Team:	Yvonne Thornton	Advanced Courses Administrator
	Julie Ormond	Advanced Courses Laboratory Manager
	Darren Hughes	Advanced Courses Programme Officer
	Nicola Stevens	Advanced Courses Assistant Administrator
	Kate Waite	Advanced Courses Assistant Lab Manager
	Pamela Black	Advanced Courses Education Officer
	Martin Aslett	Advanced Courses IT Manager

URL: www.wellcomegenomecampus.org/coursesandconferences
email: advancedcourses@wellcomegenomecampus.org

ACKNOWLEDGEMENTS

We are very grateful to the following companies for their support in the loaning of equipment, gifts of consumables and/or technical support.

Company	Web site	UK Telephone
Qiagen	www1.qiagen.com	01293-422-911
Illumina	www.illumina.com	01799 532 300
Oxford Nanopore	www. nanoporetech.com	0845 034 7900

Speakers

We would like to thank the following for giving seminars during the course:

Professor Julian Parkhill
Dr Jennifer Gardy
Dr Estée Török
Dr Nick Loman
Dr Xavier Didelot
Dr Leila Luheshi

Wellcome Genome Campus Advanced Course

Genomics and Clinical Microbiology

22 January–27 January 2017

Seminar Programme

Monday 23 January

14:00 Kendrew Theatre

‘The Outbreak that Launched a Thousand Analyses: TB in a Canadian Homeless Shelter’

Dr Jennifer Gardy

BC Centre for Disease Control, Canada

Tuesday 24 January

14:00 Kendrew Theatre

‘Translating microbial genomics into clinical practice’

Dr Estée Török

University of Cambridge

Wednesday 25 January

14:00 Kendrew Theatre

‘Identifying signatures of recent pathogen emergence’

Professor Julian Parkhill

Wellcome Trust Sanger Institute

Thursday 26 January

14:00 Kendrew Theatre

‘Real-time genomic surveillance in epidemics’

Dr Nick Loman

University of Birmingham

16:30 Kendrew Theatre

‘Pathogen genomics into practice: a health systems perspective’

Dr Leila Luheshi

PHG Foundation, UK

Friday 27 January

14:00 Kendrew Theatre

‘Within-host evolution of bacterial pathogens and implications for transmission analysis’

Dr Xavier Didelot

Imperial College London

Genomics and Clinical Microbiology 22-27 January 2017							
	Sunday 22/01/2017	Monday 23/01/2017	Tuesday 24/01/2017	Wednesday 25/01/2017	Thursday 26/01/2017	Friday 27/01/2017	
08:30		Scenario Briefings and Updates	Mcobacterial 2a	Mycobacterium 3a	Scenario Briefings/Updates	Mycobacterium WGS analysis	08:30
09:00		B230			B230		09:00
09:30		Mycobacterial 1a	Mycobacterial 2b	Interpretation of MinION data	Web-based genome analysis		09:30
10:00							10:00
10:30				IT room	IT room		10:30
11:00		Tea/Coffee C3-02	Tea/Coffee C3-02	Tea/Coffee C3-02	Tea/Coffee C3-02	Tea/Coffee C3-02	11:00
11:30		Mycobacterial 1b	Meningococcal Serogrouping	Single gene sequence data	Genome upload & gene by gene analysis	Participant report preparation	11:30
12:00			Gastro bioinf theory				12:00
12:30		RT PCR detection (rapid ID example)					12:30
13:00	Registration Buffet Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	13:00
13:30	Restaurant, Hinxton Hall	Kendrew Foyer	Kendrew Foyer	Kendrew Foyer	Kendrew Foyer	Kendrew Foyer	13:30
14:00	Introduction -WTAC Introduction Instructors & Participants	Seminar Jennifer Gardy	Seminar Estée Torok	Seminar Julian Parkhill	Seminar Nick Loman	Seminar Xavier Didelot	14:00
14:30		Kendrew Lecture Theatre	Kendrew Lecture Theatre	Kendrew Lecture Theatre	Kendrew Lecture Theatre	Kendrew Lecture Theatre	14:30
15:00	Library, Hinxton Hall	Library prep (Tagment and Thermal cycler over break)	Gastro bioinf practical Interpretation of hospital data	Mycobacterium 3b	Genome upload & gene by gene analysis	Participant presentations of the clinical scenarios	15:00
15:30	Tea/Coffee		IT room		IT room		15:30
16:00	Cath Arnold Seminar Library, Hinxton Hall	C3-02	C3-02	C3-02	C3-02	IT Room	16:00
16:30	Break	Library prep (clean up)	Interpretation of national data	Genome Assembly & SNP based analysis Tim et al.	Seminar Leila Luheshi	Instructor summaries of clinical scenarios IT Room	16:30
17:00	Martin Maiden Seminar Library, Hinxton Hall	Bioanalyser	IT room		Kendrew Lecture Theatre	Departure	17:00
17:30	Break		MinION practical				17:30
18:00	Stephen Gillespie Seminar Library, Hinxton Hall	Normalisation			Free time		18:00
18:30	Welcome drinks Bar			IT room			18:30
19:00	Supper	Supper & discussion groups	Supper & discussion groups	Supper & discussion groups	Pre-dinner drinks New Space Bar		19:00
19:30	Restaurant, Hinxton Hall	Restaurant	Restaurant	Restaurant	Course Dinner		19:30
20:00		Load MiSeq	PCR design and Implementation	RT Workshop			20:00
20:30				MM, SG, CA			20:30
21:00							21:00
21:30							21:30
22:00					Restaurant		22:00
	Sunday 22/01/2017	Monday 23/01/2017	Tuesday 24/01/2017	Wednesday 25/01/2017	Thursday 26/01/2017	Friday 27/01/2017	

Chapter 4

Introduction

Welcome to the eleventh Genomics and Clinical Microbiology Course at the Wellcome Trust Genome Campus. There has been progress in revolutionising clinical bacteriology with molecular diagnostic tests, but much needs to be done. More molecular assays are entering the portfolios of many hospital laboratories and whole genome approaches are now part of the mainstream of methods to investigate some outbreak situations. Over the last few years we have adapted our course to allow for both scientific developments and diagnostic practice to ensure that we are presenting cutting edge of technology and its implementation in the clinical setting.

For more than a hundred years microbiology diagnosis and epidemiology was dominated by the culture of organisms on artificial media. The smell of agar and the autoclave still permeate every bacteriology laboratory and this has scarcely developed from beyond that of pioneers like Koch and Ehrlich. Many laboratories are spending large sums of money to automate this 19th century technology. Unlike virologists, for whom the isolation of pathogens has been both difficult and insensitive, there has always been the perception that bacteriological culture is an effective way to make a diagnosis. The truth is, however, that this is not really the case anymore if it ever was. The extensive use of antibiotics in domiciliary practice means that patients are usually admitted having already received a potent antibacterial agent. The patient may not be cured but the cultures are likely to be sterile.

The ability to culture micro-organisms has led to a focus among bacteriological laboratories mainly on the organisms that will grow on agar. This has meant that clinical bacteriologists have practiced the art of the possible rather than considering the possibilities of their art. We now know that many important pathogens are not readily cultivatable on artificial media. This is perhaps best illustrated by lower respiratory tract infection. Most laboratories invest major resource to culture *S. pneumoniae*, *Haemophilus influenzae*, *Moraxella catarrhalis* and other organism associated with hospital acquired pneumoniae. The diagnosis of *Mycoplasma pneumoniae* responsible for regular epidemics is often ignored and other bacteria such as *Chlamydia pneumoniae*, *Legionella pneumophila* and *Chlamydia psittaci* are equally neglected despite their prevalence and seriousness. The deficiency of the “agar only” approach to clinical bacteriology is perhaps illustrated by the scramble to improve respiratory diagnosis in the face of novel respiratory organisms such as the SARS and MERS coronaviruses.

Molecular technology has much to offer the diagnostic bacteriologist enabling a diagnosis to be made in infections that are dangerous to grow (e.g., anthrax, plague or *C. psittaci*), where the speed of conventional diagnosis is too slow (*Mycoplasma pneumoniae*, *Mycobacterium tuberculosis*) and where the organism cannot be grown (readily or at all; *Treponema pallidum*, *Tropheryma whippelii*). It can significantly improve the pick-up rate (e.g., *Chlamydia trachomatis* and *Neisseria gonorrhoeae*) in screening programmes or improve the speed of detection of a pathogen to permit rapid isolation of patients and enhance infection control procedures (e.g., real-time PCR for MRSA). In the very immediate future we anticipate that molecular methods will be employed to speed blood culture diagnosis. Also, the application of 16S and 18S PCR to diagnosis has permitted detection and identification of a pathogen in a single reaction from a specimen that is often sterile. This has proved especially

valuable when linked to cloning for difficult mixed infections from deep-seated pus. It has permitted very fastidious anaerobes to be detected and identified the complex flora in specimens from empyema and brain abscess.

Sequencing was once difficult and expensive but systems capable of sequencing a whole bacterial genome in less than 24 hours are increasingly available in clinical laboratories transforming our ability to understand the population genetics and molecular epidemiology of pathogenic organisms and determine the epidemiology of outbreaks in real time. The availability of tools such as Minlon has taken sequencing technology to locations without sophisticated laboratory facilities. Improved bio-informatics techniques mean that the clinical use of whole genome sequencing in the clinical environment is now entering clinical service. To achieve the gains that are within our grasp we need a paradigm shift in the minds of those who practice bacteriology every day. We need to think how we will incorporate these changes into routine practice and how they will affect future research projects. Thus, it is essential for those working in clinical bacteriology to understand the techniques of molecular detection and whole genome sequencing as these approaches start to take over routine diagnosis.

For the infectious diseases epidemiologist molecular methods have already transformed our ability to follow the transmission of infectious agents. This is perhaps best illustrated by *M. tuberculosis* and *N. meningitidis*. Until the description of the IS6110 RFLP methodology there was no effective way of distinguishing strains of *M. tuberculosis*. This robust portable method permitted bacteriologists to detect outbreaks and transmission chains and epidemiologists to determine routes of transmission in larger communities. This method was superseded by MIRU/VNTR and now sequence based methodologies will soon be the norm. This change in practice is reflected in our practical class where you will use whole genome sequencing to unpick a tuberculosis outbreak. Although sero-typing has been available for *N. meningitidis* since the 1930s it is a difficult, expensive technique limited to reference laboratories and did not divide the species up into enough groups that would be useful to study the epidemiology of the disease. Serotyping has never been sufficiently granular to be useful in the study of meningococcal disease or to monitor the implementation of new vaccines that have become available. Multilocus sequence (MLST) typing transformed this situation and has enabled chains of transmission to be followed and has aided our understanding of pathogenicity of this organism and its population genetics. Whole genome sequencing is now becoming the methodology of choice as it is now possible to apply these techniques to outbreaks rapidly and study the transmission routes of this important pathogen for public health purposes.

In this course we have set ourselves the ambitious task of training individuals in the key methodologies of molecular bacteriology. It is impossible, of course, to give a comprehensive account of all of the ways that molecular methods can be applied to clinical bacteriology in just one week but we hope that you will gain an insight into the way that the modern bacteriologist needs to think to implement molecular diagnostics and the growing genomic data for the benefit of your patients. We have chosen to focus on key skills that can be applied generally. Beyond the formal practical classes there will be ample opportunity to talk to the senior faculty who have extensive experience of putting these techniques to use into day-to-day practice. They will also be available during the practical classes so make sure you take the opportunity to discuss your bug/assay with them and explore the implications for the complete change in approach to bacteriology. In particular you should take the opportunity of

discussing how these techniques can be implemented in the clinical environment with the faculty who have been actively engaged in this process for many years.

This course is about equipping you to be leaders in the molecular revolution. To do this we will show you a range of DNA amplification techniques and how the products of these reactions can be variously sequenced, digested and cloned. We will also show you methods that can readily be implemented into clinical practice. In addition, you will have demonstrations of the techniques that may be making a difference to diagnostics now.

It is surprisingly easy to produce data from the molecular laboratory, the real skill is to make sense of it! It is easy to drown in the tide of whole genome sequence results or MLST patterns. An important part of this course will be, therefore, the demonstration of bioinformatics techniques and tricks. New algorithms and fast computers enable us to calculate complex relationships between strains and to interrogate data-bases to identify strains or species. It is important, though, to ensure that the trees and diagrams produced are rooted in the biology and are plausible in the clinical setting.

The increase in the availability of whole genome sequencing in the clinical environment has prompted a comprehensive revision of the course with the inclusion of many more clinical scenarios. This has been done to demonstrate how the techniques can be employed in the diagnostic environment. It will be up to you to discuss with the faculty how to take the information you are gaining in these classes into the environment in which you are working.

One of the most important parts of this course, in addition to the practical classes is the “state of the art” talks. These are delivered by world-leading scientists and will bring some exciting new research areas to your attention. You will have the opportunity of hearing scientists working on the cutting edge of the technological and bioinformatics revolution. The atmosphere of these Wellcome courses is informal and you may gain your best tips from talking to these leading scientists.

If old fashioned microbiologists used to “do it with culture and sensitivity” then molecular biologists use just four techniques in different combinations to weave their magic: endonuclease digestion, ligation, amplification and hybridisation. It is with these four building blocks that we will transform bacterial diagnosis and epidemiology in the 21st Century and we will show you how.

“The future is bright but the future is molecular.” We hope you have a great learning week.

Chapter 5: Gastrointestinal

Genomics and Clinical Microbiology 2017 - Shiga Toxin-producing

Escherichia coli (STEC)

Background

Shiga toxin-producing *Escherichia coli* (STEC), also known as Verocytotoxin-producing *E. coli* (VTEC), is defined by the presence of the phage encoded Shiga toxin (Stx) genes (stx1, stx2 or both). There are more than 400 different serotypes of STEC and over 100 of these are known to cause symptoms of gastrointestinal (GI) disease in humans, including severe bloody diarrhoea and haemolytic uraemic syndrome (HUS).

Previous studies have indicated that the presence of stx2, specifically the stx2a subtype, is more frequently associated with severe disease. Many STEC associated with human disease also have the intimin-encoding gene *eae* (*E. coli* attaching and effacing), located on a pathogenicity island called the locus of enterocyte effacement (LEE), and associated with the intimate attachment of the bacteria to the human gut mucosa. Recently, strains of STEC that do not have the *eae* gene but carry a plasmid encoding *aggR*, associated with the enteroaggregative *E. coli* group, have also been associated with causing HUS. Shiga toxin-producing *Escherichia coli* (STEC) are considered to be a significant threat to public health due to the severity of gastrointestinal symptoms associated with human infection and the risk of cases developing Haemolytic Uraemic Syndrome (HUS). STEC are zoonotic; transmission occurs by direct contact with animals or their environment, or by consumption of contaminated food or water. The infectious dose is low (<10 organisms) and person-to-person spread is common.

In England, the current Standards for Microbiology Investigations protocols are specific for the isolation of non-sorbitol fermenting colonies of *E. coli* serogroup O157 on cefixime tellurite sorbitol MacConkey (CT-SMAC) agar. STEC serogroups other than O157 (non-O157 STEC) are not detected using this method. However, since 2012 the implementation of commercial PCR assays for the detection of STEC in faecal specimens from cases with symptoms of gastrointestinal infection, at a twelve local hospital laboratories, has resulted in an increase in the detection of non-O157 STEC in the UK. Faecal specimens that are PCR positive for the Shiga Toxin (stx) genes at the local hospital laboratories in England are sent to the Gastrointestinal Bacterial Reference Unit (GBRU) at Public Health England (PHE) for isolation of STEC and subsequent serotyping. Recent advances in whole genome sequencing (WGS) have led to the development of a method for high throughput sequencing of bacterial genomes at low cost. WGS was implemented at GBRU for real-time surveillance of STEC in June 2015.

Scenario - story

This scenario is designed to give you an insight into how whole genome sequencing may be useful from different clinical and public health perspectives, as well as the practical details of DNA extraction, whole genome sequencing (WGS) library preparation and bioinformatics analysis.

Chapter 1

You are a clinical microbiologist in a large UK hospital. It is Monday 23rd January and since Friday 20th January, five patients have been admitted to ITU with symptoms of Haemolytic Uraemic Syndrome (HUS). Faecal specimens submitted to the local laboratories were culture negative for the common bacterial gastrointestinal (GI) pathogens (Salmonella, Campylobacter, Shigella and STEC O157) and were couriered to GBRU for further testing. This morning the reference laboratory reported detection of non-O157 STEC from three cases identified Friday 13th January. You need to perform whole genome sequencing on these isolates and analyse the results in order to determine the serotype and virulence profile of the strain associated with the outbreak.

This will involve:

- WGS library prep
- Using the Galaxy bioinformatics platform
- KmerID to speciate from raw sequencing data
- Identification of serotype and virulence profile from the genome
- Performing whole genome SNP analysis & interpretation of the phylogeny of the outbreak strain

Chapter 2

You are an epidemiologist at Public Health England. It is Wednesday 25th January and what started as a cluster of cases linked to a hospital in one region has become a large national outbreak. You will analyse phylogenetic trees to determine the relationship between cases in the national outbreak and generate epidemiological hypotheses from based on the data. This will involve:

- Interpretation of national phylogenetic trees

Chapter 3

You are the clinical microbiologist from chapter 1 again. Public Health colleagues in France and Germany have both reported an increase in the number of HUS cases over the week-end. The French have reported PCR data suggesting that the Shiga toxin subtype is a highly pathogenic type and the Germans have reported that the isolate is multidrug resistant. You know that both these properties are difficult to analyse using

short read Illumina data so you decide to try out the new-fangled MinION machine to sequence the outbreak strain so determine whether the data enables you to improve your analysis of the accessory genome. This will involve the following:

- MinION™ nanopore sequencing Library Preparation
- MinION data analysis

Scenario – practical

Isolating Genomic DNA from Gram negative bacteria

Story:

This morning the reference laboratory reported detection of non-O157 STEC from three cases identified Friday 13th January. You need to perform whole genome sequencing on these isolates and analyse the results in order to determine the serotype and virulence profile of the strain associated with the outbreak.

Principle of the procedure

The Wizard Genomic DNA Purification Kit is designed for isolation of DNA is based on a four step process. The first step in the purification procedure lyses the cells and the nuclei. The cellular proteins are then removed by a salt precipitation step, which precipitates the proteins but leaves the high molecular weight genomic DNA in solution. Finally, the genomic DNA is concentrated and desalted by isopropanol precipitation.

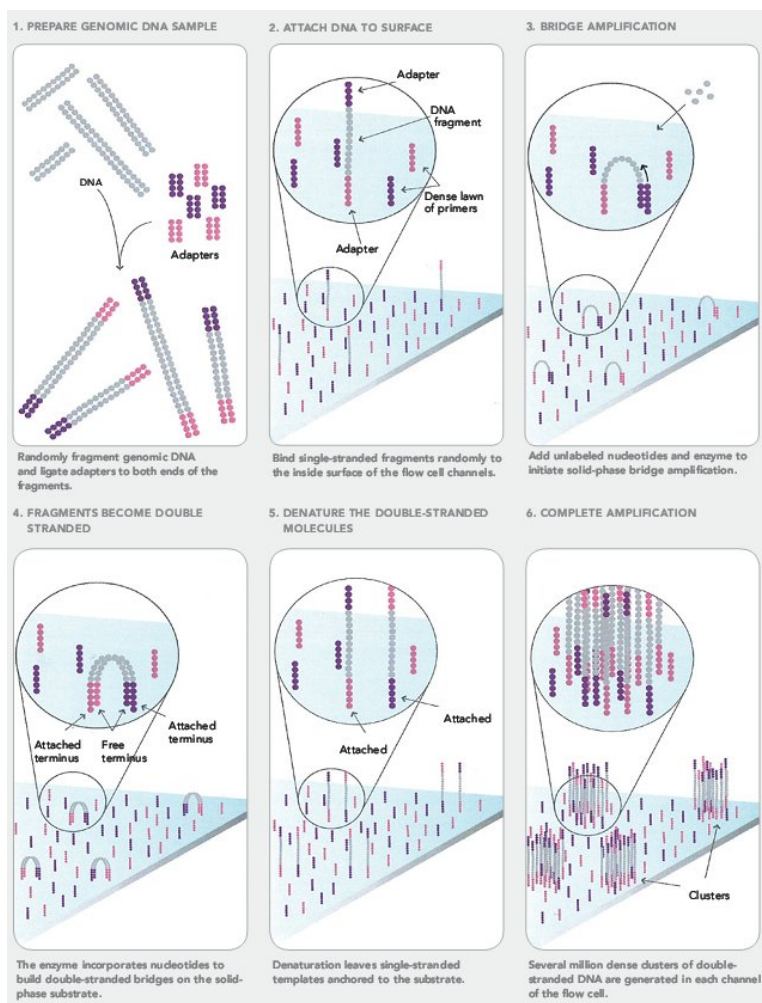
1. Add 1ml of an overnight culture to a 1.5ml microcentrifuge tube.
2. Centrifuge at 13, 000 rpm for 2 min to pellet the cells and remove the supernatant.
3. Add 600ul Nuclei Lysis Solution and gently pipet until the cells are resuspended.
4. Incubate at 80°C for 5 mins to lyse the cells; then cool to room temperature.
5. Add 3ul of RNase solution to the cell lysate. Invert the tube 2-5 times to mix.
6. Incubate at 3 °C for 15-60 mins. Cool the sample to room temperature.
7. Add 20 ul of protein precipitation solution to the lysate. Vortex vigorously at high speed for 20 seconds to mix the protein precipitation solution with the cell lysate.
8. Incubate the sample on ice for 5 mins.
9. Centrifuge at 13, 000 rpm for 3 mins.
10. Transfer the supernatant containing the DNA to a clean microcentrifuge tube containing 600ul of isopropanol. Some supernatant may remain in the original tube containing the protein pellet. Leave this residual liquid in the tube to avoid contaminating the DNA solution with the precipitating protein.
11. Gently mix by inversion until the thread like strands of DNA form a visible mass.
12. Centrifuge at 13,000 rpm for 2 mins.

13. Carefully pour off the supernatant and drain the tube on clean absorbent paper. Add 600 μ l of room temperature 70% ethanol and gently invert the tube several times to wash the DNA pellet.
14. Centrifuge at 13, 000 rpm for 2 mins and carefully aspirate the ethanol.
15. Drain the tube on clean absorbent paper and allow the pellet to air-dry for 10-15 min.
16. Add 100 μ l of DNA Rehydration solution to the tube and rehydrate the DNA by incubating at 65°C for 1 h. Periodically mix the solution gently.
17. Store the DNA at 4°C.

Whole Genome Sequencing – Nextera XT protocol

PRINCIPLE OF PROCEDURE

Using a single transposase enzymatic reaction, sample DNA is simultaneously fragmented and tagged with sequencing adapters. Short sample specific oligonucleotide barcodes are attached to the fragmented DNA. Libraries containing different indexed adapters are then constructed, quantified, pooled in equimolar amounts, and sequenced. Deconvoluting the bar codes informatically then allows multiple libraries to be sequenced on a single flow cell.



Overview of Illumina sequencing. Reproduced from SEQanswers.com

Fragmentation and Tagmentation of Genomic DNA

Personnel Protective Equipment required

Lab coat, nitrile gloves and safety specs

Hazardous substances

80% Ethanol – Flammable, Irritant

Sodium Hydroxide – Corrosive, Irritant, Skin Sensitizer

TD buffer (contains Formamide) – Toxic, Irritant, Teratogen – expectant or new mothers should avoid handling this chemical.

LDR Formamide (MiSeq reagent kit) – Toxic, Teratogen – expectant or new mothers should avoid handling this chemical.

PR2 Incorporation Buffer (MiSeq reagent kit) – Irritant, Skin Sensitizer

Library Normalisation Wash 1 (Nextera XT sample prep. kit) – Mutagen, Teratogen – expectant or new mothers should avoid handling this chemical.

Library Normalisation Additives 1 (Nextera XT sample prep. kit) – Mutagen, Teratogen – expectant or new mothers should avoid handling this chemical.

Dye Concentrate (Agilent high sensitivity DNA kit) (contains DMSO) – Irritant, Flammable

Item	Quantity	Storage
Amplicon Tagment Mix (ATM)	1 tube (7µl)	Ice Bucket
Tagment DNA Buffer (TD)	1 tube (12µl)	Ice Bucket
Neutralize Tagment Buffer (NT)	1 tube (7µl)	Room Temp
1ng Input DNA	Provided at: 5µl @ 0.2ng/µl	-15°to -25°

1. Ensure all reagents are adequately mixed by gently inverting the tubes 3–5 times, followed by a brief spin in a microcentrifuge.
2. Retrieve the tube containing your genomic DNA (Labelled as your group number).
3. Add 10µl of TD Buffer to the DNA sample.
4. Add 5µl of ATM and gently pipette up and down 5 times to mix.
5. Place the sample in a thermocycler and run the following program (with heated lid):

Thermal Cycler Setting	
Program Name:	Nextera XT Tagmentation
Total Sample Volume:	20µl
Parameters:	55°C for 5 minutes Hold at 10°C

6. Once the tubes have reached 10°C immediately remove from the thermal cycler.

7. Add 5µl of NT buffer and pipette mix gently 5 times to ensure that the sample is thoroughly mixed.
8. Pulse spin in a microcentrifuge.
9. Incubate the sample at room temperature for 5 minutes.

PCR Amplification

Retrieve the following reagents and consumables from the ice bucket:

Item	Quantity	Storage
Nextera PCR Master Mix (NPM)	1 tube (17µl)	Ice Bucket
Index 1(i7) & 2 (i5) Primer Mix (index)	1 tube (10µl)	Ice bucket

1. To your DNA sample, add 15µl of NPM and pipette mix 5 times.
2. Add 10µl of the Index Primer Mix.
3. Replace the lid and pulse spin in a microcentrifuge.
4. Place the sample tube onto a thermal cycler using the following parameters:

Thermal cycler settings	
Program Name:	Nextera XT PCR
Total Volume:	50 µl
Parameters:	72°C for 3 minutes 95°C for 30 seconds 12 cycles of: 95°C for 10 sec 55°C for 30 sec 72°C for 30 sec 72°C for 5 minutes Hold at 10°C

Ampure Clean-up

Item	Quantity	Storage
Resuspension Buffer	1 tube (55µl)	Room Temperature
AMPure XP beads (XP)	1 tube	Room Temperature
Fresh 80% ethanol	1 x 15 ml tube	N/A

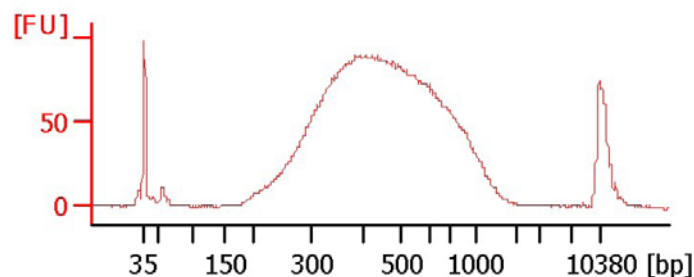
1. Retrieve your sample from the thermal cycler. Briefly vortex and pulse spin in a microcentrifuge.
2. Label a new tube with your group number.
3. Transfer 50µl of the PCR product from the original sample tube to the new clean tube.

4. Vortex the AMPure XP beads for 30 seconds to ensure that the beads are evenly dispersed.
5. Add 30µl of AMPure XP beads to the tube. Pipette mix 10 times.
6. Incubate at room temperature for 5 minutes.
7. Place the tube on a magnetic stand for 2 minutes or until the supernatant has cleared.
8. Carefully remove and discard all the supernatant from each well.

(Please note: Your DNA is now bound to the Ampure XP beads, avoid disturbing the beads. If any beads are inadvertently aspirated into the tips, dispense the beads back into the tube and let it rest on the magnet for 2 minutes and confirm that the supernatant has cleared)
9. Add 200µl of 80% ethanol to the tube whilst still on the magnetic stand. Incubate for approx 30 secs and carefully remove and discard the supernatant. Do not remove the beads.
10. Repeat the ethanol wash in step 9.
11. If required use a P10 pipette to remove excess ethanol, so as to not remove the beads.
12. With the samples still on the magnetic stand, allow the beads to air-dry for 5 minutes.
13. Remove the tube from the magnetic stand and add 52.5µl of RSB.
14. Gently pipette mix up and down 10 times.
15. Incubate at room temperature for 2 minutes.
16. Place the tube back on the magnetic plate for 2 minutes (or until the supernatant has cleared).
17. Label a new tube 'CAN' (Clean Amplified NTA) & your group number.
18. Carefully transfer 30µl of the supernatant to the CAN tube.

Library Quality and Quantity Check (Bioanalyzer)

- The size distribution of your library can be checked by running 1µl of it on an Agilent Technologies 2100 Bioanalyzer using a High Sensitivity DNA chip.
- Typical libraries show a broad size distribution from ~250-1000bp, with an average of ~400–500bp.



EXAMPLE OF DNA LIBRARY SIZE DISTRIBUTION

Setting up the Chip Priming Station

1. Insert the syringe into the clip

2. Slide it into the hole of the luer lock adapter and screw it tightly to the chip priming station.
3. Check base plate is in position C.
4. Adjust the syringe clip to the lowest position.

Checking the Chip Priming Station for Good Seal — Seal Test

1. Make sure the syringe is tightly connected to the Chip Priming Station.
2. Pull the plunger of the syringe to the 1.0ml position (plunger pulled back).
3. Place an empty chip in the Chip Priming Station.
4. Close the Chip Priming Station and make sure to lock it by pressing the cover, the lock of the latch will audibly click when it closes.
5. Press the plunger down until it is locked by the clip.
6. Wait for 5 seconds and press the side of the clip to release the plunger.
7. Appropriate sealing is verified if the plunger moves back up to the 0. ml mark within less than 1 second.

Preparing the Gel-Dye Mix

1. Allow High Sensitivity DNA dye concentrate (blue) and High Sensitivity DNA gel matrix (red) to equilibrate to room temperature for 30 min.
2. Add 15µl of High Sensitivity DNA dye concentrate (blue) to a High Sensitivity DNA gel matrix vial (red).
3. Vortex solution well and spin down. Transfer to spin filter.
4. Centrifuge at 2240 g \pm 20 % for 10 min. Protect solution from light. Store at 4 °C.

Loading the Gel-Dye Mix

1. Allow the gel-dye mix equilibrate to room temperature for 30 min before use.
2. Put a new High Sensitivity DNA chip on the chip priming station.
3. Pipette 9.0µl of gel-dye mix in the well, marked G.
4. Make sure that the plunger is positioned at 1ml and then close the chip priming station.
5. Press plunger until it is held by the clip.
6. Wait for exactly 60seconds then release clip.
7. Wait for 5seconds, and then slowly pull back the plunger to the 1 ml position.
8. Open the chip priming station and pipette 9.0µl of gel-dye mix in the two additional wells marked G.

Loading the Marker

1. Pipette 5µl of marker (green) in all sample and ladder wells. Do not leave any wells empty.

Loading the Ladder and the Samples

1. Pipette 1µl of High Sensitivity DNA ladder (yellow) in the ladder well.

2. In each of the 11 sample wells pipette 1µl of sample (used wells) or 1µl of marker (unused wells).
3. Put the chip horizontally in the adapter and vortex for 1 min at the indicated setting (2400 rpm).
4. Run the chip in the Agilent 2100 Bioanalyzer within 5 min.

Starting the Run

1. Select the High Sensitivity assay from the Assay menu.
2. Enter details in the sample name table.
3. Click the Start button in the upper right of the window to start the chip run.
4. The incoming raw signals are displayed in the Instrument context.
5. After the run is finished, remove the chip.

Average Library Fragment Size

1. When viewing the results of the run, navigate to the 'Region Table Bar'
2. Move the blue bars to either side of the curve.
3. The average length in bp will be displayed
4. Use this value to calculate the molarity using the values from the Qubit quantitation assay.

Library Normalisation

NB: REAGENTS CONTAIN FORMAMIDE: TO BE CONDUCTED INSIDE FUME HOOD

Item	Quantity	Storage
Library Normalisation Additives1 (LNA1)	1 tube	Ice bucket
Library Normalisation Beads 1 (LNB1)	1 tube	Ice bucket
Library Normalisation Wash 1 (LNW1)	1 tube	Ice bucket
Library Normalisation Storage Buffer 1 (LNS1)	1 tube	Room Temperature
Fresh 0.1 N NaOH	1 tube	Room Temperature
96-well plate	1 plate	N/A
15 ml conical tube	1 tube	N/A

1. Remove the LNA1, LNB1 and LNW1 from the ice bucket (and LNS1 from its storage location) and bring to room temperature. Vortex for approximately 1 minute prior to use, ensuring any precipitate is fully resuspended.

2. Retrieve the 'CAN' tubes and, pulse-spin in a microcentrifuge. Pipette mix the sample thoroughly.
3. Label a 96-well plate 'Normalisation'
4. Transfer 20µl of supernatant from each 'CAN' tube to a separate well in the 'Normalisation' plate.
5. Volumes of LNA1 and LNB1 required:

Sample Number	Reagents	Volume
12	LNA1	550µl
	LNB1	100µl

6. Ensure the LNA1 and LNB1 is thoroughly mixed just prior to use.
7. Combine the required volumes of LNA1 and LNB1 in a 15ml conical tube. Vortex thoroughly for 30 seconds and invert 15-20 times.
8. Add 45µl of LNA1/LNB1 mix to each sample in the 'Normalisation' plate.
9. Seal the plate and shake at 1800rpm for 30 minutes.
10. Place the plate on the magnetic stand and remove the plate seal. Incubate on the magnet for 2 mins and carefully remove all of the supernatant (with multichannel).
(Please note: Avoid disturbing the beads during this step).
11. Remove the plate from the magnetic stand and add 45µl of LNW1 to each sample well.
12. Seal the plate and place onto a plate shaker at 1800rpm for 5mins.
13. Place the plate on the magnetic stand and remove the plate seal. Incubate on the magnet for 2mins and carefully remove the supernatant (with multichannel).
14. Repeat wash steps 11 – 13.
(Please note: Ensure excess LNW1 is removed, if required use a P10 pipette or multi-channel to remove residues).
15. Remove the plate from the magnetic stand and add 30µl of 0.1 N NaOH to each sample well to elute sample.
(Please note: Only use freshly prepare 0.1N NaOH and do not store).
16. Seal the plate with and shake at 1800rpm for 5 minutes.
17. Label a new 96-well plate 'FINAL'.
18. Pipette 30µl of LNS1 to the appropriate wells on the 'FINAL' plate.
19. Remove the 'Normalisation' plate from the shaker and check to ensure all samples are completely resuspended, if not, pipette mix and place back onto the shaker for a further 5minutes @1800rpm.
20. Place the 'Normalisation' plate on a magnetic stand and remove the plate seal. Incubate for 2 minutes.
21. Transfer 30µl the supernatant from the 'Normalisation' plate to the 'FINAL' plate.
22. Seal the 'FINAL' plate and centrifuge at 1000xg for 1 minute.

Loading the MiSeq

Preparing the 20pM PhiX (control) Library

- Gently vortex and pulse the 10nM PhiX Library.
- Combine 2µL 10nM PhiX Library and 3µL EBT buffer to make a 4nM PhiX library.
- Add 5µL 0.2M NaOH to the above Eppendorf to make 2nM PhiX library.
- Vortex the Eppendorf and pulse spin.
- Incubate for 5 minutes at room temp to denature the PhiX library.
- Add 990µL HT1 to the Eppendorf to make 20pM denatured PhiX library.
Store the 20pM denatured PhiX library between -15°C to -25°C. Dispose after 3weeks.

Preparing the 'POOLED AMPLICON LIBRARY' (PAL) tube

- Retrieve the 'FINAL' plate and place plate shaker and mix for 1 minute at 1500rpm and pulse-spin in a plate centrifuge
- Place onto a magnetic stand (to collect any remaining beads)
- Pool 5µL of sample from the 'FINAL' plate into an eppendorf labelled 'Pool'.

Preparing the 'DILUTED AMPLICON LIBRARY' (DAL) tube

- Set a heat block to 96°C and prepare an ice water bath
- Gently vortex and pulse spin the PAL tube.
- Vortex HT1 to remove all trace of precipitate.
- Label a clean 1.5mL LoBind Eppendorf with 'DAL'.
- Add 22µL of 'PAL' and 588µL HT1 to the 'DAL' tube.
- Gently vortex and pulse-spin.
- Incubate the 'DAL' tube at 96°C for 2 minutes.
- Invert the 'DAL' tube twice and immediately place into the ice water bath.
- Incubate the 'DAL' tube for 5 minutes.
- Add 18µL denatured 20pM PhiX library to the 'DAL' tube.
- Gently vortex and pulse spin – keep on ice
- The 'DAL' tube is now ready for loading onto the MiSeq

Preparing MiSeq Reagent Cartridge

ALLOW APPROXIMATELY ONE HOUR FOR THAWING CARTRIDGE

1. Place the reagent cartridge in a water bath containing enough room temperature deionized water to submerge the base of the reagent cartridge up to the water line printed on the reagent cartridge. Do not allow the water to exceed the maximum water line.
2. Allow the reagent cartridge to thaw in the room temperature water bath for approximately one hour or until completely thawed.
3. Remove the cartridge from the water bath and gently tap it on the bench to dislodge water from the base of the cartridge. Dry the base of the cartridge. Make sure that no water has splashed on the top of the reagent cartridge.
4. Invert the reagent cartridge to mix the thawed reagents, and then visually inspect that all positions are thawed.
5. Visually inspect the reagent marked IMF (Position 1) to make sure that it is fully mixed and free of precipitates.
6. **NOTE:** The MiSeq sipper tubes go to the bottom of each reservoir to aspirate the reagents, so it is important that the reservoirs are free of air bubbles.
7. Place the reagent cartridge on ice or set aside at 2° to 8°C until you are ready to set up your run.

MiSeq Instrument Prep:

Clean the Flow Cell

1. Wash the flow cell with Millipore water
2. Dry any excess water with a lint-free lens cleaning tissue, and visually inspect to make sure that the flow cell ports are free of obstructions and that the gasket is well seated around the flow cell ports.

3. If the gasket appear to be dislodged, gently press it back into place until it sits securely around the flow cell ports.

Loading the Flow Cell

1. Raise the flow cell compartment door, and then press the release button to the right of the flow cell latch. The flow cell latch opens.
2. Visually inspect the flow cell stage to make sure it is free of lint. If lint or other debris is present, clean the flow cell stage using an alcohol wipe or a lint-free tissue moistened with ethanol or isopropanol. Carefully wipe the surface of the flow cell stage until it is clean and dry.
3. Hold the flow cell by the edges of the flow cell cartridge near the Illumina label.
4. Make sure the label is facing upward and place the flow cell on the flow cell stage.
5. Gently press down on the flow cell latch to close it over the flow cell. You will hear a click when the flow cell latch is secure.
6. As you close the flow cell latch, two alignment pins near the hinge of the flow cell latch properly align and position the flow cell.
7. Check the lower-left corner of the screen to confirm that the flow cell RFID was successfully read.
8. Close the flow cell compartment door.
9. Select Next on the Load Flow Cell screen. The Load Reagents screen opens

Loading Reagents

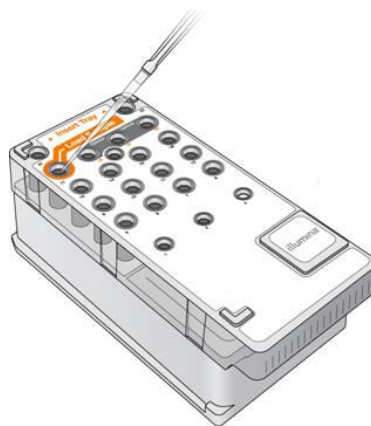
1. Remove the bottle of PR2 from fridge. Gently invert the bottle to mix the PR2 bottle, and then remove the lid.
2. Open the reagent compartment door.
3. Raise the sipper handle until it locks into place.
4. Place the PR2 bottle in the indentation to the right of the reagent chiller.
5. Make sure that the waste bottle is empty. If required, empty the contents into the appropriate waste container.
6. Slowly lower the sipper handle. Make sure that the sippers lower into the PR2 and waste bottles.
7. Check the lower-left corner of the screen to confirm that the RFID of the PR2 bottle was read successfully.
8. Select Next on the Load Reagents screen.

Load Sample Libraries onto Cartridge

1. Use a clean 1 ml pipette tip to pierce the foil seal over the reservoir labelled Load Samples.

NOTE: Do not pierce any other reagent positions. Other reagent positions are pierced automatically during the run.

2. Pipette 600µl of your sample libraries into the Load Samples reservoir. Take care to avoid touching the foil seal as you dispense your sample.



3. Proceed directly to the run setup steps using the MiSeq Control Software (MCS) interface.

Load the Reagent Cartridge

NOTE Do not leave the reagent chiller door open for extended periods of time.

1. Open the reagent chiller door.
2. Hold the reagent cartridge on the end with the Illumina label, and slide the reagent cartridge into the reagent chiller until the cartridge stops.
3. Close the reagent chiller door.
4. Check the lower-left corner of the screen to confirm that the RFID of the reagent cartridge was read successfully.
5. Close the reagent compartment door.

Select Next on the Load Reagents screen. The Review screen opens.

Sample Sheet Set-up (Illumina Experiment Manager)

The Illumina Experiment Manager is a wizard-based application that guides you through the steps to create your sample sheet.

Starting the Run

- After you have loaded the flow cell and reagents, the MCS interface prompts you to review run parameters and perform a pre-run check before beginning the run.

Getting started with Galaxy bioinformatics platform

For the bioinformatics section of the analysis, we will be using the Galaxy bioinformatics website. Galaxy is an application that allows bioinformaticians to host their own applications for users to interact with via a web browser rather than the command line. There is a public instance available at <https://usegalaxy.org/>, today we are going to be using the PHE instance.

1. Open the Chrome internet browser via the Start menu or desktop shortcut.
2. In Chrome navigate to <http://bioinformatics-galaxy.phe.org.uk/root>
3. You should be able to see something like the below, without the coloured boxes.

Menu bar – you will only need two of these sections today

1. 'Analyze Data' is the main section, if you get lost, this is 'home'
2. 'Shared Data' is where we have some data ready for you to analyse

The screenshot shows the Galaxy / PHE web interface. At the top is a menu bar with 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. Below the menu bar, the interface is divided into three main sections. On the left is a 'Tools' panel with a search bar and a list of tool categories including 'Get Data', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Operate on Genomic BED Intervals', 'Convert Formats', 'Genome Annotation', 'Extract Features', 'Picard tools', 'Statistics', 'RNA Seq - TESTING!', 'Graph/Display Data', 'NGS: VCF processing', 'Metagenomics', 'NGS: BAM Tools', 'NGS: GATK2', 'NGS: Simulation', 'Multiple Alignments', 'Phylogenetics', 'FASTA manipulation', 'NCBI BLAST+', 'NGS: PHE internal tools', 'NGS: QC and manipulation', 'NGS: Assembly', 'NGS: Mapping', 'NGS: SAM Tools', and 'Workflows'. In the center is the 'Main analysis window' which contains a 'Welcome to the PHE Galaxy environment' message, an 'IMPORTANT!' notice stating the application is for research and training purposes only, the Public Health England logo, and a paragraph about the Galaxy project. On the right is a 'History' panel with a search bar and a message stating 'This history is empty. You can load your own data or get data from an external source'.

Tool menu – where you can choose which analysis to run

Main analysis window where you will specify options for your analysis

History – you can see the progress and results of your analysis here.

Story:

It is Tuesday 24th January and the Illumina sequencing data from the run you prepared yesterday will be available tomorrow morning. Meanwhile, the Germans have asked for help analysing the WGS data they have from their outbreak strain. First of all, you will confirm that you have sequenced an isolate of *E. coli*, and then you want to identify the serotype and virulence profile.

KmerID to speciate from raw sequencing data

Principle of the procedure

For this we will use a PHE method called KmerID, the code for this process is available from <https://github.com/phe-bioinformatics> but we will use Galaxy to run this software, rather than the command line. KmerID determines a similarity index between the FASTQ reads and each of the 1769 published reference genomes by calculating the percentage of 18-mers in the reference that are also present in the FASTQs. Only 18-mers that occur at least twice in the FASTQ are considered present. Mixed cultures are detected by comparing the list of similarities between the sample and the references with the similarities of the references to each other, and filtering this comparison for inconsistencies. Other publicly available tools that perform a similar function are Kraken (<https://ccb.jhu.edu/software/kraken/>) and One Codex (<https://www.onecodex.com/>), which I would recommend looking at because you don't need to install anything and it has a very nice user interface.

Questions to answer during this section

1. What is the top hit in the Kmer ID analysis
2. Is the sample mixed?


Practical steps

1. Go to 'Shared Data' on the menu bar.
2. Select 'Data Libraries' from the drop-down menu
3. On the new page that loads, select 'WTAC' for Wellcome Trust Advanced Course and then '2017'
4. Click on the small blue arrow next to 'German Data'
5. Check the tick boxes next to **both** fastqs (these are sequence data files and end with .fastq)
6. Then, check that the current action next to 'For selected datasets' is 'import to current history' and click 'Go'.



7. You should get a green notification bar saying that 2 datasets have been imported into your current history.
8. On the menu bar, click 'Analyze data' to return to the Home screen. You should now see two items in your history bar on the right hand side of the screen, these are your two fastqs.
9. On the left hand side of the Home screen is the tool menu, near the bottom there is an option 'NGS: PHE internal tools', clicking on this option will reveal a lot of options, including 'Kmer ID'. Click on the underlined Kmer ID section (as below).

Kmer ID Quickly get whole genome similarity between reads and references

10. The main analysis window should now look like the below. By default, one of the fastqs will be selected for analysis. It does not matter that only one fastq is being analysed, or which half of the pair you analyse. You can leave all the options as default and 'Execute'. This will add two items to your history bar which will initially be grey and then turn yellow.

 **Kmer ID** Quickly get whole genome similarity between reads and references (Galaxy Tool Version 1.0.0) Options

NGS reads

Must be in fastq format

Compare against top genera

Use 5 unless instructed to do otherwise.

Amount of top hits to show

Anything >100 is not likely to be useful.

Give your sample a name.

Do not leave this blank, please.

Conduct a rudimentary check for sample mixing.

Potentially useful information for an additional 2 minutes runtime.


The tool will compute a similarity between the reads in your NGS sample and a predefined set of reference genomes from NCBI.

It will report back the closest matches in your reference sequences.

The mixing option will use the list of top similarities and compare it to a list of similarities between reference genomes.

It will list genomes for which the similarity between the reads and the references is not explained by the similarities between the reference genomes.

Author: ulf.schaefer@phe.gov.uk

11. After about 10 minutes the yellow boxes will turn green and the analysis will be complete. Click on the eye symbol  next to the item in your history that says 'Kmer ID on data 2: Mixing check'. This should change the contents of the main analysis window. What is the top hit?

Characterising the Genome

Principle of the procedure

The presence or absence of specific loci or specific allelic variants are important to characterise the strain and its pathogenic potential.

MLST

Achtman et al. proposed a sequenced based approach, multilocus sequence typing (MLST), based on the sequences of multiple house-keeping genes. Isolates that possess identical alleles for the seven gene fragments analysed are assigned a common sequence type (ST) and related STs from clonal complexes (CCs).

SEROTYPE

Pathogenic *E. coli* strains can be categorized based on elements that can elicit an immune response in animals, namely: the O antigen (part of the lipopolysaccharide layer) and the H antigen (the flagellin). The O antigen is used for serotyping *E. coli* and these O group designations go from O1 to O181 and are encoded by the *rfb* gene cluster. The H antigen is a major component of flagella, involved in *E. coli* movement. It is generally encoded by the *fliC* gene. There are 53 identified H antigens, numbered from H1 to H56.

VIRULENCE FACTORS

Several virulence factors have been identified in *E. coli* and are used to characterise strains into pathotypes. Virulence factors include toxins (e.g. shiga toxin and enterotoxins) and adherence mechanisms (e.g. intimin and fimbriae).

ANTIMICROBIAL RESISTANCE

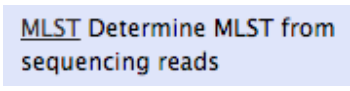
Antimicrobial resistance is a growing problem in microbiology and especially significant in gram negative bacteria such as *E. coli*. Antimicrobial sensitivity of isolates to a battery of therapeutics is often used to guide treatment of infection.

Questions to answer during this section

1. What MLST sequence type is the isolate?
2. What is the serotype?
3. What is the pathotype?
4. What is the Shiga toxin subtype?
5. What is the genotypic antimicrobial profile of the isolate?
6. Do these characteristics tell you anything about the origin of this strain?

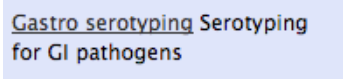
Practical steps

MLST

1. As we already have the fastqs, we can go straight to running the analysis. On the tool menu, in NGS: PHE internal tools, select MLST.
A blue rectangular button with the text "MLST Determine MLST from sequencing reads" in a sans-serif font.
2. In the main analysis window, put the fastq that ends R1.fastq into the 'Read dataset for direction 1' box, and the fastq that ends R2.fastq into the 'Read dataset for direction 2'. Under 'Select the organism' choose *Escherichia coli*. Click Execute.
3. After 10-15 minutes, the three boxes in your history should have turned Green. Click the eye next to the box with the title that ends 'MLST result XML'. The ST result is the fourth line in the main analysis screen <result type="MLST" value="X"> - what is the MLST value?

SEROTYPE & PATHOTYPE

1. On the tool menu, in NGS: PHE internal tools, select Gastro Serotyping. This will scan our FASTQ reads against a database of O & H determinant genes and some virulence genes.

A blue rectangular button with the text "Gastro serotyping Serotyping for GI pathogens" in a sans-serif font.

2. In the main analysis window, put the fastq that ends R1.fastq into the 'Read dataset for direction 1' box, and the fastq that ends R2.fastq into the 'Read dataset for direction 2'. Click Execute.
3. After 10-15 minutes, the box in your history should have turned Green. Click the eye next to the box with the title that ends 'gastro serotyping XML'. What are the results in the "o" and "h" fields? Are there any other positive matches?

SHIGA TOXIN SUBTYPE

1. On the tool menu, in NGS: PHE internal tools, select Stx subtyping.

Stx subtyping Shiga toxin
subtyping for *E. coli* and
Shigella

2. In the main analysis window, put the fastq that ends R1.fastq into the 'Read dataset for direction 1' box, and the fastq that ends R2.fastq into the 'Read dataset for direction 2'. Click Execute.
3. After 10-15 minutes, the box in your history should have turned Green. Click the eye next to the box with the title that ends 'stx subtyping XML'. How many *stx* genes does this strain have and what are the subtypes?
4. What pathotype of *E. coli* is this sample?

ANTIMICROBIAL GENOTYPE

1. On the tool menu, in NGS: PHE internal tools, select Gastro resistance finder.

Gastro resistance finder Detect
resistance genes for GI
pathogens

2. In the main analysis window, put the fastq that ends R1.fastq into the 'Read dataset for direction 1' box, and the fastq that ends R2.fastq into the 'Read dataset for direction 2'. Click Execute.
3. After 10-15 minutes, the box in your history should have turned Green. Click the eye next to the box with the title that ends 'resistance finder XML'. What classes of antibiotic have resistance mechanisms identified? What antibiotic would you recommend for treatment?

Performing whole genome SNP analysis & interpretation of the phylogeny

Story:

Scientists at Public Health England (PHE) curate a database containing over 3,000 isolates of STEC from clinical cases, animals and food. The majority of the cases acquired their infection in the UK but between 20-30% of cases travelled abroad in the week prior to onset of infection. Travel history and other epidemiological data are stored in the STEC Enhanced Surveillance System (SESSy). In the next part of the practical, you will compare the sequence from the German case with those sequences in the PHE database.


Principle of the procedure

Whole genome Single Nucleotide Polymorphism analysis involves ‘mapping’ the raw fastqs of a sample against an appropriate reference genome (i.e. same or closely related Sequence Type). The result of the mapping is then analysed to find differences (specifically, single nucleotide mutations), between the reference genome and the strain being analysed (the fastq). All the differences between a group of strains and the reference genome are collated and used to create a ‘pseudo-sequence’ of variant positions. This sequence can then be processed by phylogenetic algorithms to produce a phylogenetic tree. This process involves three main bioinformatics steps (mapping, SNP calling, collating variant positions) which are going to be carried out for you by SnapperDB. You will use two functions of SnapperDB, ‘fastq_to_db’ to upload the variant information into the database, and ‘get_the_snps’ to collate the variant positions on a set of interest.

Questions to answer during this section

1. How many SNPs is the outbreak from the most closely related isolate in the PHE database
2. Using the epidemiological data linked to the other closely related isolates, can you infer anything about the source of the outbreak strain?




Practical steps

1. First, we need to upload our data to the SNP database. On the tool menu, go to NGS: PHE internal tools and select ‘fastq to db’ from the SnapperDB section. In the main analysis window, select the R1 reads in the forward menu and the R2 read in the reverse menu. In the config file box, type ‘O26_wtac_config.txt’ – this must be **EXACTLY** correct (don’t include the quotation marks), or SnapperDB will not work. Then execute.
2. To see the progress and the different sub programs called click on the  icon. After 15-30 mins the boxes in the history window will turn green and the job will be completed, we are now ready to ‘get_the_snps’.

- On the menu bar go to 'shared data' -> Data libraries -> WTAC -> 2017->German Data and add the 'strain_list_1' file to your current history. If you need more details on how to do this, consult the first part of the KmerID practical steps as they are very similar.
- Return to the 'Analyze Data' page. On the tool menu, go to NGS: PHE internal tools and then 'get the snps'. In the main analysis window select 'strain_list_1' as the strain list, set 'O26_wtac_config.txt' as the config file, select alignment type as 'accessory'. Select 'No' to the question 'Would you like the reference genome in alignment' and 'Yes' to produce a distance matrix of SNP distances and list of annotated variants. Click execute.

get the SNPs retrieve SNPs from a database (Galaxy Version 1.0) Options

Strain list

   79: list ▼

Please upload strain list file.

The name of a config file in the user_configs directory

O26_wtac_config.txt

SNP cut off; strains more than this number of SNPs from the reference will be excluded from the analysis.

3000

Integer type. A sensible starting point is 3000.

Alignment type

accessory ▼

Accessory alignment cutoff

80

The percentage of sequences that must have no N in the relevant positions.

Would you like the reference genome in the alignment?

☐ Yes ☒ No

Produce distance matrix of SNP differences?


☐ Yes ☒ No

Produce an annotated list of variants?

☐ Yes ☒ No

For further information please contact: tim.dallman@phe.gov.uk

- After 1-2 minutes, the items in your history should turn from yellow to green. Click on the eye to inspect the results. There is a 'pseudo-sequence' of all the variant positions, a distance matrix of the pairwise SNP differences and a list that tells you where those SNPs are in the genome.
- How many SNPs are there across the alignment? What percentage are in genes? What percentage are synonymous, non-synonymous?
- How many SNPs between the German isolate and the closest isolate?
- Click on the history item title to expand the history item. You should now be able to see a floppy disk two thirds down the left hand side of the expanded history item. Click on this floppy disk to download the pseudosequence, save it somewhere easily retrievable e.g. your Desktop.

9. Now we need to analyse the pseudo-sequence to generate a phylogenetic tree. On your computer, open the free application MEGA , (available from <http://www.megasoftware.net/>) which is a Graphic User Interface (or GUI, pronounced gooey, you may hear bioinformatics types referring to software 'having a gooey') for phylogenetic analysis.
10. In MEGA, go to file -> open a file/session -> select your saved pseudo-sequence file -> select 'Analyze' in the pop up box -> ensure 'nucleotide sequences' selected in the next pop up box, press ok -> select No when MEGA asks you 'Protein coding nucleotide sequence data?'. Then you are ready to analyse your data.
11. In the MEGA window select the 'Phylogeny' box -> Construct/Test Maximum Likelihood Tree -> click yes in the dialogue box -> leave all the options as their defaults -> click on 'Compute'
12. After 3-5 minutes (this is why bioinformaticians are always on twitter, lots of 5 minute breaks) your tree should be finished. Now we need to analyse the tree.

MinION™ nanopore sequencing Library Preparation

Story:

Public Health colleagues in France and Germany have both reported an increase in the number of HUS cases over the week-end. The French have reported PCR data suggesting that the Shiga toxin subtype is a highly pathogenic type and the Germans have reported that the isolate is multidrug resistant. You know that both these properties are difficult to analyse using short read Illumina data so you decide to try out the new-fangled MinION machine to sequence the outbreak strain so determine whether the data enables you to improve your analysis of the accessory genome. This will involve the following:

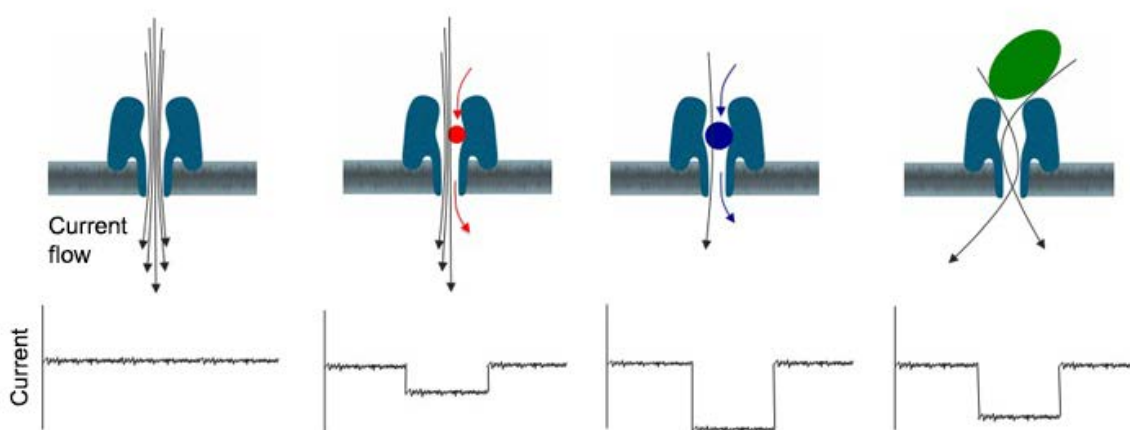
- MinION™ nanopore sequencing Library Preparation
- MinION data analysis

In contrast to the Illumina protocol, the Oxford nanopore MinION uses intact DNA strands and data is analysed in real time.

PRINCIPLE OF PROCEDURE

Sample DNA is prepared so that it has a hairpin structure at its end, allowing both stands to be read (sense and anti-sense).

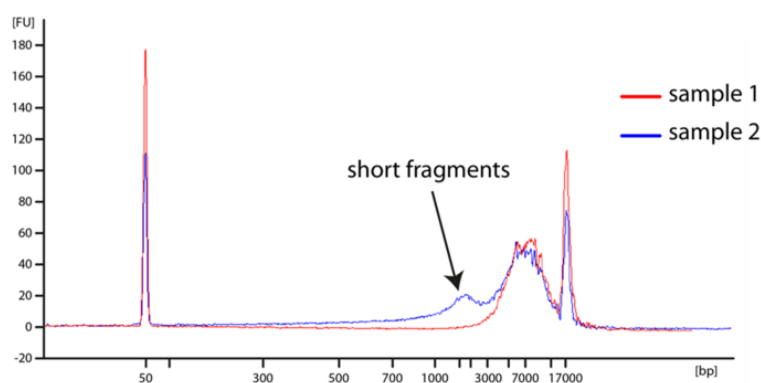
Samples are loaded onto the flowcell which contain nanopores set in a polymer membrane. An ionic current passes through the nanopores and as single molecules enter the pore this current is disrupted. Each of the DNA bases, G, A, T and C, creates a characteristic disruption in current allowing the molecule to be identified.



DNA Fragmentation (Optional)

Item	Quantity	Storage
<1µg DNA sample	1 tube	Ice bucket
Covaris g-TUBE	-	Room Temperature

1. Transfer <1µg genomic DNA in 46µl to the Covaris g-TUBE
2. Spin the g-TUBE for 1 minute
3. Remove and check all the DNA has passed through the g-TUBE
4. If DNA remains in the upper chamber, spin again for 1 minute at the same speed
5. Remove g-TUBE, invert the tube and replace into the centrifuge
6. Spin the g-TUBE for 1 minute to collect the fragmented DNA
7. Remove and check the DNA has passed into the lower chamber
8. If DNA remains in the upper chamber, spin again for 1 minute
9. Transfer the 46 µl fragmented DNA to a clean 1.5 ml Eppendorf DNA LoBind tube
10. Analyse 1 µl of the fragmented DNA for fragment size, quantity and quality using the Agilent Bioanalyzer (optional)
11. Below is an example of a successful fragmentation (sample 1) and an unsuccessful fragmentation (sample 2). The trace obtained for sample 2 shows a shoulder as a result of the presence of smaller fragments; this is indicative of substantial shearing/degradation of the input material and is likely to reduce the quality of the library preparation and the read length distribution.



End-prep

Item	Quantity	Storage
NEBNext End repair / dA-tailing	1 tube	Ice bucket
Freshly prepared 70% ethanol	500µl	Room Temperature
Agencourt AMPure XP beads	1 tube	Room Temperature

1. In a 1.5ml DNA LoBind Eppendorf tube mix together the following:

Reagent	Volume
<1 ug DNA	45µl
Nuclease Free water	5µl
Ultra II End-prep reaction buffer	7µl
Ultra II End-prep enzyme mix	3µl
Total	60µl

2. Mix gently by inversion and spin down.
3. Using a thermal cycler, incubate for 5 minutes at 20°C and 5 minutes at 65°C
4. Resuspend AMPure XP beads by vortexing.
5. Add 60µl of resuspended AMPure XP beads to the end-prep reaction and mix by pipetting.
6. Incubate on a rotator mixer for 5 minutes.
7. Spin down the sample and pellet on a magnet. Keep the tube on the magnet, and pipette off the supernatant.
8. Keep on magnet, wash beads with 200µl of freshly prepared 70% ethanol without disturbing the pellet. Remove the ethanol using a pipette and discard. Repeat.
9. Spin down and place the tube back on the magnet. Pipette off any residual ethanol. Briefly allow to dry.
10. Remove the tube from the magnetic rack and resuspend pellet in 31µl nuclease-free water. Incubate for 2 minutes.
11. Pellet beads on magnet until the eluate is clear and colourless.
12. Remove and retain 31µl of elute into a clean 1.5ml Eppendorf DNA LoBind tube.
13. Quantify 1µl of end-prepped DNA using a Qubit fluorimeter – aiming to recover >700ng.

Ligation of Barcode Adapter

Item	Quantity	Storage
NEB Blunt/TA ligase master mix	1 tube	Ice bucket
Barcode Adapter	1 tube	Ice bucket
Freshly prepared 70% ethanol	500µl	Room Temperature
Agencourt AMPure XP beads	1 tube	Room Temperature

1. Label a 1.5ml Eppendorf DNA LoBind tube with your group number.
2. Add the following reagents, mixing by inversion between each sequential addition:

Reagent	Volume
End prep DNA	30µl
Barcode Adapter	20µl
Blunt/TA Ligase Master Mix	50µl
Total	100µl

3. Mix gently by inversion and spin down.
4. Incubate the reaction for 10 minutes at room temperature.
5. Resuspend the AMPure XP beads by vortexing.
6. Add 40µl of the resuspended beads to the reaction and mix by pipetting.
7. Incubate on a rotator mixer for 5 minutes.
8. Spin down the solution and pellet on a magnet. Once it is clear and colourless aspirate off supernatant.
9. Keep on the magnet, wash beads with 200µl of 70% ethanol without disturbing the pellet.
10. Remove the 70% ethanol using a pipette and discard. Repeat.
11. Spin down and place the tube back on the magnet. Pipette off any residual 70% ethanol. Briefly allow to dry.
12. Remove the tube from the magnetic rack and resuspend pellet in 25µl nuclease-free water. Incubate for 2 minutes.
13. Pellet beads on magnet until the elute is clear and colourless.
14. Remove and keep 25µl of elute into a clean 1.5ml Eppendorf DNA LoBind tube labeled with your group number.

15. Quantify 1µl of end-prepped DNA using a Qubit fluorimeter.
16. Dilute the library to a concentration of 10ng/µl with nuclease-free water.

Barcoding PCR

1. In a fresh 1.5ml Eppendorf tube labeled with your group number, set up a barcoding PCR reaction as follows for each library:

Reagent	Volume
PCR Barcode (one of BC1-BC12)	2µl
10ng/µl adapter ligated template	2µl
LongAmp Taq 2x master mix	50µl
Nuclease-free water	46µl
Total	100µl

2. Mix gently by inversion and spin down.
3. Amplify using the following cycling conditions:

Thermal cycler settings	
Total Volume:	100 µl
Parameters:	95°C for 3 minutes 15 cycles of: 95°C for 15 sec 62°C for 15 sec 65°C for 24 sec 65°C for 10 minutes Hold at 4°C

4. Quantify the barcoded libraries using standard techniques, and pool all barcoded libraries in the desired ratios in a 1.5mL DNA LoBind Eppendorf tube.
5. Prepare 1µg of pooled barcoded libraries in 45µl nuclease-free water.

End-prep

1. In a 1.5ml Eppendorf DNA LoBind tube, set up the end-prep reaction using NEBNext Ultra II End Repair/dA-tailing module and 1µg pooled DNA in 45µl as follows:

Reagent	Volume
DNA	45µl
Ultra II End-Prep buffer	7µl
Ultra II End-Prep enzyme mix	3µl
DNA CS (control strand)	5µl
Total	60µl

- Mix gently by inversion and spin down.
- Using a Thermal cycler incubate at 20°C for 5 minutes and at 65°C for 5 mins.
- Resuspend the AMPure XP beads by vortexing.
- Add 60 µl of the resuspended beads to the End-Prep reaction and mix by pipetting.
- Allow the DNA to bind to the beads by rotating for 5 minutes on a Rotating wheel.
- Spin down the solution and pellet on a magnet. Once it is clear and colourless aspirate off supernatant.
- Keep on the magnet, wash beads with 200µl of freshly prepared 70% Ethanol without disturbing the pellet. Remove the 70% ethanol using pipette and discard. Repeat.
- Briefly spin the tube to collect residual liquid at the bottom of the tube. Return the tube to the magnet and aspirate residual wash solution. Briefly allow to air dry.
- Remove the tube from the magnetic rack and resuspend the pelleted beads in 31µl nuclease-free water, then incubate for 2 minutes at room temperature.
- Place the tubes on a magnet to pellet the beads, once it is clear and colourless remove and keep 31µl of the eluate into a clean 1.5ml Eppendorf DNA LoBind tube.
- Quantify 1µl of eluted sample using the Qubit Fluorometer. There should be more than 700ng of material.

Adaptor Ligation

- Ensure that the Blunt/TA master mix is mixed thoroughly before use.
- In a DNA LoBind 1.5ml tube add the following, mix by inversion after each:

Reagent	Volume
Nuclease free water	8µl
End-Prepped DNA	30µl
Adapter Mix	10µl

HP Adapter	2µl
Blunt/TA Ligase Master Mix	50µl
Total	100µl

- Briefly spin down in a microfuge.
- Incubate for 10mins at room temperature.
- Add 1µl HP tether, mix by inversion, briefly spin down in a microfuge and incubate for 10 mins at room temperature.

MyOne C1 bead preparation

- Resuspend MyOne C1 beads by vortexing until homogenous.
- Take 50µl of resuspended MyOne C1-beads and transfer to a clean 1.5ml DNA LoBind tube. Pellet the beads on a magnet and aspirate off and discard the supernatant.
- Add 100µl Bead Binding buffer to the pelleted beads. Resuspend beads by vortexing. Place the tube on a magnet, allow beads to pellet and aspirate off and discard supernatant.
- Repeat the wash in step 3.
- Add 100µl Bead Binding buffer to the pelleted washed beads. Resuspend beads by vortexing. These are the 'Washed beads' required for the subsequent purification.

Library Purification

- To the adapter-ligated, tether-bound DNA add 100µl 'washed beads', carefully mix by pipetting and incubate at room temperature for 5 mins on a rotating wheel.
- Place the tube on a magnetic rack; allow beads to pellet and pipette off the supernatant.
- Resuspend the pelleted beads in 150µl Bead Binding Buffer. Place the tube on a magnet, allow beads to pellet. Aspirate off the Bead Binding Buffer using a pipette and discard.
- Repeat step 3.

Elution of library from MyOne C1-beads

- Resuspend the pelleted beads in 25µl of Elution Buffer by pipetting up and down. Incubate for 10 minutes at 37°C.

3. Pellet the beads on a magnet until the eluate is clear and colourless.
4. Remove and retain 25µl of eluate which contains the library into a clean 1.5ml Eppendorf DNA LoBind tube.
5. Place the tube of library on ice until required for library loading.
6. Quantify 1µl of eluted sample using the Qubit fluorometer. One should expect to retain more than 100ng of material.

Priming the Flow Cell

IMPORTANT: Thoroughly mix the contents of the RBF tube by inversion or pipetting, and spin down briefly. The library is loaded dropwise without putting the pipette tip firmly into the port.

Take care to avoid introducing air during pipetting.

1. Flip back the MinION lid and slide the sample port cover clockwise to that the sample port is visible.



2. Ensure that the sample port cover is fully opened (a 90 ° clockwise turn).
3. Check for small bubbles under the cover. Draw back a small volume to remove any bubbles. Check that there is continuous buffer from the priming port, across the sensor array to the outlet channel of the flow cell.



4. Prepare the flow cell priming mix in a clean 1.5 ml Eppendorf DNA LoBind tube.

Reagent	Volume
RBF	500µl
nuclease-free water	500µl
Total	1000µl

- Load 800µl of the priming mix into the flow cell via the priming port and wait 5 minutes, avoiding the introduction of air bubbles.
- Gently lift the activator to make the SpotON sample port accessible.
- Load 200µl of the priming mix into the Flow Cell via the priming port.

Loading a Library

- Prepare the library for loading as follows:

Reagent	Volume
RBF	37.5µl
LLB	25.5µl
Adapted and tethered library	12µl
Total	75µl

- Mix gently by pipetting just prior to loading.
- Load 75µl of the prepared library into the flow cell via the SpotON sample port in a dropwise fashion. Ensure each drop falls into the port before adding the next.
- Gently replace the activator, making sure the bung enters the SpotON port, close the sample port and replace the MinION lid.

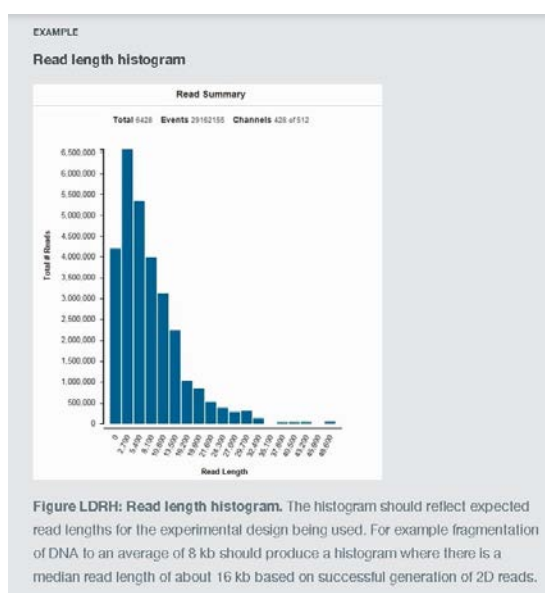
Starting a sequencing run

- Once a MinION and Flow Cell are connected, a Label Experiment dialogue box appears. Click into the Sample ID box and name your sample using free text in alphanumeric format only.
- Click into the FlowcellID box and enter the Flow Cell ID, which is the code found on a sticker on the top side of a Flow Cell.
- Select the appropriate protocol script.
 - Experiment type: **Choose Operation**
 - Flow Cell product code: Choose the Flow Cell type under **Flow cell product code**
 - Sequencing kit: Choose the sequencing kit version you have used to prepare the library
 - Whether or not live basecalling is enabled

- A drop-down menu will appear with the most relevant sequencing scripts. Select the script you need.
4. Start the script using the Execute button at the bottom of the Connections page.
 5. Select the appropriate protocol script using the Start Protocol dialogue box.
 - In the MinKNOW GUI click Start Protocol and the Run Protocol Script window opens
 - Click on the down arrow in the Select Protocol Script box to reveal the protocol script choice
 - Select the protocol script required

Progression of MinKNOW protocol script

1. Check that the number of active pores reported in the MUX scan are similar to those reported at the end of the Platform QC.
 - If there is a significant reduction in the numbers try to restart MinKNOW from the control panel.
 - If the numbers are still significantly different close down the host computer and reboot.
 - When the numbers are similar to those reported at the end of the Platform QC the experiment being carried out can be restarted on the Connection page; there is no need to load any additional library after restart.
2. Check the heatsink temperature is approximately 34 °C.
 - The MinION is able to maintain a heatsink temperature of 34°C on a typical lab bench when the local ambient conditions are between 19.5 °C and 24.5 °C. However, there are a number of external factors which can disrupt the local conditions and which need to be taken into account, for example warm air expelled from laptops, or cool air from a fan or air conditioning system increasing airflow around the MinION.
 - The MinION takes approximately 10 minutes to get to temperature.
3. Monitor the development of the read length histogram.



4. Check pore occupancy by looking at the panel at the top of the Status or Physical Layout views.

EXAMPLE

Channel Panel screen



Figure LDCP: Channel Panel screen at the start of a Lambda DNA experiment. A good library will be indicated by a high proportion of light and dark green channels. The combination of light and dark green indicate the number of active pores at any point in time and the dark green indicate the proportion of pores in strand (or sequencing) at a particular time point. A low proportion of dark green channels will reduce the throughput of the sequencing.

5. Ongoing monitoring of the experiment is best achieved using the VIEW REPORT in Desktop Agent.

It is Wednesday 25th January and the Illumina sequencing data from the run you prepared on Monday is ready for analysis. Isolates from nine cases of HUS from five different hospitals in the UK are ready for analysis.

Next generation sequencing is already having a profound impact on clinical microbiology with the potential to resolve pathogens to the resolution of a single strain. Assemblies of pathogen whole genomes allows for the genetic repertoire to be unveiled providing insights on pathogenicity, antibiotic resistance and other clinically important features. The genetic fingerprint derived from the whole genome sequence can also be used to compare strains to provide unparalleled resolution for elucidating outbreaks and transmission routes.

Data from Illumina machines (HiSeq/MiSeq) will usually be received as FASTQ files.

The first line contains the sequence identifiers:

- The second line contains the sequence of the read (ACTG or N for unresolved)

The final line is the quality score for each base. The quality is the probability that the corresponding base call is incorrect. It is referred to as a Phred quality score and is derived as follows:

$$Q_{\text{sanger}} = -10 \log_{10} p$$

The Phred score is encoded as an ASCII character by adding 64 to the Phred value.

Q=30 means probability base is incorrect is 1/1000

Q=20 means probability base is incorrect is 1/100

For paired end libraries (as prepared yesterday) you will receive two FASTQ files per sample one for each pair.

Exercise 1. Assessing the quality with FastQC

Load the FASTQs

- 1 Log on to the Public Health England Galaxy Server
- 2 Click on the 'Shared Data' Tab
- 3 Click on 'Data libraries drop down'
- 4 Click on the 'WTAC' & '2017' & 'MiSeq' Folder
- 5 Find your groups FASTQ files that you produced on Monday (if in the unlikely scenario your MiSeq run failed grab the two FASTQs STEC.R1.fastq.gz and STEC.R2.fastq.gz)
- 6 Click the check boxes next to them and select 'Import to current history'
- 7 Click on the 'Analyze Data'
- 8 Notice the FASTQ files in your history – Let's have a look at one.

Assess the quality with FASTQC

- 9 From the Tools on the left hand side click 'NGS: QC and manipulation'
- 10 Click 'FastQC:Read QC'
- 11 Select one of the FASTQ reads and Click 'Execute'
- 12 Note the job running in your History – this will take a couple of minutes.....
- 13 When it's finished (GREEN) click on the 'eye' icon to see the results

Q1. How many sequences in the file?

Q2. What is the %GC?

Q3. At what position does the average quality fall below Q30?

Explore the other statistics provided by FASTQC

It may be necessary to perform operations on the FASTQ files to exclude or correct low quality regions. This can involve trimming the end of reads where the quality is low (see FASTX toolkit) or using K-mer frequency distributions to correct or discard reads (see QUAKE/MUSKET/trimmomatic).

Remove bad quality reads with Trimmomatic

1. From the Tools on the left hand side click 'NGS: QC and manipulation'
2. Click Trimmomatic
3. Select the **R1.fastq** as Direction 1 fastq reads to trim
4. Select the **R2.fastq** as Direction 2 fastq reads to trim
5. Change the Quality encoding to **phred33**
6. Deselect Perform Sliding Window trimming
7. Change the minimum quality to trim leading bases to **30**
8. Change the minimum quality to trim trailing bases to **30**
9. Change the minimum read length to **50**
10. Click 'Execute'
11. Note the job running in your History – this will take a couple of minutes.....

Let's reassess the quality with FASTQC

- 1 From the Tools on the left hand side click 'NGS: QC and manipulation'
- 2 Click 'FastQC:Read QC'
- 3 Select one of the Trimmomatic paired output FASTQ reads and Click 'Execute'
- 4 Note the job running in your History – this will take a couple of minutes.....
- 5 When it's finished (GREEN) click on the 'eye' icon to see the results

Q1. How many sequences in the file?

Q2. What is the %GC?

Q3. At what position does the average quality fall below Q30?

Explore the other statistics provided by FASTQC – Our sequence data is much nicer now!

Analysing the outbreak isolates through the PHE *E. coli* pipeline

Now we have high quality data it's time to process the FASTQs as previously completed for the German data. This will involve as before; k-mer identification, MLST typing, *in silico* serotyping, pathotyping, Shiga toxin sub-typing and AMR typing. If you can't remember how to run these components refer to yesterday's section in the manual.

1. Based on these results do you think your isolate is likely to be the part of the same outbreak as the German isolate?
2. Based on these results do you think that all the isolates from the UK are likely to be part of the same outbreak?

Phylogenetic Analysis of Outbreak Data

Questions to answer during this section

1. How many SNPs different from each other are the UK isolates?
2. How many SNPs different are they from the German outbreak strain?
3. What can you conclude about (i) the links between all the UK cases and (ii) the relationship between the UK cases and the German outbreak?

Practical steps

1. We are going to assume that we have already carried out the mapping and SNP calling for all these isolates, so they are already in the database waiting to be queried.
2. On the menu bar go to 'shared data' -> Data libraries -> WTAC -> 2017-> Outbreak. From the outbreak folder add the 'strain_list_2' file to your current history Return to the 'Analyze Data' page.
3. On the tool menu, go to NGS: PHE internal tools and then 'get the snps'. In the main analysis window select 'strain_list_2' as the strain list, set 'O26_wtac_config.txt' as the config file, select alignment type as 'accessory'. Select 'No' to the question 'Would you like the reference genome in alignment' and 'Yes' to produce a distance matrix of SNP distances and list of annotated variants. Click execute.
4. After 1-2 minutes, the item in your history should turn from yellow to green. Click on the eye to inspect the results. This is your 'pseudo-sequence' of all the variant positions.

5. Download the pseudo-sequence and load it into MEGA. Follow the same steps as in the previous section to derive a tree.

***de novo* assembly and MinION data analysis**

1. *de novo* Assembly

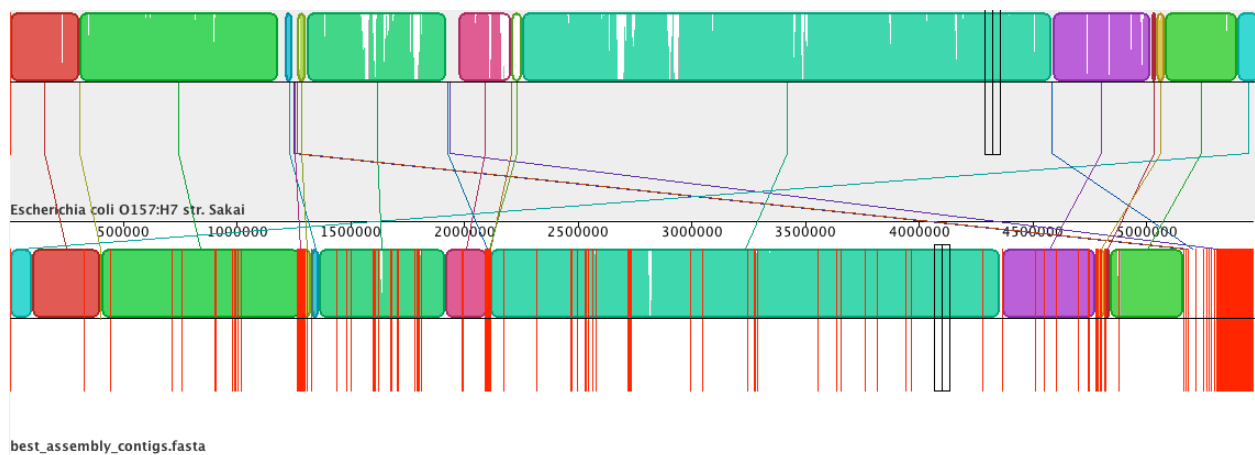
de novo assembly refers to the assembly of a genome with no reference to guide the process. A common analogy is putting a jigsaw of several hundreds thousands of pieces together without the picture to refer to.

Vocabulary:

- **Read** Any sequence that comes out of the sequencer
- **Paired read** read1, gap < 500 bp, read2
- **Mate-pair** read1, gap > 1 kbp, read2
- **k-mer** Any sequence of length k
- **Contig** gap-less assembled sequence
- **Scaffold** sequence which may contain gaps (N)

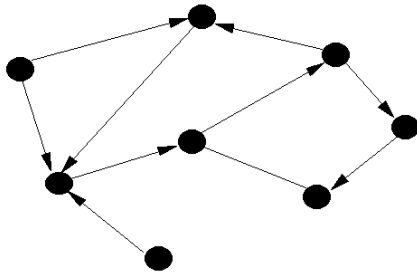
Example of a reference genome and an assembly aligned to it.

Nb. The assembly is fragmented into many contigs



Assembly Theory

de novo assembly uses graph theory. A graph is a set of nodes and a set of edges. The edges can have direction.



The graphs used to assemble short read data (illumina) uses de Bruijn graphs. Graphs allow the representation of overlaps between reads.

Example – de Bruijn graphs

Nodes represent all k-mers (k-length sub-strings) present in the reads

Edges represent the overlap between the k-mers observed.

E.G. 1 k=3 Single read

Read ACTG
Graph ACT → CTG

E.G. 2 k=3 3 reads

Read1 ACTG
Read2 CTGC
Read3 TGCT
Graph ACT → CTG → TGC → GCT

E.G. 3 k=3 4 reads – 1 has an error

Read1 ACTG
Read2 CTGC
Read3 CTGA
Read4 TGCT
Graph ACT → CTG → TGC → GCT
 ↓ TGA

E.G. 4 k=3 6 reads – What is the effect of repeats?

Read1	ACTG
Read2	CTGC
Read3	TGCT
Read4	GCTG
Read5	CTGA
Read6	TGAC
Graph	ACT→CTG→TGC ↓ ↓ ↓ GAC←TGA←GCT

How does one assemble using a graph?

To generate contigs from our graph we calculate all node-disjoint paths through the graph.

Node-disjoint means that two different path cannot share a node.

Example – Contig construction

Graph	ACT→CTG→TGC ↓ ↓ ↓ GAC←TGA←GCT	
Contig1	ACT→CTG→TGC ↓ ↓ ↓ GAC←TGA←GCT	CTGCT
Contig2	ACT→CTG→TGC ↓ ↓ ↓ GAC←TGA←GCT	TGAC
Contig3	ACT→CTG→TGC ↓ ↓ ↓ GAC←TGA←GCT	ACT

Choosing the k-mer value is important! Smaller k-mers increase sensitivity as more likely to observe an overlap, large k-mers increase accuracy.

Velvet (other *de novo* assemblers are available!)

Zerbino, D. R.; Birney, E. (2008). "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs". *Genome Research* 18 (5): 821–829.

Exercise 2. Assemble your genome.

Load the FASTQs

- 1 Click on the 'Shared Data' Tab
- 2 Click on 'Data libraries drop down'
- 3 Click on the 'WTAC'->'2017' & 'Assembly' Folder

- 4 There are two STEC FASTQs called “illumina” – click the check boxes next to them and select ‘Import to current history’

Run Velvet

- 1 From the Tools on the left hand side click ‘NGS: Assembly’
- 2 Click ‘Velvet Optimiser’
- 3 Set Start k-mer value as **45**
- 4 Set End k-mer value as **45**
- 5 Set k-mer search step size as **2**
- 6 Click Add new Input read libraries
- 7 Select File type: **shortPaired**
- 8 Click the tick box are the reads paired and in to different files
- 9 Select **_1.fastq** for Read dataset for direction 1:
- 10 Select **_2.fastq** for Read dataset for direction 2:
- 11 Click execute

How do we assess the assembly?

- Number of contigs/scaffolds
- Total length of the assembly
- Length of the largest contig/scaffold
- Percentage of gaps in scaffolds (‘N’)
- N50 of contigs/scaffolds
- Internal consistency
- Number of predicted genes

N50

N50 length is defined as the length N for which half of all bases in the sequences are in a sequence of length $L < N$
or

Half of the assembled bases reside in contig having a length of at least the n50 contig

Assess assembly

1. How many contigs were produced for each assembly?

2. What was the n50 reported at the end of the logfile
3. What is the size of the largest contig?
4. What is the total size of the assembly?

MinION read analysis

We are also going to compare Illumina and MinION reads that have been mapped to the same reference as the best way to understand MinION data. We will do this using the Tablet alignment viewer <https://ics.hutton.ac.uk/tablet/>.


Questions to answer during this section

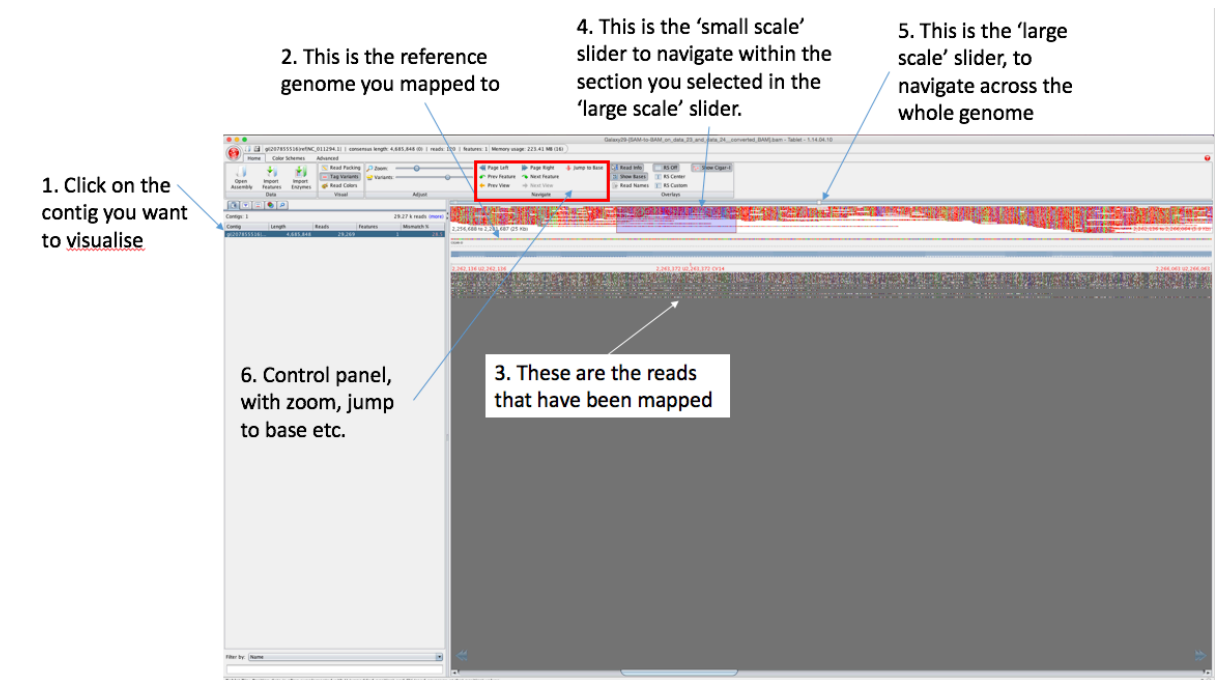
1. How many Illumina reads are mapped to the reference genome? How many MinION reads?
2. Find at least 5 SNPs in the Illumina mapping. Are these also SNPs in the MinION analysed sample? Do the two technologies agree?
3. Would you rather call SNPs directly from the MinION data, or just look for ones you have previously characterised from the Illumina data?

Practical steps

1. If you don't already have them, from Shared data -> WTAC -> 2017 -> Assembly add illumina.R2.fastq and illumina.R1.fastq to your current history.
2. From Shared data -> WTAC -> 2017 -> Assembly add stec_reference.fa to your current history
3. In the tool menu, go to NGS:Mapping -> map with BWA -> 'Will you select a reference genome from your history or use a built-in index?' - Use one from history -> 'Select a reference from history' - ensure you have stec_reference.fa selected -> 'Is this library mate-paired?' - paired-end -> 'forward fastq' - illumina .R1.fastq -> 'reverse fastq' illumina.R2.fastq -> Execute
4. After 2-3 minutes, your job will be finished. Then, we need to convert the output file (a SAM file) to a BAM file (remember - bioinformatics is basically just advanced file conversion) so that we can view it in Tablet. In the tools menu -> NGS: SAM Tools -> SAM-to-BAM -> 'Choose the source for the reference list' - History -> 'Convert SAM file' - the output of 'Map with BWA' -> 'Using reference file' - stec_reference.fa -> Execute
5. When the SAM -> BAM conversion has finished, click on the output -> then click on the floppy disk to download -> then download the dataset and the bam_index. An index is a 'helper file' that lets other

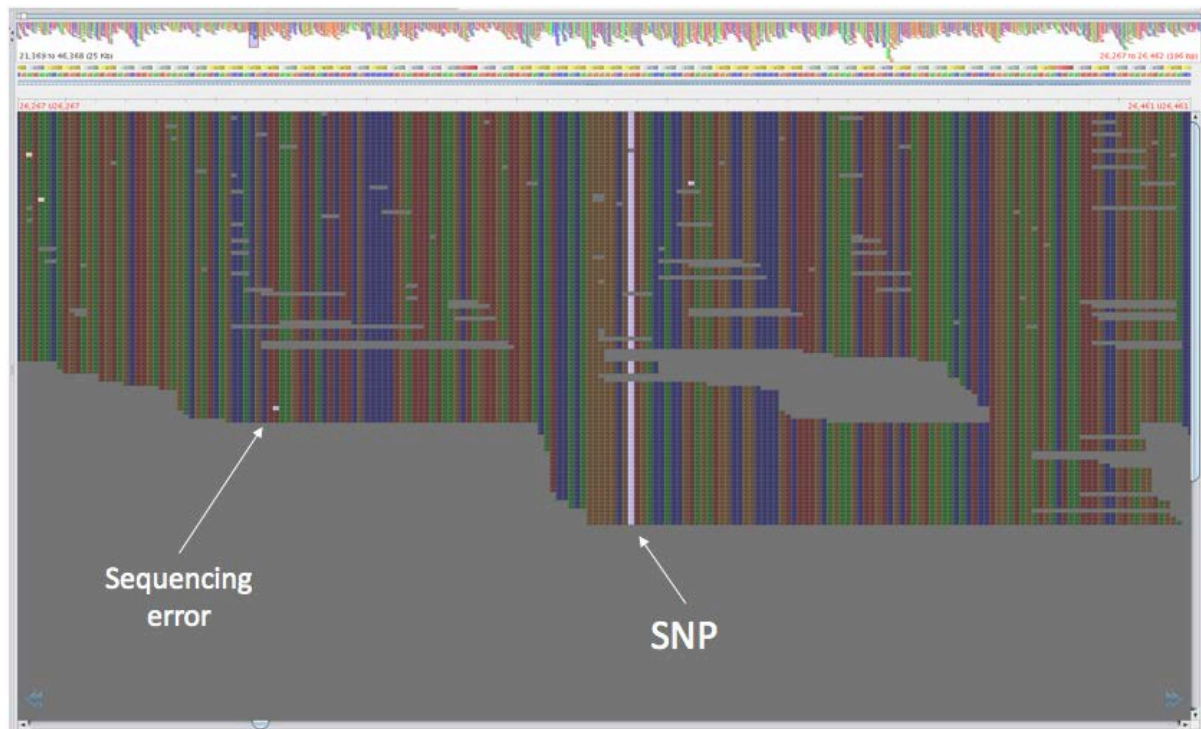
programs quickly access the main file. You also need to download the stec_reference.fa in the same way.

6. Then, open Tablet  via the Start menu. From the menu bar select 'open assembly' -> in 'Primary assembly' click 'Browse' then navigate to the bam file you just downloaded -> in 'Reference' click 'Browse' then navigate to the stec_reference.fa file you just downloaded -> Open
7. See below for a quick guide to the key parts of the Tablet software. This is one of my favourite pieces of bioinformatics software, I hope you like it too!



8. Scroll across the genome, looking for SNPs. Try to find at least 5 SNPs, and not the same ones as the person sat next to you. Mapped bases that are 'lighter' are different from the reference genome.

9. See below for an example.



10. Go back to Galaxy, from Shared data -> WTAC -> 2017 -> Assembly add minion.fasta to your current history.
11. From the tool menu -> NGS: PHE internal tools -> mapwithlast -> in 'Read data from your current history' select minion.fasta -> in 'Reference genome from your history' select stec_reference.fa -> Execute.
12. After 5 minutes, the alignment should be finished, we need to convert to BAM again. In the tools menu -> NGS: SAM Tools -> SAM-to-BAM -> 'Choose the source for the reference list' – History -> 'Convert SAM file' – the output of 'Map with BWA' -> 'Using reference file' - stec_reference.fa -> Execute
13. Once the job has finished, download the BAM and bam_index and open in Tablet.
14. Take the variant positions you identified in step 8 and navigate to the same locations in this bamfile using Tablet's 'jump to base' function. Does the MinION data agree with the Illumina data?

MinION assembly analysis

We are also going to compare Illumina and MinION assemblies to see if the longer reads help us understand the accessory genome. To do this we are going to use the program

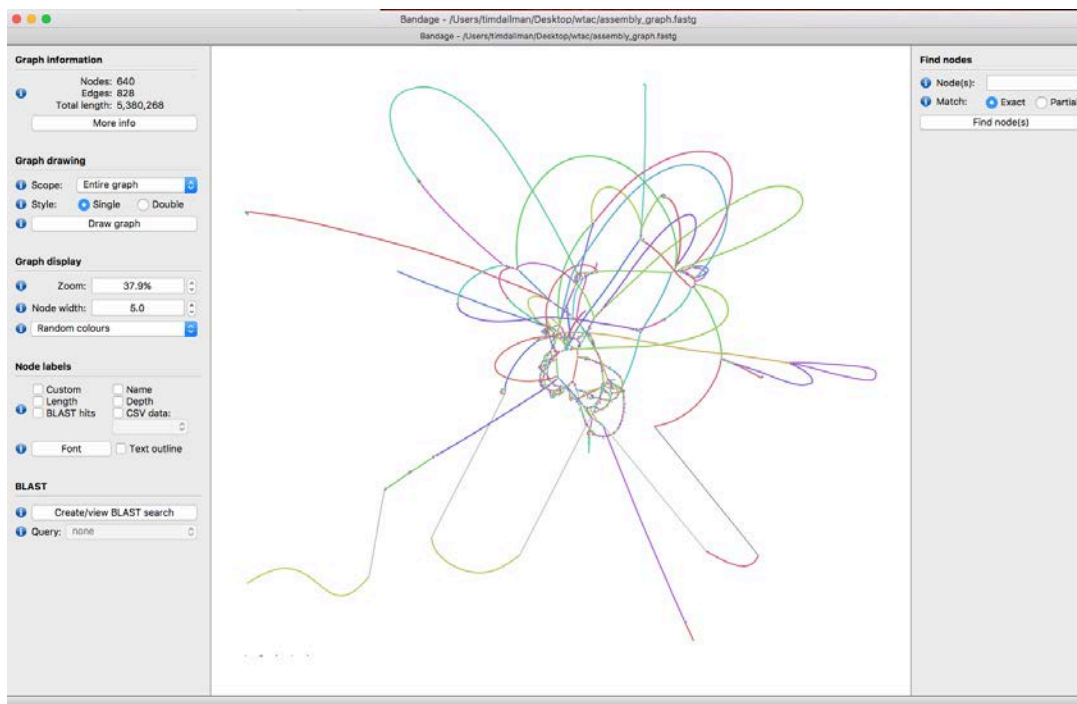
Bandage <http://rrwick.github.io/Bandage/>. Bandage is a program for visualising *de novo* assembly graphs. By displaying connections which are not present in the contigs file, Bandage opens up new possibilities for analysing *de novo* assemblies.

Questions to answer during this section

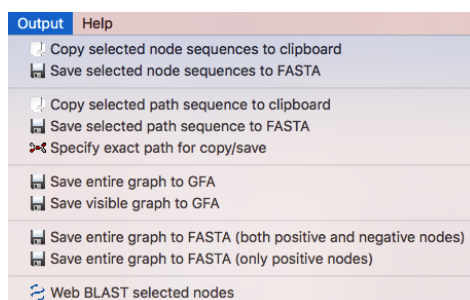
1. How many contigs (nodes) are there in the illumine assembly compared to the minion assembly?
2. What is the longest node and the n50 from each assembly method?
3. How big is the bacteriophage containing the Stx toxin?
4. How many plasmids does the isolate have?
5. Are the AMR elements on a plasmid or the chromosome?
6. Does the accessory genome analysis tell us anymore about the possible origin of the strain?

Practical Steps

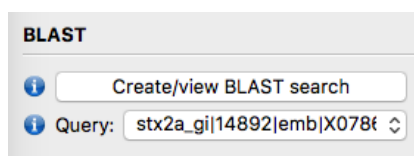
1. Add the illumina.fastg and minion.fastg assembly graphs and the blast database blast.fa from Shared data -> WTAC -> 2017 -> Assembly to your current history.
2. Click back to Analyze data and as previous click on the floppy disk icon to download the files to your Desktop.
3. Open Bandage and load the Illumina assembly graph by clicking 'Load Graph' from the File dropdown.
4. Click on the 'More Info' button to get some information about the assembly.
5. Let's take a look at the assembly graph – by clicking the 'Draw Graph' button – see below for an example. Zoom in examine the connectivity of the contigs.



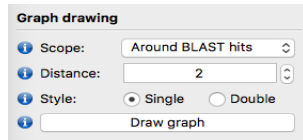
6. You can select nodes and save the underlying sequence to the clipboard or to a file, we can then BLAST this sequence against the NCBI database by clicking the 'Output' tab and clicking 'Web BLAST selected nodes'.



7. We can also provide our own BLAST database to search for features in the assembly. Click on the 'Create/view BLAST search button' to bring up a new window, as below. Click 'Build BLAST database' first. Then click 'from FASTA file' and select blast.fa that you downloaded to your desktop and you will see the list of AMR and virulence genes loaded. Finally click the 'Run BLAST search button' and the BLAST results will appear.
8. Let's see where some of these genes are in our assembly. Back on the main viewer on the side menu select the 'stx2a' gene in the BLAST query drop down menu.



9. Under the Graph drawing section change the Scope: to 'Around BLAST hits' and the Distance to 2. Click the 'Draw graph' button.



10. The node with the BLAST match is coloured blue. How big is it? Let's BLAST the node and surrounding ones against the NCBI database? What are the matches?
11. Let's look at the other BLAST matches – can you spot any AMR genes?
12. Now let's repeat with the minion data and load the minion.fastg assembly graph.
13. How has the assembly changes with the different sequence data? What is the biggest contig now?
14. Let's repeat the BLAST analysis. How big is the contig with the *stx2a* gene now? Can you spot any plasmids?
15. BLAST a contig containing an AMR gene against NCBI?

Appendix

Protocol for Real-time multiplex PCR assay for the detection of *VTEC*, *Shigella* and *Campylobacter*

PRINCIPLES OF THE RT-PCR PROCEDURE

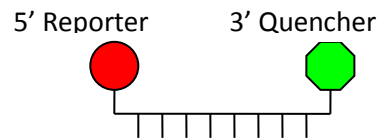
This procedure is for the molecular detection of *verocytotoxin-producing E. coli*, *Shigella* and *Campylobacter* using a multiplex real-time PCR assay. Each reaction contains an amplification internal control (IC). The purpose of this control is to identify potential inhibition from the processed specimen.

Simultaneous amplification and detection of the different targets in the multiplex PCR assay is achieved through the use of target specific primers and dual-labelled probes. Dual-labelled probes, also known as Taqman™ probes, are oligonucleotides designed to the internal region of a target. They are labelled with two different dyes, a reporter dye at the 5' end and a quencher dye at the 3' end. When excited the reporter dye emits fluorescence that is quenched by the quencher dye by Förster-type energy transfer (FRET), resulting in no fluorescence. However, following hybridisation of the probe to target DNA during PCR, the 5'-3' exonuclease activity of *Taq* polymerase cleaves the hybridised probe. This separates the reporter dye from the quencher dye, preventing FRET and resulting in the emission of fluorescence (Figure 5.1). Fluorescence increases in proportion to the rate of probe cleavage during each PCR cycle, which is displayed on the real-time PCR instrument (Figure 5.2). In order to compare and interpret the fluorescent data a threshold is set. This enables a cycle threshold (ct) value to be assigned to each sample, which is the point at which the fluorescence of the sample crosses the threshold. A sample with no ct value is considered negative and a sample with a ct value is considered positive. The probes for the different targets in this assay are labelled with different reporter dyes so that they can be detected on different channels of the real-time PCR instrument (see table 5.1). There are a number of different real-time PCR instruments (platforms) on the market including Smartcyclers, Lightcyclers, Taqmans. Each instrument has different advantages and disadvantages. For this assay a RotorGene is used.

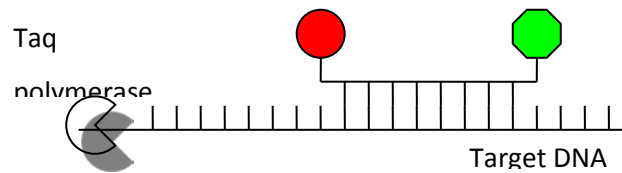
Table 5.1: Labels used for the multiplex probes and the channels they are detected in.

Organism	Label	Max Em (nm)	Max Abs (nm)	Channel
Campylobacter	Fam	520	494	Green
VTEC	Yak	549	530	Yellow
Shigella	Cy5	662	646	Red
GFP	Rox	605	575	Orange

1. Dual-labelled probe – no fluorescence due to FRET



2. Dual-labelled probe binds to target DNA



3. Taq polymerase cleaves the probe – fluorescence, no FRET

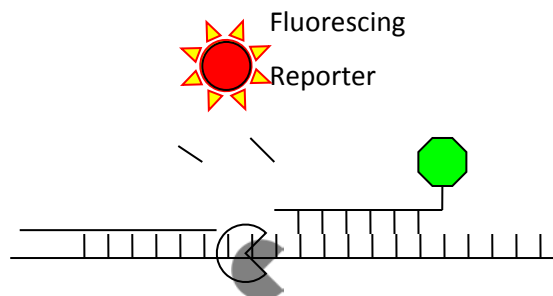


Figure 5.1 Diagram of a dual-labelled probe

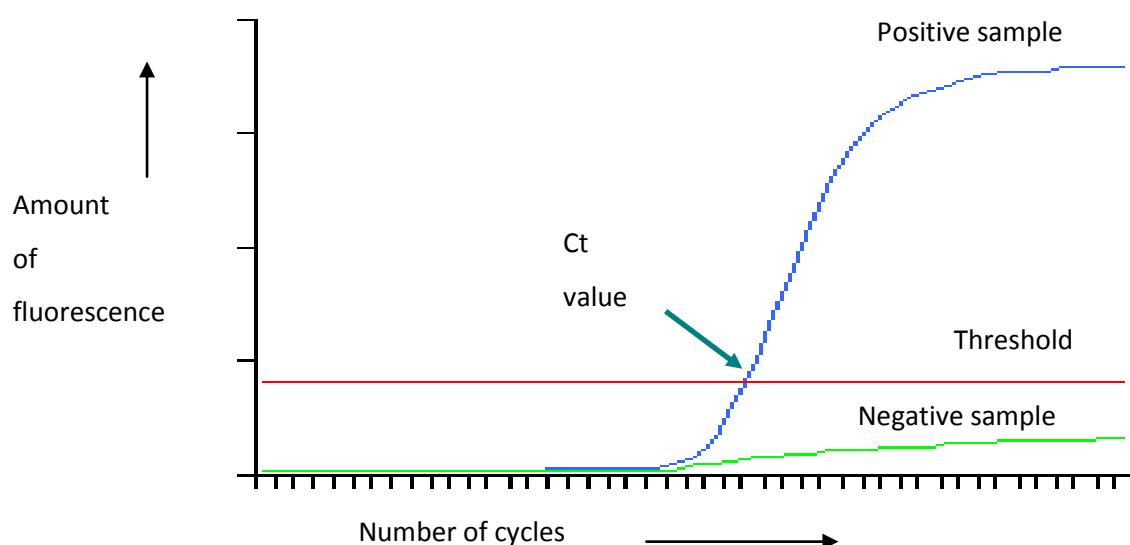


Figure 5.2 Diagram illustrating RT-PCR data

1. The Y-axis is the amount of fluorescence detected, the X-axis is the number of PCR cycles and the ct (cycle threshold) value is the point at which the positive sample's (in blue) fluorescence crosses the set threshold (straight line in red). The negative sample's (green) fluorescence does not cross the threshold and therefore does not have a ct value.

Real-time probe based PCR assays can be performed using exactly the same reagents used for conventional PCR assays, however there are a number of companies that produce specifically designed mixes, referred to as quantitative PCR (QPCR) mixes. These are most commonly supplied as two-times concentrates that solely require the addition of the primers and probe, thus reducing pipetting errors. The Invitrogen Platinum® QPCR SuperMix-UDG is used for this assay. It contains dUTPs instead of dTTPs and uracil DNA Glycosylase (UDG). These components reduce contamination by amplicon cross over. If amplicons generated with dUTPs are carried over into a new reaction the UDG is able to render the amplicon unamplifiable by cleaving the N-glycosidic bonds between the uracil bases and the phosphodiester backbone (Longo *et al*, 1990).

Enteric pathogens multiplex PCR assay protocol

Equipment:

- Gilson pipettes (P1000, P200, P20, P10)
- Plugged tips (1 ml, 200 µl, 20 µl & 10 µl)
- 0.2 ml flat cap thin walled PCR tubes
- 1.5 ml Eppendorf tubes
- Tube racks
- 4°C refrigerator
- 20°C freezer
- Disposable gloves
- Rotor-Gene (or other Real-time PCR instrument)
- Discard jar

Reagents:

- DNA extracts (samples and controls)
- Invitrogen Platinum[®] QPCR SuperMix-UDG (-20°C)
- PCR grade sterile distilled water
- Primers
- Probes

Procedure:

1. Remove a tube of Invitrogen Platinum QPCR mix and probes from the freezer to defrost. (The probes should be in an amber tube or kept in the dark as they are light sensitive).
2. Determine the number (N) of PCR reactions as below:
N = the number of samples + 4 controls (1 negative control, 1 *Campylobacter* positive control, 1 *Salmonella* positive control, 1 *Shigella* positive control) + 2 extra (to allow for mix loss during pipetting).
Each group will test two samples, therefore N = 8 (enough for 6 reaction tubes)
3. Place 6 x 0.2 ml flat capped tubes into a tray and label them appropriately as below including your group number:

S 1 = Unknown sample

S2 = Unknown sample

NEG = Negative control

CAM = *Campylobacter species* positive control

SAL = *Salmonella species* positive control

SHI = *Shigella species* positive control

In a sterile 1.5 ml eppendorf prepare the mastermix:

	x 1	x 8
Invitrogen Plat Supermix	12.5 µl	
MgCl ₂	1.5 µl	
Primer mix (2.5µM)	2.5 µl	
Probe mix (2 µM)	2.5 µl	
IC	1 µl	

Mix the mastermixes by inverting 10 times

Pipette 20 µl of mastermix into the 0.2 ml tubes.

Add 5 µl of the sample / controls to the appropriately labelled reaction tubes.

Place the tubes into the Rotor-Gene. Ensure that the tubes are at an angle so they flat in the rota, secure them with the metal ring that screws on top of them and then close the RotorGene.

Then set up the RotorGene:

- Select new run
- Select the 'Capsular screen' program.
- Select the 36 well run and confirm there are no 0.2 ml domed cap tubes in the rotor.
- Ensure the reaction volume is set to 25 µl
 - Go to edit profile to check the parameters of the program are correct:

Hold: 95°C 3 minutes

Cycling: 95°C 15 seconds

60°C 45 seconds (acquiring on green, yellow, red, and orange channels)

- Click on 'Start run'.
- Go to the sample option on the main menu and enter the appropriate reaction tube details for each position on the rota.

The RotorGene will then display the program details, the stage it is at and the amount of fluorescence obtained for each sample.

When the run is complete remove the tubes from the RotorGene and discard them in to a discard jar.

Then select 'Analysis' from the main menu, followed by the 'Quantification' option and then click on show. Set the threshold to 0.05. Repeat this process for each channel.

Record which samples have ct values and for which channels (it is useful to look at the raw data as well as the analysed data to ensure amplification has really occurred). Determine if the assay has worked (all controls correct) and if the sample was positive for any of the targets.

	Campy PCR	Salmonella PCR	Shigella PCR	IC	Result
S1					
S2					
NEG					
CAM					
SAL					
SHI					

Chapter 6 Encapsulated Bacteria

6.1 Introduction

Haemophilus influenzae, *Streptococcus pneumoniae* and *Neisseria meningitidis*, are very different bacteria which nevertheless cause similar invasive bacterial diseases, especially meningitis and septicaemia. Non-culture diagnosis is important for disease caused by these organisms, as most patients are likely to have been treated with broad-spectrum antimicrobials prior to a clinical specimen being taken. For example, in England and Wales, around half of laboratory-confirmed diagnoses of meningococcal septicaemia and meningitis are made with non-culture, molecular techniques.

In addition to causing similar disease pathologies, all three bacteria are normally members of the commensal microbiota of the human nasopharynx and can be thought of as 'accidental' pathogens in that they gain no advantage in causing disease. Pathogenic variants of these bacteria are normally encapsulated with polysaccharides which are encoded by a capsule region of the chromosome. Isolates of these organisms can only express one capsule, although each species has a repertoire of different capsules (over 90 in the case of the pneumococcus). They also are competent for DNA transformation which results in an essentially non-clonal population structure and the ability to acquire different capsules and antimicrobial resistances by lateral gene transfer.

There are effective polysaccharide-conjugate vaccines against *H. influenzae* type b polysaccharide (Hib), various combinations of pneumococcal capsules (also called serotypes), and some but not all meningococcal capsules (confusingly referred to as serogroups!), therefore determination of the capsule expressed by an invasive encapsulated bacteria is important. Multilocus sequence typing (MLST) schemes have been developed for all of these organisms and, while the most Hib isolates belong to one genetic group, both pneumococcal and meningococcal populations comprise numerous different genotypes, recognized as clonal complexes by MLST. These clonal complexes vary in their pathogenicity, antimicrobial resistance, antigenicity and likelihood of being involved in epidemic outbreaks, so knowledge the clonal complex of a particular isolate is important. As there is appreciable disease caused by a limited number of clonal complexes, epidemiological resolution of a particular disease outbreak may require characterization at additional loci, up to whole genome analysis.

This chapter contains methods usable in the clinical laboratory for the identification and characterization of these organisms at the levels of diagnosis (6.2), genetic characterization of pneumococcal (6.3) and meningococcal capsular (6.4) operons. The assembly of whole genome sequences is described (6.5) along with the use of the assembled data for single locus (6.6) and whole genome analysis.

Reference

Harrison OB, Brueggemann AB, Caugant DA, van der Ende A, Frosch M, Gray S, Heuberger S, Krizova P, Olcen P, Slack M, Taha MK, Maiden MC. (2011). Molecular typing methods for outbreak detection and surveillance of invasive disease caused by *Neisseria meningitidis*, *Haemophilus influenzae* and *Streptococcus pneumoniae*, a review. *Microbiology*, 157:2181-95.

6.2 Molecular diagnosis of meningitis

Protocol for Real-time multiplex PCR assay for the detection of *Streptococcus pneumoniae*, *Haemophilus influenzae* and *Neisseria meningitidis*.

PRINCIPLES OF THE PROCEDURE

This procedure is for the molecular detection of *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Neisseria meningitidis* using a multiplex real-time PCR assay. Each reaction contains an amplification internal control (IC). The purpose of this control is to identify potential inhibition from the processed specimen.

Simultaneous amplification and detection of the different targets in the multiplex PCR assay is achieved through the use of target specific primers and dual-labelled probes. Dual-labelled probes, also known as TaqMan™ probes, are oligonucleotides designed to the internal region of a target. They are labelled with two different dyes, a reporter dye at the 5' end and a quencher dye at the 3' end. When excited the reporter dye emits fluorescence that is quenched by the quencher dye by Förster-type energy transfer (FRET), resulting in no fluorescence. However, following hybridisation of the probe to target DNA during PCR, the 5'-3' exonuclease activity of *Taq* polymerase cleaves the hybridised probe. This separates the reporter dye from the quencher dye, preventing FRET and resulting in the emission of fluorescence (Figure 1). Fluorescence increases in proportion to the rate of probe cleavage during each PCR cycle, which is displayed on the real-time PCR instrument (Figure 2). In order to compare and interpret the fluorescent data a threshold is set. This enables a cycle threshold (Ct) value to be assigned to each sample, which is the point at which the fluorescence of the sample crosses the threshold. A sample with no Ct value is considered negative and a sample with a Ct value is considered positive. The probes for the different targets in this assay are labelled with different reporter dyes so that they can be detected on different channels of the real-time PCR instrument (see Table 1). There are a number of different real-time PCR instruments (platforms) on the market including: Smartcycler; Lightcycler; TaqMan; AB Vii7 etc. Each instrument has different advantages and disadvantages. For this assay a Qiagen Rotor-Gene is used.

Table 1: Labels used for the multiplex probes and the channels they are detected in.

Organism	Label	Max Emission (nm)	Max Absorbance (nm)	Channel
<i>H. influenzae</i>	FAM	520	494	Green
<i>N. meningitidis</i>	JOE	549	530	Yellow
<i>S. pneumoniae</i>	ROX	605	575	Orange
Inhibition control	Cy5	662	646	Red

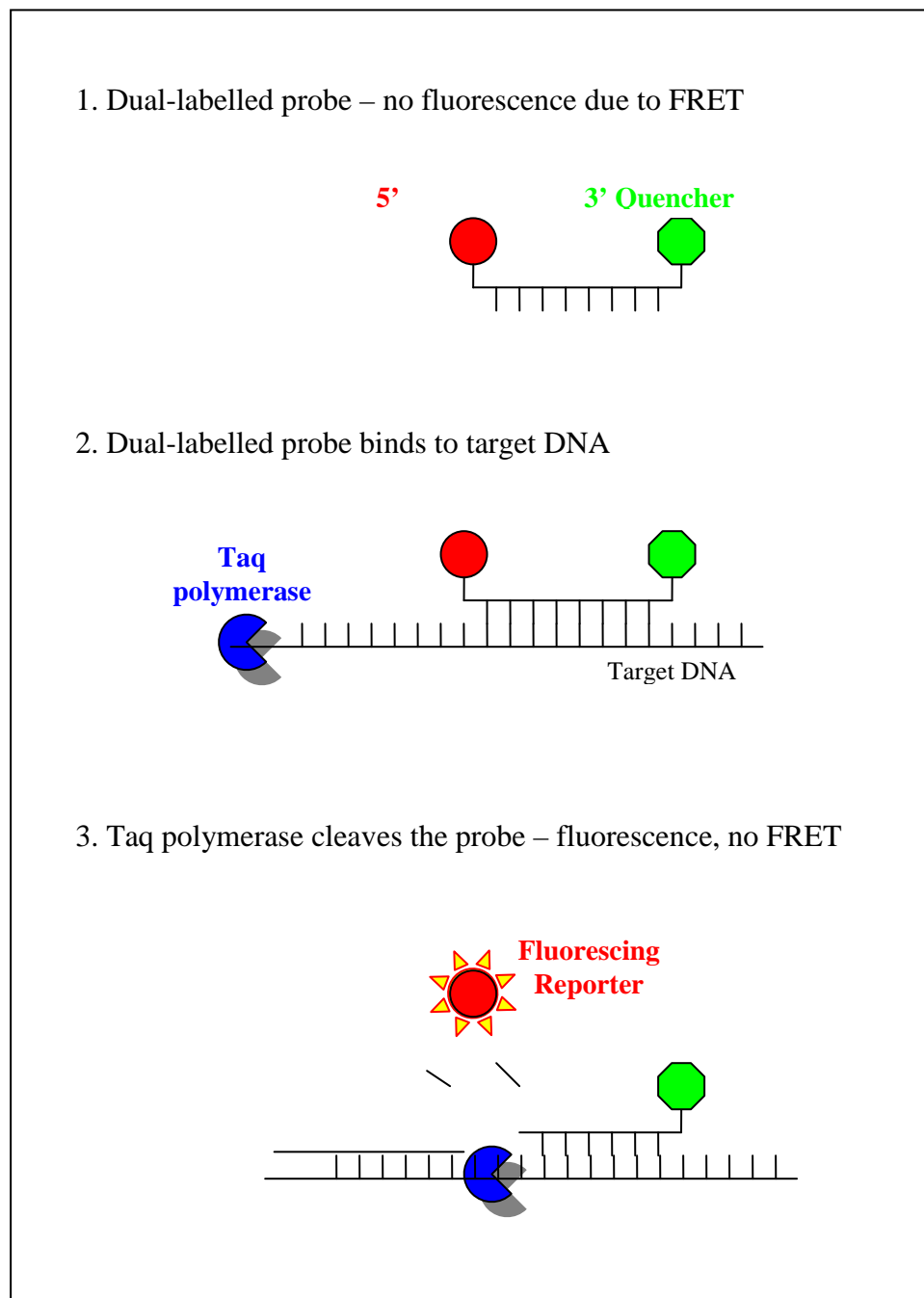


Figure 1: Diagram of a dual-labelled probe

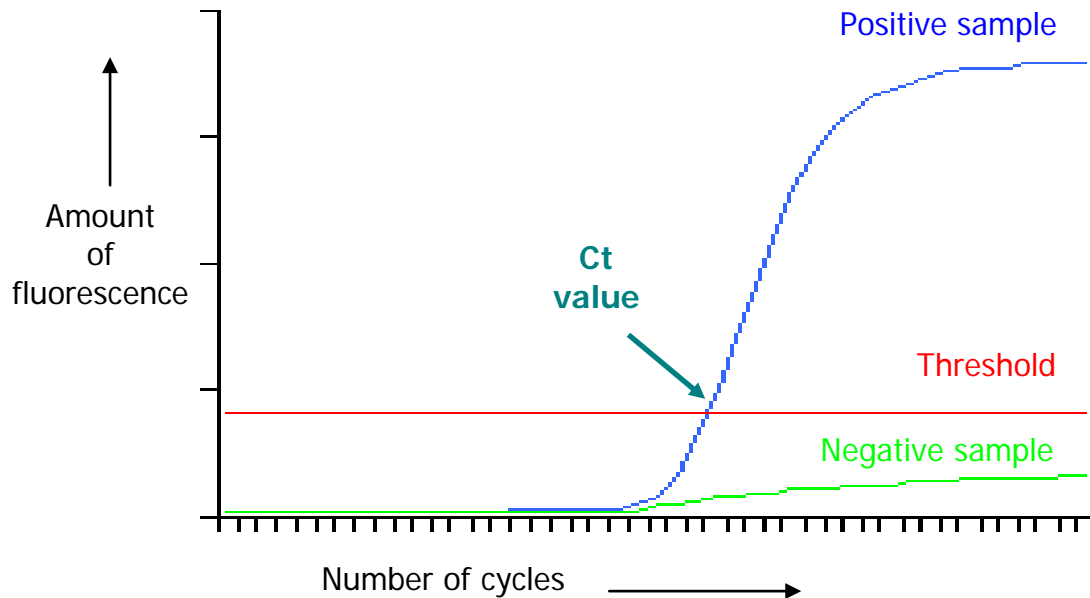


Figure 2: Diagram illustrating Real-time PCR data

The Y-axis is the amount of fluorescence detected, the X-axis is the number of PCR cycles and the Ct (cycle threshold) value is the point at which the positive sample's (in blue) fluorescence crosses the set threshold (straight line in red). The negative sample's (green) fluorescence does not cross the threshold and therefore does not have a Ct value.

Real-time probe based PCR assays can be performed using exactly the same reagents used for conventional PCR assays, however there are a number of companies that produce specifically designed mixes, referred to as quantitative PCR (QPCR) mixes. These are most commonly supplied as two-times concentrates that solely require the addition of the primers and probe, thus reducing pipetting errors. The Invitrogen Platinum® QPCR SuperMix-UDG is used for this assay. It contains dUTPs instead of dTTPs and uracil DNA Glycosylase (UDG). These components reduce contamination by amplicon cross over. If amplicons generated with dUTPs are carried over into a new reaction the UDG is able to render the amplicon unamplifiable by cleaving the N-glycosidic bonds between the uracil bases and the phosphodiester backbone [1].

Capsular pathogens multiplex PCR assay protocol

Equipment:

- Gilson pipettes (P1000, P200, P20, P10)
- Plugged tips (1 ml, 200 µl, 20 µl & 10 µl)
- 0.1 ml strip cap PCR tubes
- 1.5 ml Eppendorf tubes
- Tube racks
- 4°C refrigerator
- 20°C freezer
- Disposable gloves
- Rotor-Gene (or other Real-time PCR instrument)
- Discard jar

Reagents:

- DNA extracts (samples and controls)
- Invitrogen Platinum® QPCR SuperMix-UDG (-20°C)
- PCR grade sterile distilled water
- Primer mix (contains forward and reverse primers for all four targets)
- Probe mix (contains labelled probes for all four targets).

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all times during the practical

Procedure:

1. Remove a tube of Invitrogen Platinum QPCR mix and probes from the freezer to defrost. (The probes should be in an amber tube or kept in the dark e.g. using foil as they are light sensitive).
2. Determine the number (N) of PCR reactions as below:
NB: In practice, in real-time PCR all samples and controls are tested at least in duplicate, but for the purposes of space and time on the machine for the practical we will test just one of each.

N = the number of samples + 4 controls (1 non-template control (NTC) i.e. just mastermix, 1 *H. influenzae* positive control, 1 *N. meningitidis* positive control, 1 *S. pneumoniae* positive control) + 2 extra (to allow for mix loss during pipetting).
Therefore N = 10 (enough for 8 reaction tubes)

3. Place 0.1 ml strip capped tubes into a tray and label the caps appropriately as below including your group number:

NTC	= Non-template control
HI	= <i>H. influenzae</i> positive control
NM	= <i>N. meningitidis</i> positive control
SP	= <i>S. pneumoniae</i> positive control
S1	= Sample 1
S2	= Sample 2... etc

4. In a sterile 1.5 ml Eppendorf tube prepare the mastermix:

	x 1	x n
Invitrogen Plat Supermix	12.5 µl	
MgCl ₂	1.5 µl	
Primer mix (2.5µM)	2.5 µl	
Probe mix (2 µM)	2.5 µl	
IC	1 µl	

5. Mix the mastermixes by inverting 10 times
6. Pipette 20 µl of mastermix into the 0.1 ml tubes.
7. Add 5 µl of the sample/controls to the appropriately labelled reaction tubes.
8. Place the tubes into the Rotor-Gene. Ensure that the tubes are at an angle so they flat in the rotor, secure them with the metal ring that screws on top of them and then close the Rotor-Gene.
9. Then set up the Rotor-Gene:
 - Select new run
 - Select the 'Capsular screen' program.
 - Select the 72 well run and confirm there are no 0.1 ml strip cap tubes in the rotor.
 - Ensure the reaction volume is set to 25 µl
 - Go to edit profile to check the parameters of the program are correct:

Hold: 95°C 3 minutes

Cycling: 95°C 15 seconds

 60°C 45 seconds (acquiring on green, yellow, red, and orange channels)
 - Click on 'Start run'.
 - Go to the sample option on the main menu and enter the appropriate reaction tube details for each position on the rota.
10. The Rotor-Gene will then display the program details, the stage it is at and the amount of fluorescence obtained for each sample.
11. When the run is complete remove the tubes from the Rotor-Gene and discard them in to a discard jar.
12. Then select 'Analysis' from the main menu followed by the 'Quantification' option and then click on show. Set the threshold to 0.05. Repeat this process for each channel.
13. Record which samples have Ct values and for which channels (it is useful to look at the raw data as well as the analysed data to ensure amplification has really

occurred). Determine if the assay has worked (all controls correct) and if the sample was positive for any of the targets.

Results

Sample	Hi PCR	Sp PCR	Nm PCR	Result
Hi Pos				
Sp Pos				
Nm Pos				
NTC				
S1				
S2				
S3				
S4				

Interpretation

Ct values ≤ 35 are considered positive; Cts between 36 and 40 are equivocal; and Ct values >40 are negative. Equivocal results are usually retested by diluting the DNA either 1:4 or 1:10 to reduce any inhibitors that may be interfering with the polymerase. The amplification plots should be analyzed to ensure they are smooth and sigmoidal in shape. If plot is lacking these characteristics, it should be considered negative or retested.

Appendix

Organism	Name	Sequence (5' – 3')	5' label	3' label	bp	ref
<i>S. pneumoniae</i>	lytA-CDC-F	ACGCAATCTAGCAGATGAAGCA			22	[2]
	lytA-CDC-R	TCGTGCGTTTTAATTCCAGCT			21	[2]
	lytA-CDC-Probe	TGCCGAAAACGCTTGATACAGGGAG	ROX	BHQ2	25	[2]
<i>H. influenzae</i>	HelS-F	CCGGGTGCGGTAGAATTTAATAA			23	
	HelA-R	CTGATTTTTCAGTGCTGTCTTTGC			24	
	Hel-Probe	ACAGCCACAACGGTAAAGTGTCTACG	FAM	BHQ1	28	
<i>N. meningitidis</i>	ctrA-F	TGTGTTCCGCTATACGCCATT			21	[3]
	ctrA-R	GCCATATTCACACGATATACC			21	[3]
	ctr-Probe	AACCTTGAGCAATCCATTTATCCTGACGTTCT	JOE	BHQ1	32	[3]
gfp	gfp F	CCTGTCCTTTTACCAGACAACCA			23	[4]
	gfp R	GGTCTCTCTTTTCGTTGGGATCT			23	[4]
	gfp Probe	TACCTGTCCACACAATCTGCCCTTTCG	Cy5	BHQ3	27	[4]

Table A1: Primers and probes for Capsular pathogens multiplex PCR assay.

References:

1. Longo MC, Berninger MS, Hartley JL. 1990. Use of uracil DNA glycosylase to control carry-over contamination in polymerase chain reactions. *Gene* 93:125-8.
2. Carvalho Mda G, Tondella ML, McCaustland K, Weidlich L, McGee L, Mayer LW, Steigerwalt A, Whaley M, Facklam RR, Fields B, Carlone G, Ades EW, Dagan R, Sampson JS. 2007. Evaluation and improvement of real-time PCR assays targeting *lytA*, *ply*, and *psaA* genes for detection of pneumococcal DNA. *J Clin Microbiol* 45:2460-6.
3. Mothershed EA, Sacchi CT, Whitney AM, Barnett GA, Ajello GW, Schmink S, Mayer LW, Phelan M, Taylor TH, Jr., Bernhardt SA, Rosenstein NE, Popovic T. 2004. Use of real-time PCR to resolve slide agglutination discrepancies in serogroup identification of *Neisseria meningitidis*. *J. Clin. Microbiol.* 42:320-328.
4. Murphy NM(1), McLauchlin J, Ohai C, Grant KA. Construction and evaluation of a microbiological positive process internal control for PCR-based examination of food samples for *Listeria monocytogenes* and *Salmonella enterica*. *Int J Food Microbiol.* 2007 Nov 30;120(1-2):110-9.

6.3 Protocol for *Streptococcus pneumoniae* sequotyping

Serotyping of *Streptococcus pneumoniae* is used to monitor epidemiological trends and to monitor the strains causing invasive disease. Conventional serotyping by Quellung reaction is costly, labour-intensive and prone to misidentification. The sequotyping method described in this chapter is based on PCR targeting the capsular polysaccharide synthesis loci (*cps*) specific for *S. pneumoniae* and subsequent sequencing of the amplified product. This method is a rapid, accurate and robust alternative to the conventional serotyping.

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all times during the practical

PB Buffer – Harmful/Irritant/ Flammable

Ethanol – Highly flammable/Irritant

1. Genomic DNA extraction by heat lysis

(DNA extraction will have been done for you)

Principle

This is a simple technique for obtaining crude DNA lysates compatible with subsequent PCR analysis. The DNA is extracted from pure cultures of *Streptococcus pneumoniae*. The extraction is based on heat lysis in a buffer. Centrifugation at high speed separates cell debris and leaves the released DNA in the solution.

Procedure

1. Transfer 1 ml of an overnight *S. pneumoniae* culture in 1.5 ml Eppendorf tube.
2. Spin at 10,000 *g* for 10 min and discard the supernatant.
3. Resuspend the cell pellet in 100 µl of 1× TE buffer (10 mM Tris-HCl and 1 mM EDTA; pH 8.0).
4. Incubate the tubes containing cell suspensions at 95°C and 1,000 rpm for 20 min.
5. After incubation, spin the tubes at 13,000 *g* for 3 min.
6. The supernatant contains released DNA; 2 µl is used for the PCR reaction.
7. The supernatant can be transferred into a new tube and stored at 4°C or -20°C.

2. PCR amplification for sequotyping

Principle

A universal primer-binding region of the *S. pneumoniae cps* operon is amplified in a real-time SYBR Green-based PCR. The PCR product is identified as positive fluorescence signal obtained upon amplification of the target and is confirmed by the analysis of melting temperature.

Procedure

1. Label PCR tubes with your group number and the sample numbers. Each DNA sample will be analysed in duplicate.
2. Prepare the mastermix for all samples according to the table below.

Reagent	V (1 reaction)	V (n reactions)
EvaGreen Mix (2x; SsoFast, BioRad)	10 ml	
Primer cps1 (10 mM)	0.8 ml	
Primer cps2 (10 mM)	0.8 ml	
PCR water	6.4 ml	
Volume to aliquot	18 ml	

3. Aliquot 18 ml of the mastermix into PCR tubes.
4. Add 2 ml of sample to the tubes containing mastermix (use 2 ml nuclease-free water as negative control).
5. Place PCR tubes into the real-time PCR instrument (RotorGeneQ) and run the reaction with settings as follows:

Denaturation	95°C, 15 min
	95°C, 1 min
30 cycles	57°C, 1 min
	72°C, 1.5 min; fluorescence acquisition in green channel
Extension	72°C, 3 min
Melting	ramp from 60°C to 90°C; fluorescence read in green channel
6. Analyse the amplification and melting curves of the PCR products. A single distinctive melting peak should be observed for each sample.
7. Store the PCR products at 4°C (or -20°C for longer periods) or proceed immediately to clean-up and sequencing.

3. Clean-up of PCR products

Principle

The *S. pneumoniae* PCR products will be purified by the QIAquick PCR Purification Kit (Qiagen) in this practical. The kit employs spin-columns with silica membrane. DNA adsorbs to the silica membrane in the presence of high concentrations of salt. DNA fragments ranging from 100 bp to 10 kb are purified from primers, nucleotides, polymerases, and salts. DNA is then eluted to water or buffer.

Due to time constraints, this will be described and performed for you by the demonstrators. The full process is, however, detailed in the following section.

Procedure

1. Before starting the clean-up
 - a. Add ethanol (96–100%) to Buffer PE before use.

- b. Ensure the Buffer PB has the pH of ≤ 7.5 (the colour of the mixture is yellow). If the colour of the mixture is orange or violet, add 10 μl of 3 M sodium acetate, pH 5.0, and mix. The colour of the mixture will turn to yellow.
 - c. All centrifugation steps will be carried out at $17,900\times g$ (13,000 rpm) at room temperature.
2. Briefly spin the PCR tubes to ensure all liquid is at the bottom.
3. Combine the duplicate PCR products by pipetting them into one 1.5 ml Eppendorf tube.
4. Add 5 volumes of Buffer PB to 1 volume of the PCR sample and mix. For example add 250 μl of Buffer PB to 50 μl PCR sample.
5. Place a QIAquick spin column in a provided 2 ml collection tube.
6. Apply the PCR sample to the column and centrifuge for 60 s.
7. Discard flow-through by pouring it in a waste jar. Place the column back into the same tube.
8. To wash, add 0.75 ml Buffer PE to the column and centrifuge for 60 s.
9. Discard flow-through and place the column back in the same tube. Centrifuge the column for an additional 1 min. This step removes residual ethanol from buffer PE.
10. Place the column in a clean 1.5 ml Eppendorf tube.
11. To elute the DNA with increased concentration, add 30 μl of water (pH 7.0–8.5) to the centre of the spin-column membrane, let the column stand for 1 min, and then centrifuge for 1 min. The DNA is ready for the sequencing reaction.
12. N.B. Use for the elution the provided buffer instead of water if the DNA is to be stored for a long-term period at -20°C .

4. Sequencing

The expected PCR amplicons are about 1,000 bp long. The amplicons will be cycle sequenced (Sanger sequencing) using the BigDye Sequence Terminator chemistry (Applied Biosystems) according to manufacturer's protocol.

The sequencing reactions work like a PCR reaction except that the ready reaction mix contains a mixture of dNTPs and dNTPs with fluorescent dye terminators. The dye terminators stop the elongation of the DNA strand and are fluorescently labelled so that the terminal nucleotide fluoresces with a colour according to its base, i.e. T is red. This occurs on a random basis, so that there will be a fluorescing terminal residue corresponding to every base in the sequence. The sequencing reaction is then analysed by capillary electrophoresis on a sequencer that uses a laser to read the fluorescence of each dye as it passes, thus determining the sequence.

Due to time constraints, the demonstrators will perform the sequencing for you.

Sequence data will be viewed and analysed with free software tools such as Mega5, 4Peaks and FinchTV.

5. Sequotyping by sequence comparison

The obtained *S. pneumoniae* *cps* sequences will be compared against GenBank database using the BLAST tool (<http://blast.ncbi.nlm.nih.gov/>). The highest BLAST bit score (usually $>98\%$ identity) indicates the correct serotype.

This method can differentiate most of the invasive pneumococcal disease serotypes within the United Kingdom and carriage strains to at least the serogroup level. If identical highest bit score appears for more serotypes it means the *cps* gene sequence is identical or highly similar for these serotypes. Consider performing conventional serotyping.

6. Further reading

Leung MH, Bryson K, Freystatter K, Pichon B, Edwards G, Charalambous BM, Gillespie SH. Sequotyping: serotyping *Streptococcus pneumoniae* by a single PCR sequencing strategy. J Clin Microbiol. 2012 Jul;50(7):2419-27.

6.4 Multiplex real-time PCR for detection of *Neisseria meningitidis* serogroups B, C and Y

Introduction

Epidemiology

Neisseria meningitidis - the meningococcus - is a Gram-negative diplococcus which is normally a harmless commensal of the human nasopharynx. On occasion however, the organism can invade the bloodstream and/or cross the blood-brain barrier and cause severe disease, namely septicaemia and meningitis. Meningococcal disease can lead to rapid death of the individual or in the case of survivors, devastating sequelae such as limb loss, hearing loss and neurological impairment [1]. Those at highest risk of developing disease are young infants and young adults.

There are 12 meningococcal serogroups that have been identified but only six are associated with most meningococcal disease worldwide (Figure 1): A, B, C, W, Y, and X [2]. In Africa, organisms with serogroup A, and more recently serogroup X capsules, are responsible for the vast majority of disease [3]. In the UK, a meningococcal conjugate C (MCC) vaccine was introduced in 1999 in response to an increase in serogroup C-related disease [4]. Since then, serogroup C-related disease has declined dramatically and today serogroup B disease dominates followed by serogroup C and serogroup Y though serogroup W has been on the increase more recently [5][6]. These three are the most common serogroups associated with disease in other Western countries also [2].

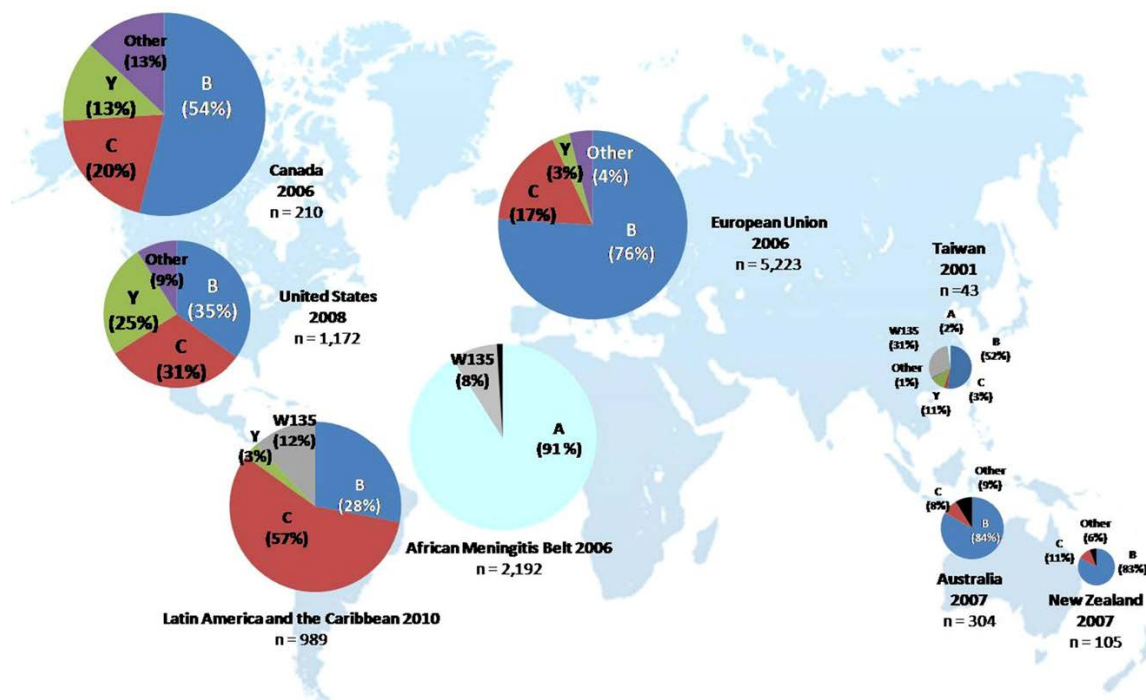


Figure 1: Proportion of meningococcal disease by serogroup by geographic region [2].

Typing

Precise characterization of bacterial isolates from cases of invasive disease is essential for informed public health responses and the management and control of disease. Conventional typing methods involve serological approaches based on the immunochemical diversity of the capsular polysaccharide of a bacteriological isolate followed by, where available and appropriate: outer-membrane protein (OMP) typing, antibiotic susceptibility profiling, PFGE fingerprinting, and/or RFLP analysis. Multi-locus enzyme electrophoresis (MLEE), the first multi-locus genetic method employed for population genetic analysis of bacteria, enabled the spread of global infections by particular genotypes to be tracked, and played an important role in revealing important aspects of the population biology of these organisms [7]. MLEE did not, however, achieve widespread application in routine typing of clinical isolates due to the technical complexity of MLEE combined with difficulties in comparing results among laboratories.

During the last decade of the 20th century and the first decade of the 21st, there was a change in emphasis in microbiological typing methods, with phenotypic and serological approaches increasingly replaced by molecular techniques that indexed genotypes [8].

For example, multi-locus sequence typing (MLST), a nucleotide sequence-based interpretation of MLEE, became the gold standard for *N. meningitidis* isolate characterization and epidemiological surveillance, and played a major part in defining the population biology of this micro-organism [9]. The development of methods for the full molecular characterization of meningococci for both culture- and non-culture-confirmed cases has provided the tools necessary for enhanced surveillance and outbreak detection and for the management of the successful implementation of major disease prevention strategies across Europe.

Serogroup determination

Serogroup determination can be achieved using slide agglutination assays when an isolate is available or from cerebrospinal fluid; however, early administration of antibiotic treatment hinders the ability to recover viable bacteria. Furthermore, human subjectivity in result interpretation, human error or poor quality antiserum may contribute to the misidentification of meningococcal serogroups [10].

Correct identification can be achieved using rapid and sensitive DNA-based methods [9]. Furthermore, a variety of techniques have been developed that identify meningococci expressing a polysaccharide capsule that contains sialic acids, including B, C, W-135 and Y meningococci. These methods include PCR-ELISA and real-time PCR protocols [11] [12]. The targeted index genes encode the respective polysialyltransferases for serogroup determination. In spite of numerous publications documenting primers and protocols for the molecular identification of meningococcal serogroups, a standardized method for molecular genogrouping has not yet been implemented.

This practical will use real-time PCR to identify serogroups. A multiplex reaction will be performed that will amplify the polysialyltransferase gene from region A of the capsule operon: the *csb*, *csc* and *csy* (previously known as *siaDb*, *siaDc* and *siaDy*) genes from serogroups B, C and Y respectively (see Figure 2 below). This assay is based on a published protocol by Wang *et al.*, 2012 [13] but others are available.

Note: Please refer to previous section on meningitis typing multiplex real-time assay for the background to TaqMan™ dual-labelled probes and real-time PCR.

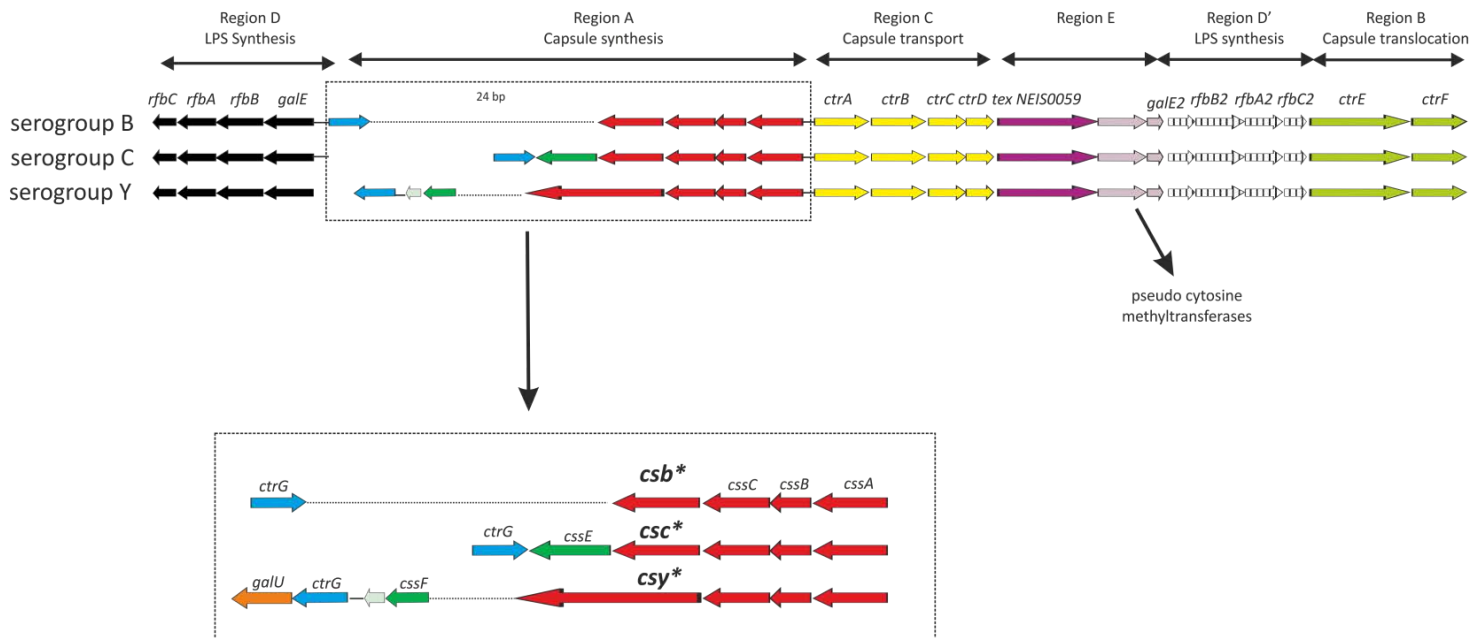


Figure 2: Capsular operon arrangement of serogroups B, C and Y with the target *csb*, *csc* and *csy* genes highlighted.

Table 1: Reporter dyes used with the multiplex probes and the channels they are detected in.

Target serogroup (gene)	Label	Max Emission (nm)	Max Absorbance (nm)	Channel
Serogroup B (<i>csb</i>)	Cy5	660±10	625±10	Red
Serogroup C (<i>csc</i>)	FAM	510±5	470±10	Green
Serogroup Y (<i>csy</i>)	HEX	555±5	530±5	Yellow

Real-time PCR reaction protocol:

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all times during the practical

1. In your tube rack you will have:

- PCR-grade sterile distilled water.
- Real-time PCR mastermix. This mix contains: Taq polymerase, buffers and dNTPs.
- Primers and probes for each of the three serogroups in the assay (B, C and Y).
The probe tubes are covered in foil to protect them from light.
- Your samples and positive controls. These are labelled as follows:

S2 = Sample 2
 S5 = Sample 5
 NTC = non-template control
 BPos = Serogroup B positive control
 CPos = Serogroup C positive control
 YPos = Serogroup Y positive control

- You will also have a plate rack with 0.1 ml strip tubes with caps.
- Label a 0.1 ml strip tube cap for each of your samples and controls (as above) and place them in the plate tray. You should also label a tube 'NTC' for your non-template control. Please also write your group number on each tube.
- Label a 1.5 ml Eppendorf tube for the serogroup assay mix.
- Prepare a mix for the assay and add to the Eppendorf. Determine the number (N) of PCR reactions as follows:

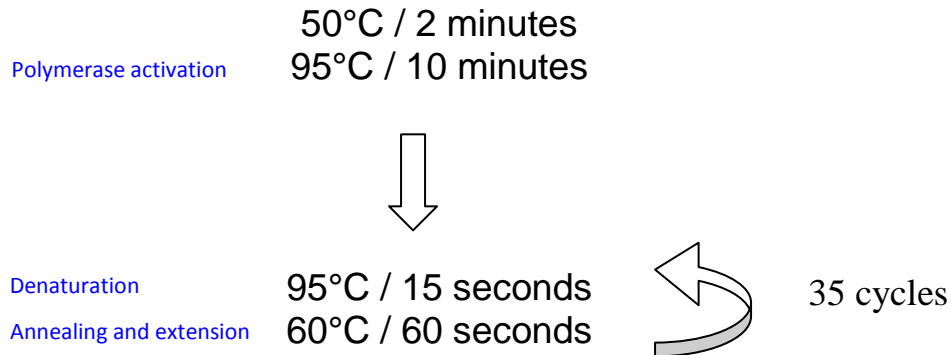
NB: In practice, real-time PCR all samples and controls are tested at least in duplicate, but for the purposes of space and time on the machine for the practical we will test just one of each.

N = the number of samples + 4 controls (1 non-template control (NTC) i.e. just mastermix, 1 serogroup B positive control, 1 serogroup C positive control, 1 serogroup Y positive control) + 2 extra (to allow for mix loss during pipetting).
Therefore N =

Table 2: BCY Assay mix per reaction tube and enough for N reaction tubes.

Reagent	Volume (µl)	
	1x tube	N x tubes
Mastermix	12.5	
H ₂ O	1.5	
B-Forward primer (7.5 µM)	1	
B-Reverse primer (7.5 µM)	1	
B-Probe (Cy5) (2.5 µM)	1	
C-Forward primer (22.5 µM)	1	
C-Reverse primer (7.5 µM)	1	
C-Probe (FAM) (2.5 µM)	1	
Y-Forward primer (22.5 µM)	1	
Y-Reverse primer (15 µM)	1	
Y-Probe (Hex) (2.5 µM)	1	
	23 µl/tube	
DNA	2	
Total	25 µl	

6. Mix gently by inverting 10 times.
7. Pipette 23 μ l into each of the **N** 0.1ml strip tubes using your P200 pipette.
8. Pipette 2 μ l of sample into the appropriate assay mix tube.
9. Place strip tube caps on tubes by pressing firmly.
10. Next, set up the Rotor-Gene:
 - Select new run.
 - Select 'Meningo grouping' program.
 - Select the 72 well run.
 - Ensure the reaction volume is set to 25 μ l.
 - Go to edit profile to check the parameters of the program are correct.
 - Real-time PCR cycling conditions:



- Click on 'Start run'.
 - Go to the sample option on the main menu and enter the appropriate reaction tube details for each position on the rotor.
11. The Rotor-Gene will then display the program details, the stage it is at and the amount of fluorescence obtained from each sample.
 12. When the run is complete, remove the tubes from the machine and discard them into a discard jar.
 13. Then select 'Analysis' from the main menu followed by the 'Quantification' option and then click on show. Set the threshold to 0.1. Repeat this process for each channel (red, green and yellow).
 14. Record the samples that have Ct values and for which channels. Determine if the assay has worked (all controls correct) and if the samples were positive for any of the targets.

Results

Sample	Serogroup B PCR	Serogroup C PCR	Serogroup Y PCR	Result
BPos				
Cpos				
YPos				
NTC				
S2				
S5				

Interpretation

Ct values ≤ 35 are considered positive; Cts between 36 and 40 are equivocal; and Ct values >40 are negative. Equivocal results are usually retested by diluting the DNA either 1:4 or 1:10 to reduce any inhibitors that may be interfering with the polymerase. The amplification plots should be analyzed to ensure they are smooth and sigmoidal in shape. If plot is lacking these characteristics, it should be considered negative or retested.

Appendix

target	Name	Sequence (5' – 3')	5' label	3' label	Amplicon size
<i>csb</i> (serogroup B)	B_F737	GCTACCCCATTTTCAGATGATTTGT			
	B_R882	ACCAGCCGAGGGTTTATTTCTAC			
	B_Pb	AAGAGATGGGYAACAACATGTAATGTCTTTATTT	Cy5	BHQ3	169 bp
<i>csc</i> (serogroup C)	C_F478	CCCTGAGTATGCGAAAAAAATT			
	C_R551	TGCTAATCCCGCCTGAATG			
	C_Pb	TTTCAATGCAATGAATACCACCGTTTTTTTGC	FAM	BHQ1	77 bp
<i>csy</i> (serogroup Y)	Y_F787	TCCGAGCAGGAAATTTATGAGAATAC			
	Y_R929	TTGCTAAAATCATTGCTCCATAT			
	Y_Pb	TATGGTGACGATATCCCTATCCTTGCCTATAAT	HEX	BHQ1	146 bp

Table A1: Primer and probes sequences of BCY assay ref. [13].

References

1. Pace, D. and A.J. Pollard, *Meningococcal disease: Clinical presentation and sequelae*. Vaccine, 2012. **30**, **Supplement 2**(0): p. B3-B9.
2. Halperin, S.A., et al., *The changing and dynamic epidemiology of meningococcal disease*. Vaccine, 2012. **30**, **Supplement 2**(0): p. B26-B36.
3. Greenwood, B., *The changing face of meningococcal disease in West Africa*. Epidemiol Infect, 2007. **135**(5): p. 703-5.

4. Miller, E., D. Salisbury, and M. Ramsay, *Planning, registration, and implementation of an immunisation campaign against meningococcal serogroup C disease in the UK: a success story*. Vaccine, 2001. **20**(Suppl 1): p. S58-67.
5. Ladhani, S.N., et al., *Invasive meningococcal disease in England and Wales: Implications for the introduction of new vaccines*. Vaccine, 2012. **30**(24): p. 3710-3716.
6. Ladhani SN, Beebeejaun K, Lucidarme J, Campbell H, Gray S, Kaczmarek E, Ramsay ME, Borrow R. Increase in Endemic Neisseria meningitidis Capsular Group W Sequence Type 11 Complex Associated With Severe Invasive Disease in England and Wales. Clin Infect Dis 2015;60(4):578-85.
7. Selander, R.K., et al., *Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics*. Applied and Environmental Microbiology, 1986. **51**: p. 837-884.
8. Maiden, M.C. and M. Frosch, *Molecular techniques for the investigation of meningococcal disease epidemiology*. Molecular Biotechnology, 2001. **18**(2): p. 119-34.
9. Maiden, M.C.J., et al., *Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms*. Proceedings of the National Academy of Sciences USA, 1998. **95**(6): p. 3140-3145.
10. Mothershed, E.A., et al., *Use of real-time PCR to resolve slide agglutination discrepancies in serogroup identification of Neisseria meningitidis*. J. Clin. Microbiol., 2004. **42**(1): p. 320-328.
11. Borrow, R., et al., *Non-culture diagnosis and serogroup determination of meningococcal B and C infection by a sialyltransferase (siaD) PCR ELISA*. Epidemiology and Infection, 1997. **118**(2): p. 111-117.
12. Tzanakaki, G., et al., *Evaluation of non-culture diagnosis of invasive meningococcal disease by polymerase chain reaction (PCR)*. FEMS Immunology and Medical Microbiology, 2003. **39**(1): p. 31-6.
13. Wang, X., et al., *Clinical Validation of Multiplex Real-Time PCR Assays for Detection of Bacterial Meningitis Pathogens*. Journal of Clinical Microbiology, 2012. **50**(3): p. 702-708.

- **N** Y if the read fails filter (read is bad), N otherwise

The second line contains the sequence of the read (ACTG or N for unresolved)

The final line is the quality score for each base. The quality is the probability that the corresponding base call is incorrect. It is referred to as a Phred quality score and is derived as follows:

$$Q_{\text{sanger}} = -10 \log_{10} p$$

The Phred score is encoded as an ASCII character by adding 64 to the Phred value.

Q=30 means probability base is incorrect is 1/1000

Q=20 means probability base is incorrect is 1/100

For paired end libraries (as prepared yesterday) you will receive two FASTQ files per sample one for each pair.

Exercise 1. Assessing the quality with FastQC

1. Open FastQC
2. Click on Basic Statistics Tab

Q1. How many sequences in the file?

NB: We can also use this command on the terminal

```
grep M00882 X.fastq > wc -l
```

Q2. What is the %GC?

3. Click on the Per base sequence quality

Q3. At what position does the average quality fall below Q30?

4. Explore the other statistics provided by FASTQC

It may be necessary to perform operations on the FASTQ files to exclude or correct low quality regions. This can involve trimming the end of reads where the quality is low (see FASTX toolkit) or using K-mer frequency distributions to correct or discard reads (see QUAKE/MUSKET).

2. *de novo* Assembly

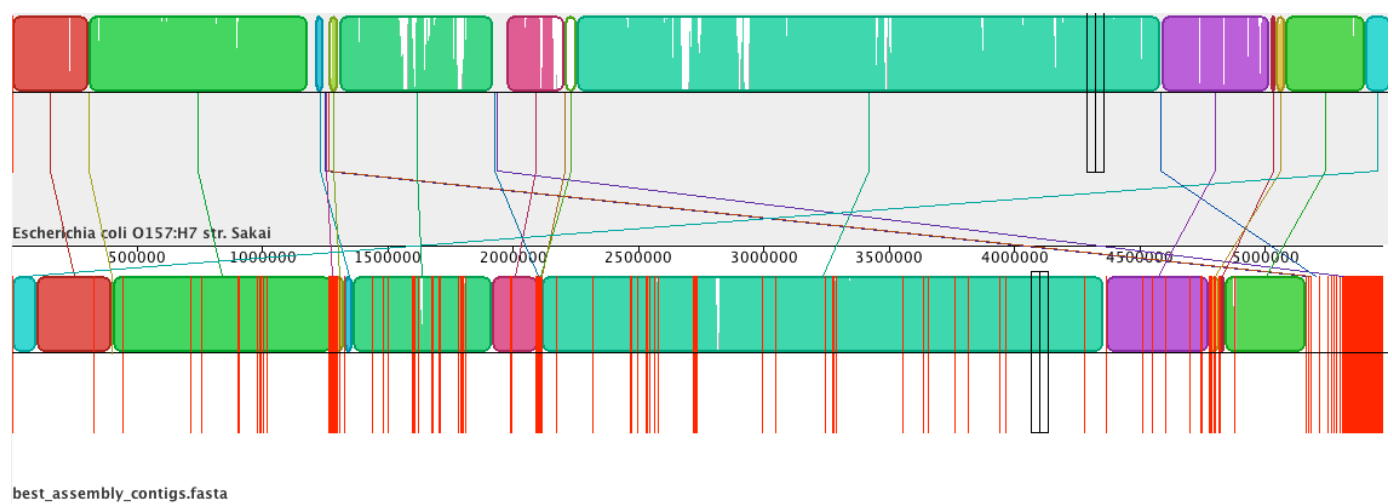
de novo assembly refers to the assembly of a genome with no reference to guide the process. A common analogy is putting a jigsaw of several hundred thousand of pieces together without the picture to refer to.

Vocabulary:

- **Read** Any sequence that comes out of the sequencer
- **Paired read** read1, gap < 500 bp, read2
- **Mate-pair** read1, gap > 1 kbp, read2
- **k-mer** Any sequence of length k
- **Contig** gap-less assembled sequence
- **Scaffold** sequence which may contain gaps (N)

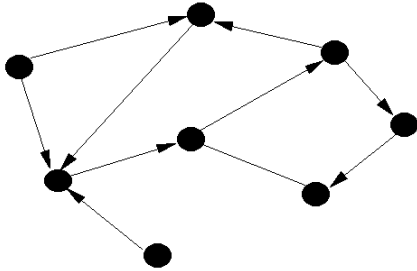
Example of a reference genome and an assembly aligned to it.

Nb. The assembly is fragmented into many contigs



Assembly Theory

de novo assembly uses graph theory. A graph is a set of nodes and a set of edges. The edges can have direction.



The graphs used to assemble short read data (illumina) uses de Bruijn graphs. Graphs allow the representation of overlaps between reads.

Example – de Bruijn graphs

Nodes represent all k-mers (k-length sub-strings) present in the reads

Edges represent the overlap between the k-mers observed.

E.G. 1 k=3 Single read

Read ACTG

Graph ACT → CTG

E.G. 2 k=3 3 reads

Read1 ACTG

Read2 CTGC

Read3 TGCT

Graph ACT → CTG → TGC → GCT

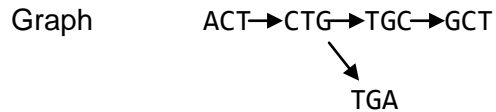
E.G. 3 k=3 4 reads – 1 has an error

Read1 ACTG

Read2 CTGC

Read3 CTGA

Read4 TGCT

**E.G. 4 k=3 6 reads – What is the effect of repeats?**

Read1 ACTG

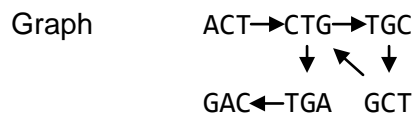
Read2 CTGC

Read3 TGCT

Read4 GCTG

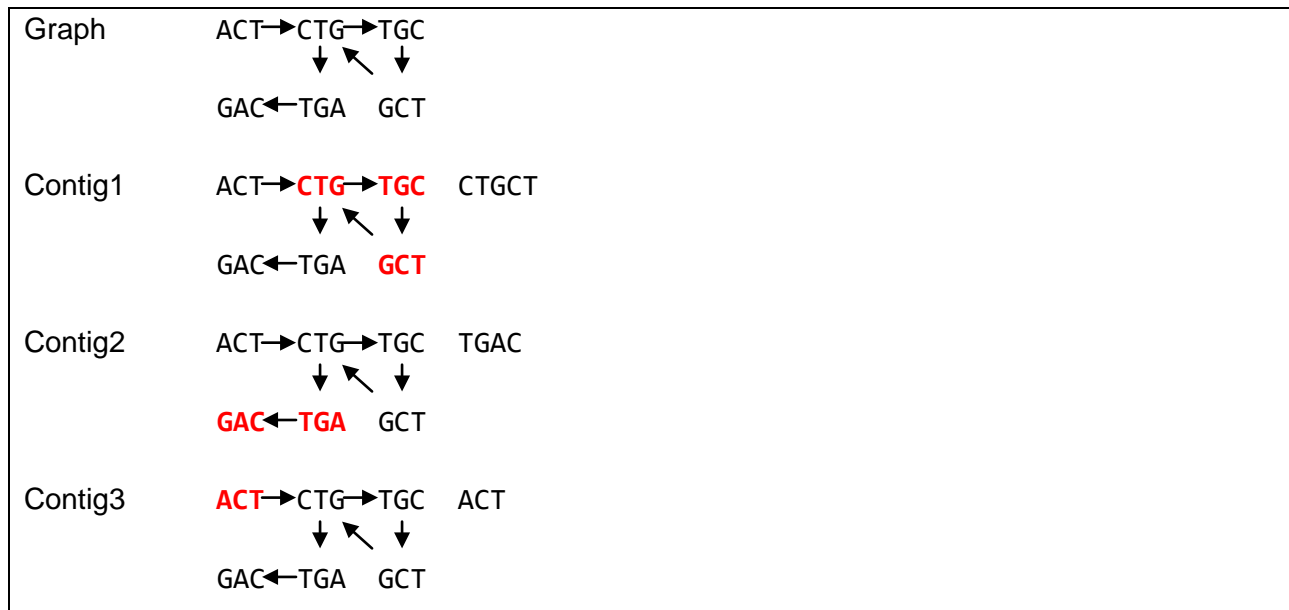
Read5 CTGA

Read6 TGAC

**How does one assemble using a graph?**

To generate contigs from our graph we calculate all node-disjoint paths through the graph.

Node-disjoint means that two different path cannot share a node.

Example – Contig construction

Choosing the k-mer value is important! Smaller k-mers increase sensitivity as more likely to observe an overlap, large k-mers increase accuracy.

VELVET (other *de novo* assemblers are available!)

Zerbino, D. R.; Birney, E. (2008). "VELVET: Algorithms for de novo short read assembly using de Bruijn graphs". *Genome Research* 18 (5): 821–829.

Exercise 2. Assemble your genome.

1. Run velveth with K-mer value of 65. This creates a data structure of all the 65-mers

Type the following command into your terminal window:

2. Create the 'Roadmap'. This creates a data structure with the position of each k-mers in the reads.

```
velveth 6930_5_14_data_65 65 -fastq -shortPaired 6930_5_14.fastq
```

3. Build the de Bruijn graph and assemble your reads into contigs

```
velvetg 6930_5_14_data_65 -ins_length 300 -exp_cov auto
```


How do we assess the assembly?

- Number of contigs/scaffolds
- Total length of the assembly
- Length of the largest contig/scaffold
- Percentage of gaps in scaffolds ('N')
- N50 of contigs/scaffolds
- Internal consistency
- Number of predicted genes

N50

N50 length is defined as the length N for which half of all bases in the sequences are in a sequence of length $L < N$

or

Half of the assembled bases reside in contig having a length of at least the n50 contig

Exercise 3.

1. How many contigs were produced for each assembly?

```
grep '>' contigs.fa | wc -l
```

2. What was the N50 reported in the logfile

```
more Log
```

3. What percentage of reads were used in the assembly?
4. What is the size of the largest contig?
5. What is the total size of the assembly?

Annotation

To understand the content of your assembled genome annotation pipeline are used.

Annotation uses a mixture of *ab initio* methods (gene prediction, signal peptide prediction) and homology inference (gene names, protein function).

There are many methods for genome annotation; here we've used PROKKA from the Victorian Bioinformatics Consortium Australia (<http://www.vicbioinformatics.com/software/prokka.shtml>).

Annotated microbial genomes are usually stored in the Genbank file format (see <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord>).

Comparing annotated assemblies with a reference

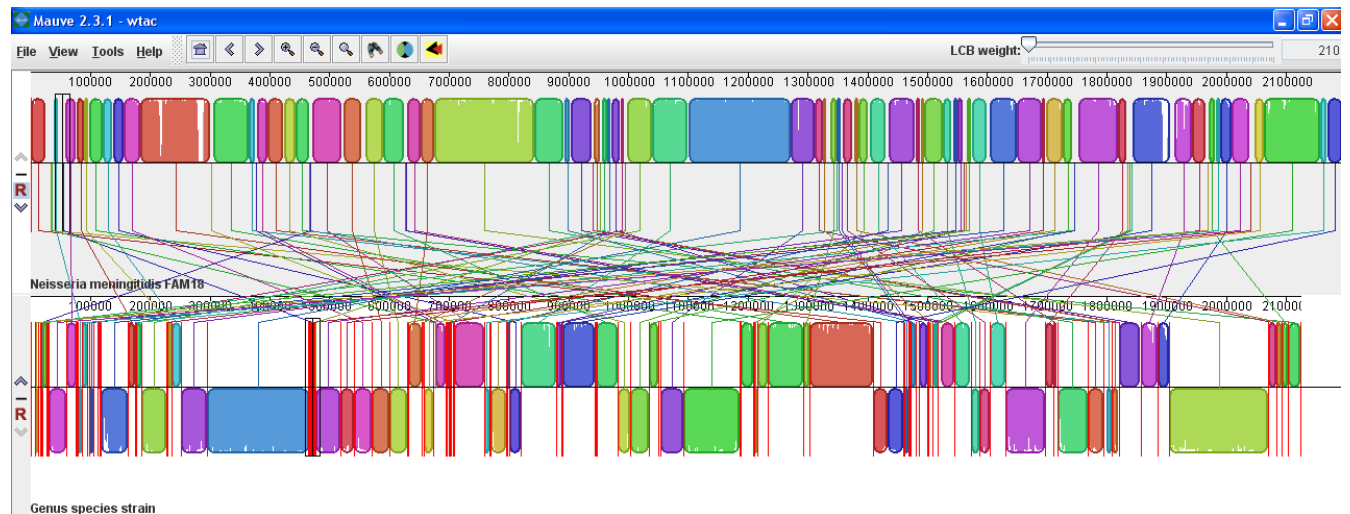
We will use MAUVE (<http://gel.ahabs.wisc.edu/mauve/>) to align our assembly against a reference genome.

1. Open MAUVE
2. Click the File->Align with progressiveMauve
3. Click add sequences and select the genomes
 - a. MyAssembly.gbk
 - b. FAM18.gbk
4. Select an output file to save your alignment

MAUVE identifies Locally Collinear Blocks (LCBs) orthologous regions that appear to be internally free from genome rearrangements. The resultant alignment shows coloured blocks of similarity.

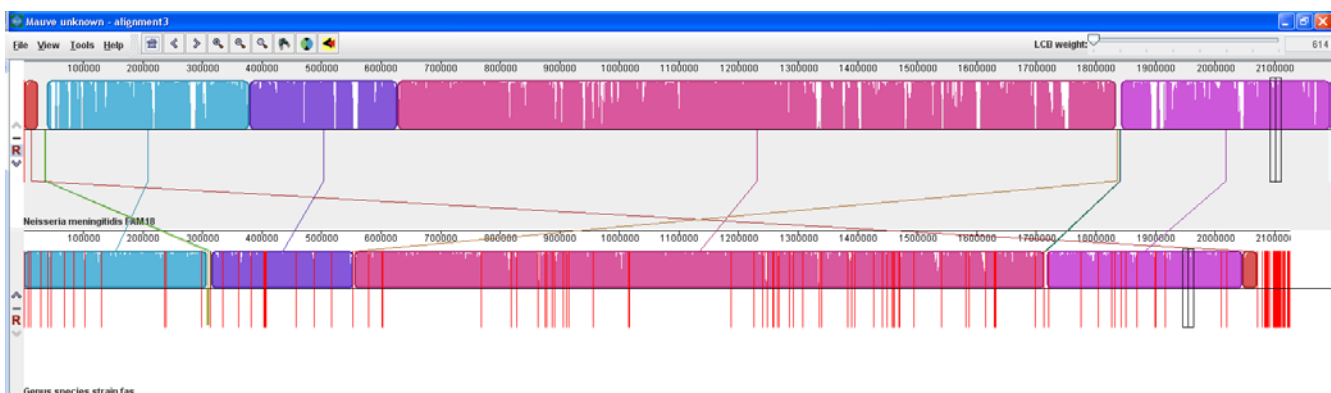
Exercise 4.

1. How many LCBs in the alignment?



We can rearrange the contigs in our assembly so they are in the same orientation and order as the reference.

2. Click Tools->Move contigs
3. Choose a location to keep the output files
4. Click add sequences and select the genomes
 - a. FAM18.gbk
 - b. MyAssembly.gbk



Exercise 5.

1. How many LCBs in the ordered alignment?

There are several regions present in our assembly that are not in the reference (indicated white at the right hand side of the assembly and *vice-versa*).

2. Find three genes that are present in our assembly but not in FAM18.

MAUVE can be used to search for specific features in the alignment.

3. Click on the binoculars icon to bring up the sequence navigator.

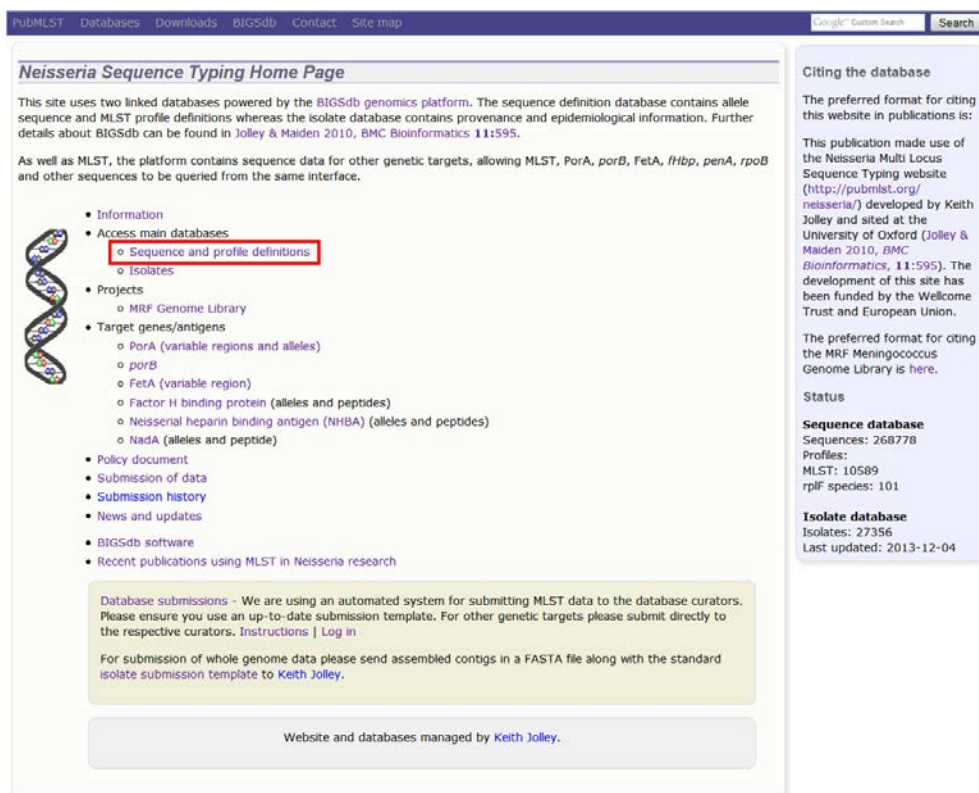
We shall now search for some genes that are important virulence determinants in these Gram negative bacteria (see separate exercise sheet).

6.6 Identifying alleles and sequence types using PubMLST

Determining allele identifier from a sequence

The allele numbers of specific sequences can be determined using the PubMLST sequence definition databases. In particular you may have sequences for MLST loci, surface antigens such as *porA*, or antibiotic resistance genes such as *dhps*. Sequence querying works on any length of sequence, including whole genome data, so you can upload contig assemblies to the website for sequence determination of individual loci.

1. From the PubMLST *Neisseria* front page (<http://pubmlst.org/neisseria/>), navigate to the sequence and profile definitions database:



Neisseria Sequence Typing Home Page

This site uses two linked databases powered by the BIGSdb genomics platform. The sequence definition database contains allele sequence and MLST profile definitions whereas the isolate database contains provenance and epidemiological information. Further details about BIGSdb can be found in Jolley & Maiden 2010, *BMC Bioinformatics* **11**:595.

As well as MLST, the platform contains sequence data for other genetic targets, allowing MLST, *PorA*, *porB*, *FetA*, *I/hbp*, *penA*, *rpoB* and other sequences to be queried from the same interface.

- Information
- Access main databases
 - Sequence and profile definitions
 - Isolates
- Projects
 - MRF Genome Library
- Target genes/antigens
 - PorA* (variable regions and alleles)
 - porB*
 - FetA* (variable region)
 - Factor H binding protein (alleles and peptides)
 - Neisseria heparin binding antigen (NHBA) (alleles and peptides)
 - NadA* (alleles and peptide)
- Policy document
- Submission of data
- Submission history
- News and updates
- BIGSdb software
- Recent publications using MLST in *Neisseria* research

Database submissions - We are using an automated system for submitting MLST data to the database curators. Please ensure you use an up-to-date submission template. For other genetic targets please submit directly to the respective curators. [Instructions](#) | [Log in](#)

For submission of whole genome data please send assembled contigs in a FASTA file along with the standard isolate submission template to [Keith Jolley](#).

Website and databases managed by [Keith Jolley](#).

Citing the database

The preferred format for citing this website in publications is:

This publication made use of the *Neisseria* Multi Locus Sequence Typing website (<http://pubmlst.org/neisseria/>) developed by Keith Jolley and sited at the University of Oxford (Jolley & Maiden 2010, *BMC Bioinformatics*, **11**:595). The development of this site has been funded by the Wellcome Trust and European Union.

The preferred format for citing the MRF Meningococcus Genome Library is [here](#).

Status

Sequence database
Sequences: 268778
Profiles:
MLST: 10589
rPLF species: 101

Isolate database
Isolates: 27356
Last updated: 2013-12-04

2. Click the 'Sequence query' link:

PubMLST Query: Sequences | Batch sequences | Compare alleles | Profile/ST | Batch profiles | List | Browse | Query
 Download: Alleles | MLST profiles
 Links: Contents | Home | PorA | FetA | Options | Isolate Database

Neisseria locus/sequence definitions database

The Neisseria PubMLST sequence definition database contains allele and profile data representing the total known diversity of Neisseria species. Every new ST deposited in this database should have a corresponding record in the isolate database.

Query database

- Sequence query - query an allele sequence.
- Batch sequence query - query multiple sequences in FASTA format.
- Sequence attribute search - find alleles by matching attributes.
- Browse MLST profiles
- Search MLST profiles
- List - find MLST profiles matched to entered list
- Batch profile query - lookup MLST profiles copied from a spreadsheet.
- Extract timespice from whole genome data (experimental)
- Search by combinations of MLST alleles - including partial matching

Downloads

- Allele sequences
- MLST profiles

Option settings

- Set general options

General information

- Number of sequences: 121681
- Number of profiles (MLST): 10026
- Last updated: 2013-01-08
- About BCSdb

Export

- Concatenate alleles
- XMFA export

Analysis

- Sequence similarity - find sequences most similar to selected allele.
- Sequence comparison - display a comparison between two sequences.
- Locus Explorer - tool for analysing allele sequences stored for particular locus.

3. Select the locus from the drop-down box and paste your sequence in to the form. Press 'Submit':

PubMLST Query: Sequences | Batch sequences | Compare alleles | Profile/ST | Batch profiles | List | Browse | Query
 Download: Alleles | MLST profiles
 Links: Contents | Home | PorA | FetA | Options | Isolate Database

Sequence query - Neisseria locus/sequence definitions

Please paste in your sequence to query against the database. Query sequences will be checked first for an exact match against the chosen (or all) loci - they do not need to be trimmed. The nearest partial matches will be identified if an exact match is not found. You can query using either DNA or peptide sequences. [?]

Please select locus/scheme: **porA (NEIS1364)** Order results by: **locus**

Enter query sequence (single or multiple configs up to whole genome in size)

```

GATGAAAGCCAAAGTACCACTCCCTTGAAGAAACCATCAAGTTACACCCCTTACCGGGCGGC
TATGAGGAGAGCGGGCTTGAATCAAGCTTGGCGGCTCACTTGGAACTTGGTGAAGAAATGCG
GACAAAGCAAAACAGTACGACCAATTCGCGGCTCACTTGGAACTTGGTGAAGAAATGCG
SCAGTTCACGCAATCAGCTATGCGCATGGTTCGACTTATCGAACGCGGTAAAGAAAGGC
GAAATACACGCTACGATCAATCATCGCGGGCTTGGATATGATTTTCCAAACGCACT
TGCGGCATCGTCTCGCGGCTTGGCTGAAGCGCAATACCGGCATCGCGCACTACACTGAA
ATTAAAGCGCGCTCGTGGCTTGGCGGCGCAAAATGCTAA
  
```

Reset Submit

If the sequence has been previously identified, the website will display the corresponding allele number:

PubMLST Query: Sequences | Batch sequences | Compare alleles | Profile/ST | Batch profiles | List | Browse | Query
 Download: Alleles | MLST profiles
 Links: Contents | Home | PorA | FetA | Options | Isolate Database

Sequence query - Neisseria locus/sequence definitions

Please paste in your sequence to query against the database. Query sequences will be checked first for an exact match against the chosen (or all) loci - they do not need to be trimmed. The nearest partial matches will be identified if an exact match is not found. You can query using either DNA or peptide sequences. [?]

Please select locus/scheme: **porA (NEIS1364)** Order results by: **locus**

Enter query sequence (single or multiple configs up to whole genome in size)

Reset Submit

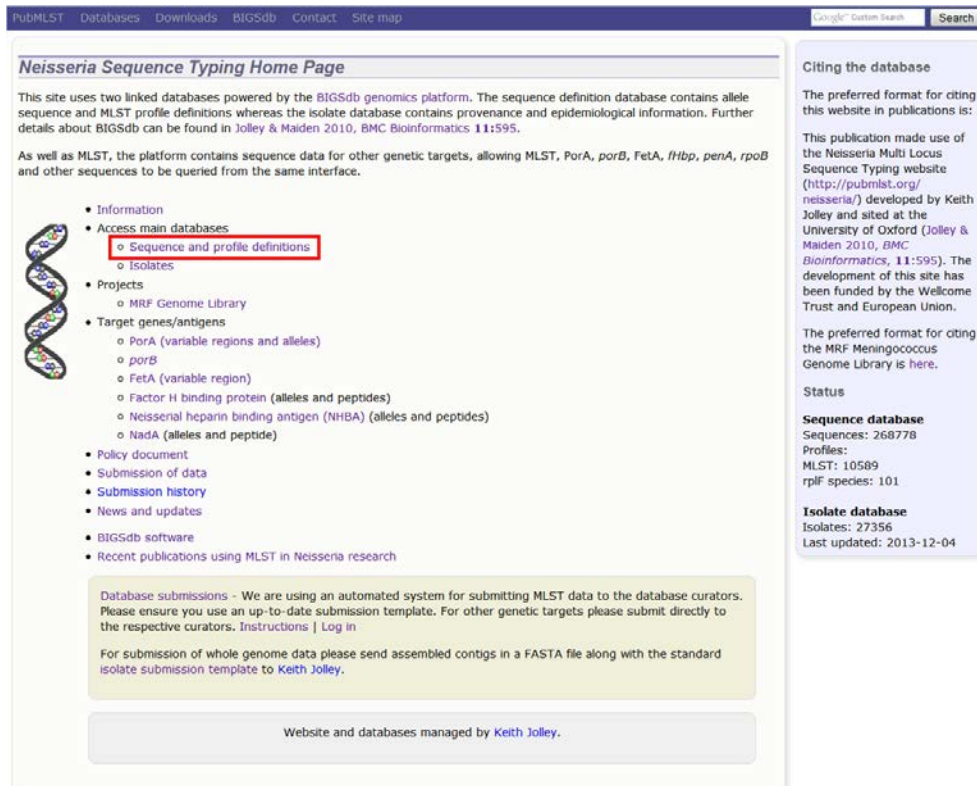
1 exact match found

Allele	Length	Start position	End position	Flags
NEIS1364 (porA)	6	1179	1	1179

Determining ST from MLST allelic profile

If you have been provided with MLST allelic profile results, you can look up the sequence type (ST) and clonal complex information as follows:

1. From the PubMLST *Neisseria* front page (<http://pubmlst.org/neisseria/>), navigate to the sequence and profile definitions database:



Neisseria Sequence Typing Home Page

This site uses two linked databases powered by the BIGSdb genomics platform. The sequence definition database contains allele sequence and MLST profile definitions whereas the isolate database contains provenance and epidemiological information. Further details about BIGSdb can be found in Jolley & Maiden 2010, *BMC Bioinformatics* **11**:595.

As well as MLST, the platform contains sequence data for other genetic targets, allowing MLST, *PorA*, *porB*, *FetA*, *fhbp*, *penA*, *rpoB* and other sequences to be queried from the same interface.

- Information
- Access main databases
 - Sequence and profile definitions
 - Isolates
- Projects
 - MRF Genome Library
- Target genes/antigens
 - PorA* (variable regions and alleles)
 - porB*
 - FetA* (variable region)
 - Factor H binding protein (alleles and peptides)
 - Neisserial heparin binding antigen (NHBA) (alleles and peptides)
 - NadA* (alleles and peptide)
- Policy document
- Submission of data
- Submission history
- News and updates
- BIGSdb software
- Recent publications using MLST in *Neisseria* research

Database submissions - We are using an automated system for submitting MLST data to the database curators. Please ensure you use an up-to-date submission template. For other genetic targets please submit directly to the respective curators. [Instructions](#) | [Log in](#)

For submission of whole genome data please send assembled contigs in a FASTA file along with the standard isolate submission template to [Keith Jolley](#).

Website and databases managed by [Keith Jolley](#).

Citing the database

The preferred format for citing this website in publications is:

This publication made use of the *Neisseria* Multi Locus Sequence Typing website (<http://pubmlst.org/neisseria/>) developed by Keith Jolley and sited at the University of Oxford (Jolley & Maiden 2010, *BMC Bioinformatics*, **11**:595). The development of this site has been funded by the Wellcome Trust and European Union.

The preferred format for citing the MRF Meningococcus Genome Library is [here](#).

Status

Sequence database
Sequences: 268778
Profiles:
MLST: 10589
rPLF species: 101

Isolate database
Isolates: 27356
Last updated: 2013-12-04

2. Click the 'Search by combinations of MLST alleles' link:



Neisseria locus/sequence definitions database

The *Neisseria* PubMLST sequence definition database contains allele and profile data representing the total known diversity of *Neisseria* species. Every new ST deposited in this database should have a corresponding record in the isolate database.

Query database

- Sequence query - query an allele sequence.
- Batch sequence query - query multiple sequences in FASTA format.
- Sequence attribute search - find alleles by matching attributes.
- Browse profiles
- List - find profiles matched to entered list.
- Search by combinations of alleles - including partial matching.
- Batch profile query - lookup profiles copied from a spreadsheet.
- Extract finetype from whole genome data (experimental) **NEW**

Downloads

- Allele sequences
- MLST

Option settings

- Set general options

General information

- Number of sequences: 268778
- Number of profiles: [Show](#)
- Last updated: 2013-12-04
- Profile update history
- About BIGSdb

Export

- Concatenate alleles
- MLPA export

Analysis

- Sequence similarity - find sequences most similar to selected allele.
- Sequence comparison - display a comparison between two sequences.
- Locus Explorer - tool for analysing allele sequences stored for particular locus.

3. Enter the allelic profile in to the web form and press submit:

Query: Sequences | Batch sequences | Compare alleles | Profile/ST | Batch profiles | List | Browse | Query
 Download: Alleles | MLST profiles
 Links: Contents | Home | PorA | FetA | Options | Isolate Database

Search *Neisseria* locus/sequence definitions database by combinations of MLST loci

Please enter your allelic profile below. Blank loci will be ignored.

abcZ	adk	aroE	fumC	gdh	pdhC	pgm
2	3	4	3	8	4	6

Autofill profile: ST:

Options: Search: Order by:
 Display: records per page

4. The ST and clonal complex (if defined) will then be displayed:

Query: Sequences | Batch sequences | Compare alleles | Profile/ST | Batch profiles | List | Browse | Query
 Download: Alleles | MLST profiles
 Links: Contents | Home | PorA | FetA | Options | Isolate Database

Search *Neisseria* locus/sequence definitions database by combinations of MLST loci

Please enter your allelic profile below. Blank loci will be ignored.

abcZ	adk	aroE	fumC	gdh	pdhC	pgm
2	3	4	3	8	4	6

Autofill profile: ST:

Options: Search: Order by:
 Display: records per page

Exact matches found (7 loci).
 1 record returned. Click the hyperlink for detailed information.

ST	abcZ	adk	aroE	fumC	gdh	pdhC	pgm	clonal complex
11	2	3	4	3	8	4	6	ST-11 complex/ET-37 complex

Analysis tools:

Export:

Querying PubMLST isolate databases

Select the isolate database from a species page on pubmlst.org (e.g. <http://pubmlst.org/neisseria/>):

Neisseria Sequence Typing Home Page

This site uses two linked databases powered by the BIGSdb genomics platform. The sequence definition database contains allele sequence and MLST profile definitions whereas the isolate database contains provenance and epidemiological information. Further details about BIGSdb can be found in Jolley & Maiden 2010, *BMC Bioinformatics* **11**:595.

As well as MLST, the platform contains sequence data for other genetic targets, allowing MLST, *PorA*, *porB*, *FetA*, *fHbp*, *penA*, *rpoB* and other sequences to be queried from the same interface.

- Information
- Access main databases
 - Sequence and profile definitions
 - Isolates**
- Projects
 - MRF Genome Library
- Target genes/antigens
 - PorA* (variable regions and alleles)
 - porB*
 - FetA* (variable region)
 - Factor H binding protein (alleles and peptides)
 - Neisserial heparin binding antigen (NHBA) (alleles and peptides)
 - NadA* (alleles and peptide)
- Policy document
- Submission of data
- Submission history
- News and updates
- BIGSdb software
- Recent publications using MLST in Neisseria research

Database submissions - We are using an automated system for submitting MLST data to the database curators. Please ensure you use an up-to-date submission template. For other genetic targets please submit directly to the respective curators. [Instructions](#) | [Log in](#)

For submission of whole genome data please send assembled contigs in a FASTA file along with the standard [isolate submission template](#) to [Keith Jolley](#).

Website and databases managed by [Keith Jolley](#).

Citing the database

The preferred format for citing this website in publications is:

This publication made use of the Neisseria Multi Locus Sequence Typing website (<http://pubmlst.org/neisseria/>) developed by Keith Jolley and sited at the University of Oxford (Jolley & Maiden 2010, *BMC Bioinformatics*, **11**:595). The development of this site has been funded by the Wellcome Trust and European Union.

The preferred format for citing the MRF Meningococcus Genome Library is [here](#).

Status

Sequence database
Sequences: 312521
Profiles:
MLST: 10802
rplF species: 108

Isolate database
Isolates: 28132
Last updated: 2014-05-09

Click the 'Search database' link:

Neisseria PubMLST database

The Neisseria PubMLST database contains data for a collection of isolates that represent the total known diversity of Neisseria species. For every allelic profile in the profiles/sequence definition database there is at least one corresponding isolate deposited here. Any isolate may be submitted to this database and consequently it should be noted that it does not represent a population sample.

Query database

- Search database** - advanced queries.
- Browse database - peruse all records.
- Search by combinations of loci (profiles) - including partial matching.
- List query - find isolates by matching a field to an entered list.

Option settings

- Set general options - including isolate table field handling
- Set display and query options for locus, schemes or scheme fields.

General information

- Isolates: 28132
- Last updated: 2014-05-09
- Update history
- About BIGSdb

Breakdown

- Single field
- Two field
- Unique combinations
- Scheme and alleles
- Publications
- Sequence bin

Export

- Export dataset
- Contigs
- Concatenate alleles
- XMFA export

Analysis

- Codon usage
- Presence/absence status of loci
- Genome comparator
- BLAST

Miscellaneous

- Description of database fields

Searching provenance data

The standard query form initially provides a means to query provenance information, e.g. isolate name, country and year of isolation etc. To search for all isolates from Africa, select 'continent' in the field dropdown box, and enter 'Africa' as the value. Click 'Submit':

PubMLST Query: [Search](#) | [Browse](#) | [Profile/ST](#) | [List](#)
 Breakdown: [Isolate fields](#) | [Scheme/alleles](#) | [Publications](#)
 Links: [Contents](#) | [Home](#) | [Options](#) | [Profiles/sequences definitions](#) | [Database submissions](#)

Toggle: [?](#) Field help: [id](#) Go

Search *Neisseria* PubMLST database

Isolate provenance/phenotype fields

continent = Africa + [?](#)

Display/sort options

Order by: [id](#) ascending

Display: 25 records per page [?](#)

Action

Reset Submit

2103 records returned (1 - 25 displayed). Click the hyperlinks for detailed information.

Page: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) > Last

Isolate fields ?								MLST		Finotyping antigens		
id	isolate	aliases	country	year	disease	species	serogroup	ST	clonal complex	PorA VR1	PorA VR2	FetA VR
19	S3131	B213; Z1213	Ghana	1973	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	4	ST-4 complex/subgroup IV	7	13-1	F1-5
31	10	B269; Z1269	Burkina Faso	1963	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	4	ST-4 complex/subgroup IV	7	13-1	F1-5
34	20	B275; Z1275	Niger	1963	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	1	ST-1 complex/subgroup III	5-2	10	F1-7
35	26	B278; Z1278	Niger	1963	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	4	ST-4 complex/subgroup IV	7	13	F1-5
46	255	B318; Z1318	Burkina Faso	1966	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	4	ST-4 complex/subgroup IV	7-2	13-1	F1-5
52	243	B362; Z1362	Cameroon	1966	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	4	ST-4 complex/subgroup IV	7	13	F1-5
56	223	B380; Z1380	Djibouti	1965	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	1	ST-1 complex/subgroup III		10	

You can build up more complex queries by adding further search terms. Additional query boxes can be added to the form by clicking the '+' button:

PubMLST Query: [Search](#) | [Browse](#) | [Profile/ST](#) | [List](#)
 Breakdown: [Isolate fields](#) | [Scheme/alleles](#) | [Publications](#)
 Links: [Contents](#) | [Home](#) | [Options](#) | [Profiles/sequences definitions](#) | [Database submissions](#)

Toggle: [?](#) Field help: [id](#) Go

Search *Neisseria* PubMLST database

Isolate provenance/phenotype fields

Combine with: AND

continent = Africa + [?](#)

id = Enter value... + [?](#)

Display/sort options

Order by: [id](#) ascending

Display: 25 records per page [?](#)

Action

Reset Submit

2103 records returned (1 - 25 displayed). Click the hyperlinks for detailed information.

Page: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) > Last

Isolate fields ?								MLST		Finotyping antigens		
id	isolate	aliases	country	year	disease	species	serogroup	ST	clonal complex	PorA VR1	PorA VR2	FetA VR
19	S3131	B213; Z1213	Ghana	1973	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	4	ST-4 complex/subgroup IV	7	13-1	F1-5
31	10	B269; Z1269	Burkina Faso	1963	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	4	ST-4 complex/subgroup IV	7	13-1	F1-5
34	20	B275; Z1275	Niger	1963	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	1	ST-1 complex/subgroup III	5-2	10	F1-7
35	26	B278; Z1278	Niger	1963	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	4	ST-4 complex/subgroup IV	7	13	F1-5
46	255	B318; Z1318	Burkina Faso	1966	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	4	ST-4 complex/subgroup IV	7-2	13-1	F1-5
52	243	B362; Z1362	Cameroon	1966	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	A	4	ST-4 complex/subgroup IV	7	13	F1-5

For example, you can combine the previous query with year of isolation, e.g. to select records for isolates isolated after the year 2000, select 'year' in the newly appeared dropdown field box, choose '>' as the modifier, and enter '2000' as the field value. Click submit:

Query: [Search](#) | [Browse](#) | [Profile/ST](#) | [List](#)
 Breakdown: [Isolate fields](#) | [Scheme/alleles](#) | [Publications](#)
 Links: [Contents](#) | [Home](#) | [Options](#) | [Profiles/sequences definitions](#) | [Database submissions](#)

Toggle: [?](#) Field help: [id](#) [Go](#)

Search *Neisseria* PubMLST database

Isolate provenance/phenotype fields Display/sort options

Combine with: **AND**

continent = Africa + [?](#)

year > 2000

Order by: **id** ascending

Display: 25 records per page [?](#)

[Modify form options](#)

Action

[Reset](#) [Submit](#)

1590 records returned (1 - 25 displayed). Click the hyperlinks for detailed information.

Page: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [>](#) [Last](#)

Isolate fields ?								MLST		Finotyping antigens		
id	isolate	aliases	country	year	disease	species	serogroup	ST	clonal complex	PorA VR1	PorA VR2	FetA VR
309	Mrs 2001047		Morocco	2001	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	B	303	ST-41/44 complex/Lineage 3			
574	2162		South Africa	2001	invasive (unspecified/other)	<i>Neisseria meningitidis</i>		1387	ST-865 complex			
575	3188		South Africa	2001	invasive (unspecified/other)	<i>Neisseria meningitidis</i>		1388	ST-4240/6688 complex			
576	S12/01		Algeria	2001	invasive (unspecified/other)	<i>Neisseria meningitidis</i>		1389	ST-23 complex/Cluster A3			
577	S40/01		Algeria	2001	invasive (unspecified/other)	<i>Neisseria meningitidis</i>		1390	ST-11 complex/ET-37 complex			
999	2093		South Africa	2005		<i>Neisseria meningitidis</i>	B	6709				

Modifying query interface

Sometimes you will want to query by more than just provenance fields. Additional search criteria can be added to the form by clicking the 'Modify form options' tab on the right hand side of the page:

Query: [Search](#) | [Browse](#) | [Profile/ST](#) | [List](#)
 Breakdown: [Isolate fields](#) | [Scheme/alleles](#) | [Publications](#)
 Links: [Contents](#) | [Home](#) | [Options](#) | [Profiles/sequences definitions](#) | [Database submissions](#)

Toggle: [?](#) Field help: [id](#) [Go](#)

Search *Neisseria* PubMLST database

Isolate provenance/phenotype fields Display/sort options

id = Enter value... + [?](#)

Order by: **id** ascending

Display: 25 records per page [?](#)

[Modify form options](#)

Action

[Reset](#) [Submit](#)

This displays a box that allows you to add additional query types:

Query: [Search](#) | [Browse](#) | [Profile/ST](#) | [List](#)
 Breakdown: [Isolate fields](#) | [Scheme/alleles](#) | [Publications](#)
 Links: [Contents](#) | [Home](#) | [Options](#) | [Profiles/sequences definitions](#) | [Database submissions](#)

Toggle: [?](#) Field help: id [Go](#)

Search *Neisseria* PubMLST database

Isolate provenance/phenotype fields

id = Enter value... [+](#) [?](#)

Display/sort options

Order by: id

Display: 25

Action

[Reset](#) [Submit](#)

Modify form parameters

Click to add or remove additional query terms:

- [Show](#) Allele designations/scheme field values
- [Show](#) Allele designation status
- [Show](#) Tagged sequence status
- [Show](#) Filters

[\[X\]](#) [Modify form options](#)

To allow searching by ST or by allele designations, click the 'Show' button next to 'Allele designations/scheme field values':

Query: [Search](#) | [Browse](#) | [Profile/ST](#) | [List](#)
 Breakdown: [Isolate fields](#) | [Scheme/alleles](#) | [Publications](#)
 Links: [Contents](#) | [Home](#) | [Options](#) | [Profiles/sequences definitions](#) | [Database submissions](#)

Toggle: [?](#) Field help: id [Go](#)

Search *Neisseria* PubMLST database

Isolate provenance/phenotype fields

id = Enter value... [+](#) [?](#)

Display/sort options

Order by: id

Display: 25

Action

[Reset](#) [Submit](#)

Modify form parameters

Click to add or remove additional query terms:

- [Show](#) Allele designations/scheme field values
- [Show](#) Allele designation status
- [Show](#) Tagged sequence status
- [Show](#) Filters

[\[X\]](#) [Modify form options](#)

Then close the modification box by clicking the [X] in the top left of the box or clicking the 'Modify form options' tab again.

Searching by MLST alleles/STs

With the allele designation/scheme field query type displayed (see previous section), you can now search by ST, clonal complex or allele designation, e.g. to search the entire database for ST-11 isolates, select 'ST (MLST)' from the field list and enter '11' as the value. Click submit:

PubMLST Query: Search | Browse | Profile/ST | List
Breakdown: Isolate fields | Scheme/alleles | Publications
Links: Contents | Home | Options | Profiles/sequences definitions | Database submissions

Toggle: ? Field help: id Go

Search *Neisseria* PubMLST database

Isolate provenance/phenotype fields

id = Enter value... + ?

Allele designations/scheme fields

ST (MLST) = 11 + ?

Display/sort options

Order by: id ascending

Display: 25 records per page

Action

Reset Submit

2175 records returned (1 - 25 displayed). Click the hyperlinks for detailed information.

Page: 1 2 3 4 5 6 7 8 9 > Last

Isolate fields							MLST		
id	isolate	aliases	country	year	disease	species	serogroup	ST	clonal complex
79	160		USA	1993	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	C	11	ST-11 complex/ET-37 complex
100	638		USA	1994	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	C	11	ST-11 complex/ET-37 complex
156	00-1008		UK	2000	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	C	11	ST-11 complex/ET-37 complex
161	00-1050		UK	2000	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	C	11	ST-11 complex/ET-37 complex
162	00-1122		UK	2000	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	C	11	ST-11 complex/ET-37 complex
164	00-1245		UK	2000	invasive (unspecified/other)	<i>Neisseria meningitidis</i>	C	11	ST-11 complex/ET-37 complex

As before, additional query terms can be combined by adding new form elements by clicking the '+' button. These query terms will be combined with any provenance field queries, e.g. all ST-11 isolates from Africa in years after 2000:

PubMLST Query: Search | Browse | Profile/ST | List
Breakdown: Isolate fields | Scheme/alleles | Publications
Links: Contents | Home | Options | Profiles/sequences definitions | Database submissions

Toggle: ? Field help: id Go

Search *Neisseria* PubMLST database

Isolate provenance/phenotype fields

Combine with: AND

continent = Africa + ?

year > 2000 + ?

Allele designations/scheme fields

ST (MLST) = 11 + ?

Display/sort options

Order by: id ascending

Display: 25 records per page

Action

Reset Submit

132 records returned (1 - 25 displayed). Click the hyperlinks for detailed information.

Page: 1 2 3 4 5 6 > Last

Isolate fields							MLST		Finotyping antigens			
id	isolate	aliases	country	year	disease	species	serogroup	ST	clonal complex	PorA VR1	PorA VR2	FetA VR
3518	Ni92		Niger	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			
3519	L683		Madagascar	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			
3765	Mrs2002011		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			
3766	Mrs2002012		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	NG	11	ST-11 complex/ET-37 complex			
3769	Mrs2002015		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			
3770	Mrs2002016		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			
3772	Mrs2002018		Niger	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			
3773	Mrs2002019		Niger	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			
3775	Mrs2002021		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			
3776	Mrs2002022		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			
3783	Mrs2002029		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			
3793	Mrs2002039		Niger	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			
3794	Mrs2002040		Niger	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex			

Filtering queries

Clicking the 'Modify form options' tab, allows you to display various query filters by clicking the 'Show' button next to 'Filters':

Query: Search | Browse | Profile/ST | List
Breakdown: Isolate fields | Scheme/alleles | Publications
Links: Contents | Home | Options | Profiles/sequences definitions | Database submissions

Toggle: [?](#) Field help: id [Go](#)

Search *Neisseria* PubMLST database

Isolate provenance/phenotype fields

id = Enter value... [+](#) [?](#)

Action

[Reset](#) [Submit](#)

Display/sort options

Order by: id

Display: 25

Modify form parameters

Click to add or remove additional query terms:

- [Show](#) Allele designations/scheme field values
- [Show](#) Allele designation status
- [Show](#) Tagged sequence status
- [Show](#) **Filters**

[Modify form options](#)

Newly appeared filters include publications, MLST profile completion and clonal complex. Any filter used will be combined with queries entered in other areas of the form:

Query: Search | Browse | Profile/ST | List
Breakdown: Isolate fields | Scheme/alleles | Publications
Links: Contents | Home | Options | Profiles/sequences definitions | Database submissions

Toggle: [?](#) Field help: id [Go](#)

Search *Neisseria* PubMLST database

Isolate provenance/phenotype fields

id = Enter value... [+](#) [?](#)

Display/sort options

Order by: id

Display: 25 records per page [?](#)

Action

[Reset](#) [Submit](#)

Filters

Publication: [?](#)

Project: [?](#)

MLST profiles: [?](#)

clonal complex (MLST): [?](#)

Ribosomal MLST profiles: [?](#)

Sequence bin: [?](#)

[Modify form options](#)

For example, to show ST-11 clonal complex African isolates from after 2000, combine the query as below:

Query: Search | Browse | Profile/ST | List
 Breakdown: Isolate fields | Scheme/alleles | Publications
 Links: Contents | Home | Options | Profiles/sequences definitions | Database submissions

Toggle: [i](#) Field help: [id](#) [Go](#)

Search *Neisseria* PubMLST database

Isolate provenance/phenotype fields

Combine with: **AND**

continent = Africa

year > 2000

Filters

Publication:

Project:

MLST profiles:

clonal complex (MLST): ST-11 complex/ET-37 complex

Ribosomal MLST profiles:

Sequence bin:

Display/sort options

Order by: **id** ascending

Display: 25 records per page

Action

[Reset](#) [Submit](#)

Modify form options

149 records returned (1 - 25 displayed). Click the hyperlinks for detailed information.

Page: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [Last](#)

id	isolate	aliases	country	year	disease	species	serogroup	ST	MLST		Finotyping antigens		
									clonal complex	ST	PorA VR1	PorA VR2	FetA VR
577	S40/01		Algeria	2001	invasive (unspecified/other)	<i>Neisseria meningitidis</i>		1390	ST-11 complex/ET-37 complex				
2923	BUFA25/02		Burkina Faso	2002	invasive (unspecified/other)	<i>Neisseria meningitidis</i>		1966	ST-11 complex/ET-37 complex				
2976	BUFA 2/06		Burkina Faso	2006	meningitis	<i>Neisseria meningitidis</i>	W	5779	ST-11 complex/ET-37 complex		5	2	F3-1
3518	NI92		Niger	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex				
3519	L683		Madagascar	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex				
3765	Mrs2002011		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex				
3766	Mrs2002012		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	NG	11	ST-11 complex/ET-37 complex				
3769	Mrs2002015		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex				
3770	Mrs2002016		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex				
3772	Mrs2002018		Niger	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex				
3773	Mrs2002019		Niger	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex				
3775	Mrs2002021		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex				
3776	Mrs2002022		Cameroon	2002	meningitis	<i>Neisseria meningitidis</i>	W	11	ST-11 complex/ET-37 complex				

Analysis tools:

Breakdown: [Fields](#) [Two Field](#) [Codons](#) [Polymorphic sites](#) [Combinations](#) [Schemes/alleles](#) [Publications](#) [Sequence bin](#) [Tag status](#)

Analysis: [BURST](#) [Presence/Absence](#) [Genome Comparator](#) [BLAST](#)

Export: [Dataset](#) [Contigs](#) [Concatenate](#) [XMFA](#) [Sequences](#)

Page: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [Last](#)

Analysing returned datasets

At the bottom of any page of results you will find a large number of buttons that will take you to analysis functions using the results of your query:

3518 NI92 Niger 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3519 L683 Madagascar 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3765 Mrs2002011 Cameroon 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3766 Mrs2002012 Cameroon 2002 meningitis *Neisseria meningitidis* NG 11 ST-11 complex/ET-37 complex

3769 Mrs2002015 Cameroon 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3770 Mrs2002016 Cameroon 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3772 Mrs2002018 Niger 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3773 Mrs2002019 Niger 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3775 Mrs2002021 Cameroon 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3776 Mrs2002022 Cameroon 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3783 Mrs2002029 Cameroon 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3793 Mrs2002039 Niger 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3794 Mrs2002040 Niger 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3799 Mrs2002058 Niger 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3800 Mrs2002059 Niger 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

3805 Mrs2002080 Senegal 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

4775 BF10 Burkina Faso 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

4776 L467 Niger 2001 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

4831 LNP19979 Burkina Faso 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

4832 LNP19998 Burkina Faso 2002 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

4844 LNP20458 Niger 2003 meningitis *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

4850 LNP20770 Burkina Faso 2003 meningitis and septicaemia *Neisseria meningitidis* W 11 ST-11 complex/ET-37 complex

Analysis tools:

Breakdown: [Fields](#) [Two Field](#) [Codons](#) [Polymorphic sites](#) [Combinations](#) [Schemes/alleles](#) [Publications](#) [Sequence bin](#) [Tag status](#)

Analysis: [BURST](#) [Presence/Absence](#) [Genome Comparator](#) [BLAST](#)

Export: [Dataset](#) [Contigs](#) [Concatenate](#) [XMFA](#) [Sequences](#)

Page: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [Last](#)

For example, you can breakdown the results by provenance field by clicking the 'Fields' button:

3518	Ni92	Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3519	L683	Madagascar	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3765	Mrs2002011	Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3766	Mrs2002012	Cameroon	2002	meningitis	Neisseria meningitidis	NG	11	ST-11 complex/ET-37 complex		
3769	Mrs2002015	Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3770	Mrs2002016	Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3772	Mrs2002018	Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3773	Mrs2002019	Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3775	Mrs2002021	Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3776	Mrs2002022	Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3783	Mrs2002029	Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3793	Mrs2002039	Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3794	Mrs2002040	Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3799	Mrs2002058	Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3800	Mrs2002059	Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
3805	Mrs2002080	Senegal	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
4775	BF10	Burkina Faso	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
4776	L467	Niger	2001	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
4831	LNP19979	Burkina Faso	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
4832	LNP19998	Burkina Faso	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
4844	LNP20458	Niger	2003	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		
4850	LNP20770	Burkina Faso	2003	meningitis and septicaemia	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex		

Analysis tools:

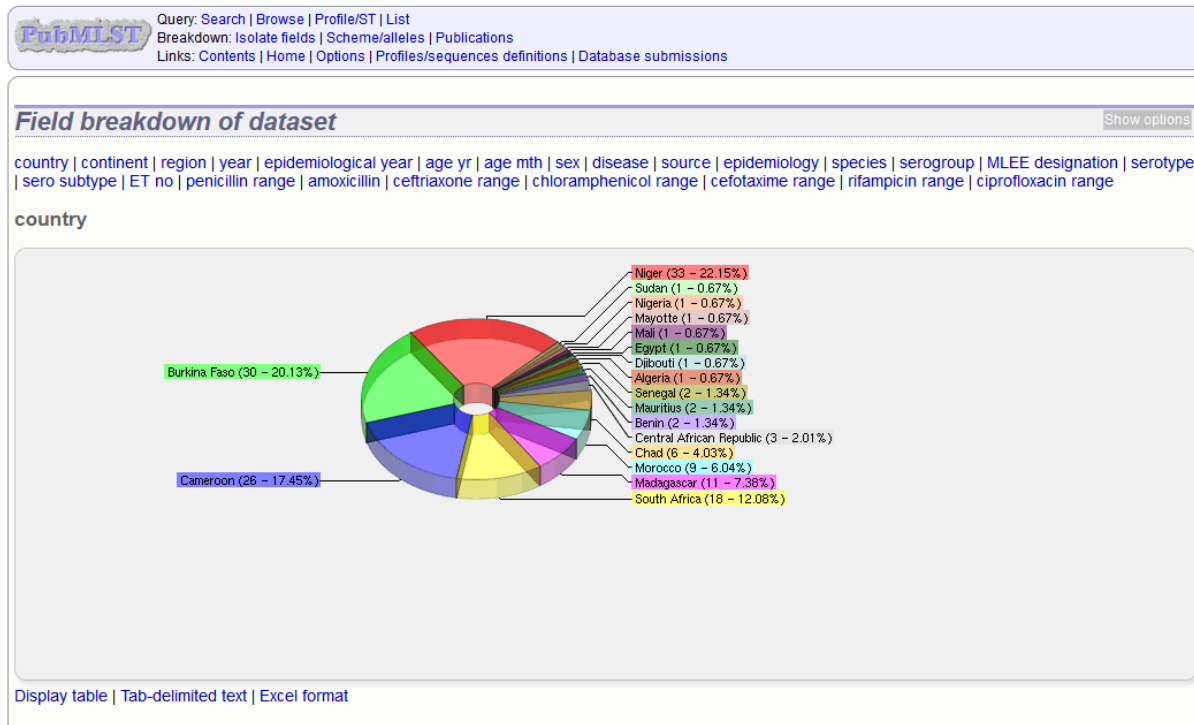
Breakdown: **Fields** Two Field Codons Polymorphic sites Combinations Schemes/alleles Publications Sequence bin Tag status

Analysis: BURST Presence/Absence Genome Comparator BLAST

Export: Dataset Contigs Concatenate XMFA Sequences

Page: 1 2 3 4 5 6 > Last

A series of charts will be displayed:



Data for these can additionally be exported in text or Excel formats.

You can also break one field down against another using the 'Two Field' breakdown:

Isolate fields							MLST		Finetyping antigens			
id	isolate	aliases	country	year	disease	species	serogroup	ST	clonal complex	PorA VR1	PorA VR2	FetA VR
577	S40/01		Algeria	2001	invasive (unspecified/other)	Neisseria meningitidis		1390	ST-11 complex/ET-37 complex			
2923	BUFA25/02		Burkina Faso	2002	invasive (unspecified/other)	Neisseria meningitidis		1966	ST-11 complex/ET-37 complex			
2976	BuFa 2/06		Burkina Faso	2006	meningitis	Neisseria meningitidis	W	5779	ST-11 complex/ET-37 complex	5	2	F3-1
3518	NI92		Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3519	L683		Madagascar	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3765	Mrs2002011		Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3766	Mrs2002012		Cameroon	2002	meningitis	Neisseria meningitidis	NG	11	ST-11 complex/ET-37 complex			
3769	Mrs2002015		Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3770	Mrs2002016		Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3772	Mrs2002018		Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3773	Mrs2002019		Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3775	Mrs2002021		Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3776	Mrs2002022		Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3783	Mrs2002029		Cameroon	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3793	Mrs2002039		Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3794	Mrs2002040		Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3799	Mrs2002058		Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3800	Mrs2002059		Niger	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
3805	Mrs2002080		Senegal	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
4775	BF10		Burkina Faso	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
4776	L467		Niger	2001	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
4831	LNP19979		Burkina Faso	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
4832	LNP19998		Burkina Faso	2002	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
4844	LNP20458		Niger	2003	meningitis	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			
4850	LNP20770		Burkina Faso	2003	meningitis and septicaemia	Neisseria meningitidis	W	11	ST-11 complex/ET-37 complex			

Analysis tools:

Breakdown:

Analysis:

Export:

Page:

This allows you to combine any field (provenance, allele designation, ST etc.). For example country vs serogroup:

Query: [Search](#) | [Browse](#) | [Profile/ST](#) | [List](#)
 Breakdown: [Isolate fields](#) | [Scheme/alleles](#) | [Publications](#)
 Links: [Contents](#) | [Home](#) | [Options](#) | [Profiles/sequences definitions](#) | [Database submissions](#)

Two field breakdown of dataset

Here you can create a table breaking down one field by another, e.g. breakdown of serogroup by year.

Select fields

Field 1:

Field 2:

Display

☒ values only

☐ values and percentages

☐ percentages only

Calculate percentages by

☒ dataset

☐ row

☐ column

Action

This will display a table of combinations.



Query: [Search](#) | [Browse](#) | [Profile/ST](#) | [List](#)

Breakdown: [Isolate fields](#) | [Scheme/alleles](#) | [Publications](#)

Links: [Contents](#) | [Home](#) | [Options](#) | [Profiles/sequences definitions](#) | [Database submissions](#)

Two field breakdown of dataset

Show options

Breakdown of country by serogroup:

Selected options: Display values only.

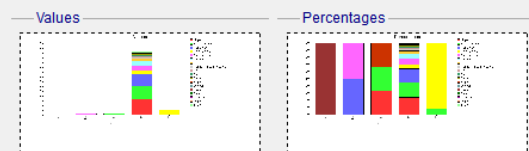
— Axes — Show —

country	serogroup					Total
	A	NG	No value	W	Y	
Algeria			1			1
Benin				2		2
Burkina Faso			1	28	1	30
Cameroon		1		25		26
Central African Republic				3		3
Chad				6		6
Djibouti				1		1
Egypt	1					1
Madagascar		1		10		11
Mali				1		1
Mauritius				2		2
Mayotte				1		1
Morocco				9		9
Niger			1	32		33
Nigeria				1		1
Senegal				2		2
South Africa				8	10	18
Sudan				1		1
Total	1	2	3	132	11	149

[Download as tab-delimited text.](#)

Charts

Click to enlarge.



6.7 Whole genome analysis using BIGSdb

A separate database instance has been set up for each pair of participants for when you come to upload and analyse your genome data. This is an empty isolate database with all *Neisseria* loci defined as currently used on PubMLST. You will have full administrator access to this database in order to upload, scan, tag and analyse sequence data for this practical.

You will be told the web address to use in order to connect to the database which will vary slightly by your pair numbering.

Extracting typing data

You have been provided with FASTA files of the sequence contigs assembled from Illumina short read data using the VELVET assembler. For full analysis you can upload these sequences to the database and associate with an isolate record and we will do this. You can, however, quickly extract standard typing information (ST, clonal complex, PorA and FetA variants) by querying these sequences against the sequence definition database. We'll do this for just one of the isolates as a demonstration.

1. From the PubMLST *Neisseria* front page, navigate to the sequence and profile definitions database:

Neisseria Sequence Typing Home Page

This site uses two linked databases powered by the BIGSdb genomics platform. The sequence definition database contains allele sequence and MLST profile definitions whereas the isolate database contains provenance and epidemiological information. Further details about BIGSdb can be found in Jolley & Maiden 2010, *BMC Bioinformatics* **11**:595.

As well as MLST, the platform contains sequence data for other genetic targets, allowing MLST, *PorA*, *porB*, *FetA*, *fHbp*, *penA*, *rfpB* and other sequences to be queried from the same interface.

- Information
- Access main databases
 - Sequence and profile definitions
 - Isolates
- Projects
 - MRF Genome Library
- Target genes/antigens
 - PorA* (variable regions and alleles)
 - porB*
 - FetA* (variable region)
 - Factor H binding protein (alleles and peptides)
 - Neisseria* heparin binding antigen (NHBA) (alleles and peptides)
 - NadA* (alleles and peptide)
- Policy document
- Submission of data
- Submission history
- News and updates
- BIGSdb software
- Recent publications using MLST in *Neisseria* research

Database submissions - We are using an automated system for submitting MLST data to the database curators. Please ensure you use an up-to-date submission template. For other genetic targets please submit directly to the respective curators. [Instructions](#) | [Log in](#)

For submission of whole genome data please send assembled contigs in a FASTA file along with the standard isolate submission template to [Keith Jolley](#).

Website and databases managed by [Keith Jolley](#).

Citing the database

The preferred format for citing this website in publications is:

This publication made use of the *Neisseria* Multi Locus Sequence Typing website (<http://pubmlst.org/neisseria/>) developed by Keith Jolley and sited at the University of Oxford (Jolley & Maiden 2010, *BMC Bioinformatics*, **11**:595). The development of this site has been funded by the Wellcome Trust and European Union.

The preferred format for citing the MRF Meningococcus Genome Library is [here](#).

Status

Sequence database
Sequences: 268778
Profiles: MLST: 10589
rplF species: 101

Isolate database
Isolates: 27356
Last updated: 2013-12-04

2. Click the 'Extract finetype from whole genome data' link:

3. You can either copy and paste the contigs in FASTA format or choose to upload the file. We'll do the latter. Click the browse button and locate the contig file for FAM18. Then click 'Submit'.

4. The extraction job gets submitted to a job queue. The status of the job and the results can be viewed by clicking the link:

5. The strain type should be displayed in the output section:

PubMLST Query: Sequences | Batch sequences | Compare alleles | Profile/ST | Batch profiles | List | Browse | Query
Download: Alleles | MLST profiles
Links: Contents | Home | PorA | Feta | Options | Isolate Database

Job status viewer

Status

Job id:	BIGSdb_19755_1348046854_61264
Submit time:	2012-09-19 10:27:34
Status:	finished
Start time:	2012-09-19 10:28:01
Progress:	100%
Stop time:	2012-09-19 10:29:00
Total time:	59 seconds

Output

Strain type

- P1.5, 2; F1-30; ST-11 (cc11)

Antibiotic resistance

- *penA* allele: 1 (penicillin MIC: >0.06 - 1 (intermediate))
- *rpoB* allele: 9 (rifampicin MIC: <=1 (susceptible))

Please note that job results will remain on the server for 7 days.

Uploading isolate records to the database

While the rapid extraction of typing data is useful for a quick look, more detailed analysis requires loading an isolate record and associated sequence data in to the database.

The first stage of this process is creating isolate records for each of the genomes we want to analyse. This can be done through the curation interface either record-by-record - which is ok if you only have one isolate to do, or more usually using a batch upload method. The batch upload can be prepared in Excel, or any other spreadsheet package, and then copy-and-pasted in to the batch add web form.

6. Enter the curation interface from the database front page:

PubMLST Query: Search | Browse | Profile/ST | List
Breakdown: Isolate fields | Scheme/alleles | Publications
Links: Contents | Home | Options | Profiles/sequences definitions | **Curate**

Welcome to the Neisseria PubMLST database

The Neisseria PubMLST database contains data for a collection of isolates that represent the total known diversity of Neisseria species. For every allelic profile in the profiles/sequence definition database there is at least one corresponding isolate deposited here. Any isolate may be submitted to this database and consequently it should be noted that it does not represent a population sample.

Query database

- Search database - advanced queries.
- Browse database - peruse all records.
- Search by combinations of loci (profiles) - including partial matching.
 - MLST
 - All loci
- List query - find isolates by matching a field to an entered list.

Option settings

- Set general options - including isolate table field handling
- Set display and query options for locus, schemes or scheme fields.

General information

- Isolates: 0
- About BIGSdb

Breakdown

- Single field
- Two field
- Polymorphic sites
- Unique combinations
- Scheme and alleles
- Publications
- Sequence bin
- Tag status

Export

- Export dataset
- Presence/absence status of loci
- Concatenate alleles
- XMFA export


Analysis

- Codon usage
- Genome comparator
- BLAST

Miscellaneous

- Description of database fields

7. If prompted, log in using the account details that you have been provided with:



[Database: Species home](#) | [Curator's page \(species\)](#) | [Curator's page \(database\)](#)
[Users: Add](#) | [Query/update](#)
[Isolates: Add](#) | [Query/update](#) | [Batch insert](#)

Not logged in.

Please log in - *Neisseria* PubMLST database

The *Neisseria* PubMLST database contains data for a collection of isolates that represent the total known diversity of *Neisseria* species. For every allelic profile in the profiles/sequence definition database there is at least one corresponding isolate deposited here. Any isolate may be submitted to this database and consequently it should be noted that it does not represent a population sample.

Please enter your log-in details. Part of your IP address is used along with your username to set up your session. If you have a session opened on a different computer, where the first three parts of the IP address vary, it will be closed when you log in here.

Log in details

Username: user2

Password: ●●●●●●

Log in

8. You should now have reached the database curator's page. Click the batch add isolates link:

Database: Species home | Curator's page (species) | Curator's page (database)
Users: Add | Query/update
Isolates: Add | Query/update | Batch insert

Logged in: User 2 (user2). Log out | Change password

Database curator's interface - Neisseria PubMLST

Add, update or delete records

Record type	Add	Batch Add	Update or delete	Comments
users	+	++	?	
user groups	+	++	?	Users can be members of these groups - use for setting access permissions.
user group members	+	++	query batch	Add users to groups for setting access permissions.
user permissions	+	++	?	Set curator permissions for individual users - these are only active for users with a status of 'curator' in the users table.
isolates	+	++	query browse list batch update	
isolate field extended attribute values	+	++	?	Add values for additional isolate field attributes.
projects	+	++	?	Set up projects to which isolates can belong.
project members	+	++	?	Add isolates to projects.
isolate aliases	+	++	?	Add alternative names for isolates.
PubMed links	+	++	?	
allele designations	+	++	?	Allele designations can be set within the isolate table functions.
sequences	+	++	?	The sequence bin holds sequence contigs from any source.
accession number links	+	++	?	Tag sequences with Genbank/EMBL accession number.
experiments	+	++	?	Set up experiments to which sequences in the bin can belong.
experiment sequence links	+	++	?	Query/delete links associating sequences to experiments.
sequence tags		scan	?	Tag regions of sequences within the sequence bin with locus information.

Database configuration

Table	Add	Batch Add	Update or delete	Comments
loci	+	++	?	
locus aliases	+	++	?	Add alternative names for loci. These can also be set when you batch add loci.
PCR reactions	+	++	?	Set up <i>in silico</i> PCR reactions. These can be used to filter genomes for tagging to specific repetitive loci.
PCR locus links	+	++	?	Link a locus to an <i>in silico</i> PCR reaction.
mutantids probes	+	++	?	Define mutantids probes for <i>in silico</i> hybridisation reaction to filter genomes for tagging to specific repetitive loci.

This takes you to a page that allows you to paste in the prepared batch data. If you were to prepare this data yourself, there is a link that provides an Excel template for you.

Database: Species home | Curator's page (species) | Curator's page (database)
Users: Add | Query/update
Isolates: Add | Query/update | Batch insert

Logged in: Keith Jolley (keith). Log out | Change password

Batch insert isolates

This page allows you to upload isolate data as tab-delimited text or copied from a spreadsheet.

- Field header names must be included and fields can be in any order. Optional fields can be omitted if you wish.
- Enter aliases (alternative names) for your isolates as a semi-colon (;) separated list.
- Enter references for your isolates as a semi-colon (;) separated list of PubMed ids (non integer ids will be ignored).
- You can also upload allele fields along with the other isolate data - simply create a new column with the locus name. These will be added with a confirmed status and method set as 'manual'.
- You can choose whether or not to include an id number field - if it is omitted, the next available id will be used automatically.
- [Download tab-delimited header for your spreadsheet](#) - use Paste special → text to paste the data.
- [Download submission template \(xlsx format\)](#)

Please select the sender from the list below:

Select sender ... Value will be overridden if you include a sender field in your pasted data.

Paste in tab-delimited text (include a field header line).

Action
Reset Submit

Back

Clicking this link provides a row of text that can be pasted in to a spreadsheet (using Paste Special) containing all the available column headings. However, an upload file has been prepared for you (available in Excel [isolate_upload.xlsx] and tab-delimited text format [isolate_upload.txt]).

9. Open either the isolate_upload.xlsx file in a spreadsheet program or isolate_upload.txt in a text editor, then copy and paste the entire set of data, including headings in to the web form. Do not worry if the formatting appears to be messed up. Select your user name from the drop-down list.

Database: Species home | Curator's page (species) | Curator's page (database)
 Users: Add | Query/update
 Isolates: Add | Query/update | Batch insert

Logged in: User 2 (user2) | Log out | Change password

Batch insert isolates

This page allows you to upload isolate data as tab-delimited text or copied from a spreadsheet.

- Field header names must be included and fields can be in any order. Optional fields can be omitted if you wish.
- Enter aliases (alternative names) for your isolates as a semi-colon (;) separated list.
- Enter references for your isolates as a semi-colon (;) separated list of PubMed ids (non integer ids will be ignored).
- You can also upload allele fields along with the other isolate data - simply create a new column with the locus name. These will be added with a confirmed status and method set as 'manual'.
- You can choose whether or not to include an id number field - if it is omitted, the next available id will be used automatically.
- [Download tab-delimited header for your spreadsheet](#) - use Paste special → text to paste the data.

Please select the sender from the list below:

2. User (user2) Value will be overridden if you include a sender field in your pasted data.
 Please paste in tab-delimited text (include a field header line).

Case 1	2839	22785191;10565901	UK	Southampton	1997	meningitis and
septicaemia		blood	Neisseria meningitidis	C		
Carrier 1	2838	22785191;10565901	UK	Southampton	1997	carrier
throat swab			Neisseria meningitidis	C		
Case 3	2837	22785191;10565901	UK	Southampton	1997	meningitis and
septicaemia			Neisseria meningitidis	C		
Case 6	2840	22785191;10565901	UK	Southampton	1997	meningitis and
septicaemia			Neisseria meningitidis	C		
Remote case 1	2844	22785191;10565901	UK	Southampton	1997	
meningitis and			Neisseria meningitidis	C		
Remote case 2	2847	22785191;10565901	UK	Southampton	1997	
meningitis and			Neisseria meningitidis	C		
Carrier 2	2845	22785191;10565901	UK	Southampton	1997	carrier
throat swab			Neisseria meningitidis	C		
Carrier 3	2846	22785191;10565901	UK	Southampton	1997	carrier
throat swab			Neisseria meningitidis	C		
Carrier 4	2843	22785191;10565901	UK	Southampton	1997	carrier
throat swab			Neisseria meningitidis	C		
Carrier 5	2842	22785191;10565901	UK	Southampton	1997	carrier
throat swab			Neisseria meningitidis	C		

Reset Submit Query

[Back](#)

10. Click 'Submit'. The data will be checked for formatting and any problems will be highlighted (there should be no errors at the file has been prepared for you). Finally click the 'Import data' button.


Uploading genome data to isolate records

- Genomics and Clinical Microbiology

Database: Species home | Curator's page (species) | Curator's page (database)
 Users: Add | Query/update
 Isolates: Add | Query/update | Batch insert

Logged in: User 2 (user2). Log out | Change password

Database curator's interface - *Neisseria* PubMLST

 Add, update or delete records

Record type	Add	Batch Add	Update or delete	Comments
users	+	++	?	
user groups	+	++	?	Users can be members of these groups - use for setting access permissions.
user group members	+	++	query batch	Add users to groups for setting access permissions.
user permissions	+	++	?	Set curator permissions for individual users - these are only active for users with a status of 'curator' in the users table.
isolates	+	++	query browse list batch update	
isolate field extended attribute values	+	++	?	Add values for additional isolate field attributes.
projects	+	++	?	Set up projects to which isolates can belong.
project members	+	++	?	Add isolates to projects.
isolate aliases	+	++	?	Add alternative names for isolates.
PubMed links	+	++	?	
allele designations	+	++	?	Allele designations can be set within the isolate table functions.
sequences	+	++	?	The sequence bin holds sequence contigs from any source.
accession number links	+	++	?	Tag sequences with Genbank/EMBL accession number.
experiments	+	++	?	Set up experiments to which sequences in the bin can belong.
experiment sequence links	+	++	?	Query/delete links associating sequences to experiments.
sequence tags		scan	?	Tag regions of sequences within the sequence bin with locus information.


 Database configuration

Table	Add	Batch Add	Update or delete	Comments
loci	+	++	?	
locus aliases	+	++	?	Add alternative names for loci. These can also be set when you batch add loci.
PCR reactions	+	++	?	Set up <i>in silico</i> PCR reactions. These can be used to filter genomes for tagging to specific repetitive loci.
PCR locus links	+	++	?	Link a locus to an <i>in silico</i> PCR reaction.
nucleotide probes	+	++	?	Define nucleotide probes for <i>in silico</i> hybridization reaction to filter genomes for tagging to specific repetitive loci.
probe locus links	+	++	?	Link a locus to an <i>in silico</i> hybridization reaction.
isolate field extended attributes	+	++	?	Define additional attributes to associate with values of a particular isolate record field.
composite fields	+	++	?	Used to construct composite fields consisting of fields from isolate, loci or scheme fields.
schemes	+	++	?	Describes schemes consisting of collections of loci, e.g. MLST.

12. Open the appropriate FASTA file in a text editor, and copy and paste this in to the web form. Select the appropriate isolate_id in the drop-down box. Leave all other options at their default and click 'Submit':

Database: Species home | Curator's page (species) | Curator's page (database)
 Users: Add | Query/update
 Isolates: Add | Query/update | Batch insert

Logged in: User 2 (user2) | Log out | Change password

Batch insert sequences

This page allows you to upload sequence data for a specified isolate record in FASTA format.

If an isolate id is chosen, then all sequences will be associated with that isolate. Alternatively, the isolate id, or any other isolate table field that uniquely defines the isolate, can be named in the identifier rows of the FASTA file. This allows data for multiple isolates to be uploaded.

Please note that you can reach this page for a specific isolate by [querying isolates](#) and then clicking 'Upload' within the isolate table.

Please fill in the following fields - required fields are marked with an exclamation mark (!).

Attributes
 isolate id: 1) FAM18
 identifier field: id
 sender: 2. User (user2)
 method:
 run id:
 assembly id:

Options
☐ Don't insert sequences shorter than 250 bps.
 Link to experiment:

Please paste in sequences in FASTA format:

```

ATCCAGCCGTAGCTTTGCGCCGCTTCGCGGATCAGCATAAAGCCTTGTGCGTAGGAAAT
GATTTTGGATGCAAGCAGGGCCTGTCTCAACGCCTCAACCCATTCTTGTTCGCGCCTTC
GACGGGGTAACGGTTCGGGCGAACAGCTTGCGGCTGTGACGCGCTGTTCTTTGAACGA
CGAAACGACAGCGGGGAATACCGCTTCGGAATCAGCGTCAACGGAATACCCAAATCCAA
AGCATTGATGCGCGTCAATTTGCGCGTGCTTTTGGCCTGCGGTATCGAGGATTTCTC
GACCAGCGTTGCGCGCTTCGTCTTATAGCCAAAATTGCGCTGTGATTCAATCAG
ATAAGAATCCAGCTCGGTTTGTCCACTCGGCAACACGCGGTGCTTGTCTGTAGGA
CAGTCCAAACCGTCTTTCATGAACGGTACGCTTCGCAATCACTGCATATCGCGTA
TTGATGCGGTATGACCACTTTTACGAAATGTCCGCAACCTCCCTGCGGACCCAAATC
GCAACACGGCTCACCTTGAGGGGTTTTCGAGCAATCGCCTGGAAAATCGGTTTAACAGC
TTCCATGCGCGCTCATCTCGGCGGCTGATAGAGCGGCGTGGCGGCACTTCCTC
TCGCGGAGACCCCATGCGCAACAAATCCCTTTTCGGAAGTAATGTGTCG
CGTGTGTCGCGGTAATGGCATGCAACCGTGCATATGATGTCGCTTTTCCAA
CAACGGAAGCAGTTGTCGATAAAGTCGTCAACACCGAACCTGGCGAACCATCATCAT
GATTTTTCGCTTTTCCAGCTTATCGACCAAGCTTTCAGGGAATACGCGCGAATAT
GCCGTCCTTTTTCGCGCGCTTTAAAATTGTCGCACTTACCGATTGTCGTTGTA
GGCGACACCTTAAATCGCAATCGTTCATATCAAAATCAGTTTTGCCCATACCGC
CAACCGATTACGCAATATCGCTTTCATTGCAAGGAGCTCCGTTATAGATTAAATTA
TCGACCGCACTCTACCTGATTACACTGTTTAACTCTTAACTTTTAAATTTT
AAAAGATGCTTTACGCTTTACCGTGCCTTCCCTGAAGG
  
```

Reset Submit Query

Back

13. A confirmation screen will be displayed. Click 'Upload':

Database: Species home | Curator's page (species) | Curator's page (database)
 Users: Add | Query/update
 Isolates: Add | Query/update | Batch insert

Logged in: User 2 (user2) | Log out | Change password

Batch insert sequences

The following sequences will be entered.

Original designation	Sequence length	Comments
3	2194961	

- Number of contigs: 1
- Minimum length: 2194961
- Maximum length: 2194961
- Total length: 2194961
- Mean length: 2194961
- N50: 2194961
- N90: 2194961
- N95: 2194961

Upload

14. Repeat for all other records.

Scanning and tagging the typing loci

Blasting the genomes against all known alleles of a particular locus is a process known as 'scanning'. Marking these identified alleles in the database is termed 'tagging'.

From the curator's index page, click 'sequence tags .. scan':

Database: Species home | Curator's page (species) | Curator's page (database)
 Users: Add | Query/update
 Isolates: Add | Query/update | Batch insert

Logged in: User 2 (user2) | Log out | Change password

Database curator's interface - Neisseria PubMLST

Add, update or delete records

Record type	Add	Batch Add	Update or delete	Comments
users	+	++	?	
user groups	+	++	?	Users can be members of these groups - use for setting access permissions.
user group members	+	++	query batch	Add users to groups for setting access permissions.
user permissions	+	++	?	Set curator permissions for individual users - these are only active for users with a status of 'curator' in the users table.
isolates	+	++	query browse list batch update	
isolate field extended attribute values	+	++	?	Add values for additional isolate field attributes.
projects	+	++	?	Set up projects to which isolates can belong.
project members	+	++	?	Add isolates to projects.
isolate aliases	+	++	?	Add alternative names for isolates.
PubMed links	+	++	?	
allele designations	+	++	?	Allele designations can be set within the isolate table functions.
sequences	+	++	?	The sequence bin holds sequence contigs from any source.
accession number links	+	++	?	Tag sequences with Genbank/EMBL accession number.
experiments	+	++	?	Set up experiments to which sequences in the bin can belong.
experiment sequence links	+	++	?	Query/delete links associating sequences to experiments.
sequence tags	+	++	scan	Tag regions of sequences within the sequence bin with locus information.

Database configuration

Table	Add	Batch Add	Update or delete	Comments
loci	+	++	?	
locus aliases	+	++	?	Add alternative names for loci. These can also be set when you batch add loci.
PCR reactions	+	++	?	Set up <i>in silico</i> PCR reactions. These can be used to filter genomes for tagging to specific repetitive loci.
PCR locus links	+	++	?	Link a locus to an <i>in silico</i> PCR reaction.
nucleotide probes	+	++	?	Define nucleotide probes for <i>in silico</i> hybridization reaction to filter genomes for tagging to specific repetitive loci.

15. Select the first isolate from the isolate selection list and the MLST and Finetyping antigens checkboxes from the schemes list. Leave other controls at their default settings and click 'Scan':

Database: Species home | Curator's page (species) | Curator's page (database)
 Users: Add | Query/update
 Isolates: Add | Query/update | Batch insert

Logged in: User 2 (user2) | Log out | Change password

Sequence tag scan

Please select the required isolate ids and loci for sequence scanning - use ctrl or shift to make multiple selections. In addition to selecting individual loci, you can choose to include all loci defined in schemes by selecting the appropriate scheme description. By default, loci are only scanned for an isolate when no allele designation has been made or sequence tagged. You can choose to rescan loci with existing designations or tags by selecting the appropriate options.

Isolates

- 1) FAM18
- 2) L934206
- 3) Case 1
- 4) Carrier 1
- 5) Case 3
- 6) Case 6
- 7) Remote case 1
- 8) Remote case 2
- 9) Carrier 2
- 10) Carrier 3
- 11) Carrier 4

Loci

- abcZ (NEIS1015)
- aceF (NEIS1279)
- acnA (NEIS1729)
- acnB (NEIS1492)
- adk (NEIS0767)
- aroE
- aroE (NEIS1810)
- aspA
- aspA (NEIS1185)

Schemes

- All loci
- Genetic Information Process
- Metabolism
- Typing
 - ☒ MLST
 - ☒ Finetyping antigens
- Antigen genes
- Ribosomal MLST
- eMLST (20 locus partic

Parameters

Min % identity: 70

Min % alignment: 50

BLASTN word size: 15

Return up to: 1 partial match(es)

Stop after: 200 new matches

Stop after: 5 minute(s)

☐ Use TBLASTX

☐ Hunt for nearby start and stop codons

☐ Rescan even if allele designations are already set

☐ Rescan even if allele sequences are tagged

☐ Mark missing sequences as provisional allele '0'

Restrict included sequences by

Sequence method:

Project:

Experiment:

Reset

Scan

Scanning takes approximately one second per locus per genome for a DNA locus, such as the MLST loci. The PorA VR and FetA VR loci are defined by peptide sequence - these take slightly longer to scan since the server has to perform a TBLASTX query to determine these.

16. When scanning is complete, you should see a list of identified allelic matches and ticked checkboxes. These indicate that you will tag these alleles in the database. Click the 'Tag alleles/sequences' button:

Sequence method:
 Project:
 Experiment:

Isolate	Match	Locus	Allele	% identity	Alignment length	Allele length	E-value	Sequence bin id	Start	End	Predicted start	Predicted end	Orientation	Designate allele	Tag sequence	Flag
1) FAM18	exact	abcZ	2	100.00	433	433	0.0	1	1005377	1005809	1005377	1005809	←	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="text"/>
1) FAM18	exact	adk	3	100.00	465	465	0.0	1	784946	785410	784946	785410	→	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="text"/>
1) FAM18	exact	aroE	4	100.00	490	490	0.0	1	1849070	1849559	1849070	1849559	→	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="text"/>
1) FAM18	exact	fumC	3	100.00	465	465	0.0	1	1395712	1396176	1395712	1396176	→	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="text"/>
1) FAM18	exact	gdh	8	100.00	501	501	0.0	1	1317578	1318078	1317578	1318078	←	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="text"/>
1) FAM18	exact	pdhC	4	100.00	480	480	0.0	1	1257934	1258413	1257934	1258413	→	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="text"/>
1) FAM18	exact	pgm	6	100.00	450	450	0.0	1	760937	761386	760937	761386	←	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="text"/>
1) FAM18	exact	PorA VR1	5	100.00	12	12	0.21	1	1357565	1357600	1357565	1357600	←	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="text"/>
1) FAM18	exact	PorA VR2	2	100.00	15	15	0.005	1	1357118	1357162	1357118	1357162	←	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="text"/>
1) FAM18	exact	FetA VR	F1-30	100.00	32	32	5e-13	1	2002034	2002129	2002034	2002129	→	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="text"/>

The database should update and display a confirmation:

Experiment:

Database updated ok.
[Back to main page](#)

Allele designations set

1) FAM18: abcZ: 2
 1) FAM18: adk: 3
 1) FAM18: aroE: 4
 1) FAM18: fumC: 3
 1) FAM18: gdh: 8
 1) FAM18: pdhC: 4
 1) FAM18: pgm: 6
 1) FAM18: PorA VR1: 5
 1) FAM18: PorA VR2: 2
 1) FAM18: FetA VR: F1-30

Allele sequences set

1) FAM18: abcZ: Seqbin id: 1; 1005377-1005809
 1) FAM18: adk: Seqbin id: 1; 784946-785410
 1) FAM18: aroE: Seqbin id: 1; 1849070-1849559
 1) FAM18: fumC: Seqbin id: 1; 1395712-1396176
 1) FAM18: gdh: Seqbin id: 1; 1317578-1318078
 1) FAM18: pdhC: Seqbin id: 1; 1257934-1258413
 1) FAM18: pgm: Seqbin id: 1; 760937-761386
 1) FAM18: PorA VR1: Seqbin id: 1; 1357565-1357600
 1) FAM18: PorA VR2: Seqbin id: 1; 1357118-1357162
 1) FAM18: FetA VR: Seqbin id: 1; 2002034-2002129

17. Repeat scanning and tagging for all other isolates.

You should now be in a position to update the isolate table (page 2). Answer the following questions about the outbreak:

Genome-wide comparison

The information you have extracted so far could have been achieved using conventional MLST and antigen gene sequencing. With the genomes available, however, we can now look at relationships between isolates at a much higher resolution.

The Genome Comparator tool can be used to compare isolates using any sets of loci or against a complete annotated genome. First we will use ribosomal MLST (rMLST).

18. From the database contents page (not curator's interface), click the 'Genome Comparator' link:

Query: Search | Browse | Profile/ST | List
Breakdown: Isolate fields | Scheme/alleles | Publications
Links: Contents | Home | Options | Profiles/sequences definitions | Curate

Welcome to the Neisseria PubMLST database

The Neisseria PubMLST database contains data for a collection of isolates that represent the total known diversity of Neisseria species. For every allelic profile in the profiles/sequence definition database there is at least one corresponding isolate deposited here. Any isolate may be submitted to this database and consequently it should be noted that it does not represent a population sample.

Query database

- Search database - advanced queries.
- Browse database - peruse all records.
- Search by combinations of loci (profiles) - including partial matching.
 - MLST
 - All loci
- List query - find isolates by matching a field to an entered list.

Option settings

- Set general options - including isolate table field handling.
- Set display and query options for locus, schemes or scheme fields.

General information

- Isolates: 12
- Last updated: 2012-09-20
- About BIGSdb

Breakdown

- Single field
- Two field
- Polymorphic sites
- Unique combinations
- Scheme and alleles
- Publications
- Sequence bin
- Tag status

Export

- Export dataset
- Presence/absence status of loci
- Concatenate alleles
- XMFA export

Analysis

- Codon usage
- Genome comparator**
- BLAST

Miscellaneous

- Description of database fields

19. Select all the isolates and choose 'Ribosomal MLST' from the schemes list (this is found within Typing). Click 'submit':

Query: Search | Browse | Profile/ST | List
Breakdown: Isolate fields | Scheme/alleles | Publications
Links: Contents | Home | Options | Profiles/sequences definitions | Curate

Genome Comparator

Please select the required isolate ids and loci for comparison - use ctrl or shift to make multiple selections. In addition to selecting individual loci, you can choose to include all loci defined in schemes by selecting the appropriate scheme description. Alternatively, you can enter the accession number for an annotated reference genome and compare using the loci defined in that.

Isolates

- 5) Case 3
- 6) Case 6
- 7) Remote case 1
- 8) Remote case 2
- 9) Carrier 2
- 10) Carrier 3
- 11) Carrier 4
- 12) Carrier 5

Loci

- abcZ (NEIS1015)
- aceF (NEIS1279)
- acnA (NEIS1729)
- acnB (NEIS1492)
- adk (NEIS0767)
- aroE

Schemes

- metabolism
- Typing
- Finely typing antigens
- Antigen genes
- Ribosomal MLST**
- eMLST (20 locus parti...
- eMLST (20 locus whol...

Reference genome

Enter accession number:

or choose annotated genome:

or upload Genbank/EMBL file:

Parameters / options

Min % identity:

Min % alignment:

BLASTN word size:

☒ Use TBlastX

☐ Produce alignments (Clustal + XMFA)

☐ Include ref sequences in alignment

☐ Align all loci (not only variable)

☒ Use tagged designations if available

☐ Disable HTML output

Restrict included sequences by

Sequence method:

Project:

Experiment:

20. The analysis will be submitted to the job queue. The status of the job and the results can be viewed by clicking the link:

Query: Search | Browse | Profile/ST | List
Breakdown: Isolate fields | Scheme/alleles | Publications
Links: Contents | Home | Options | Profiles/sequences definitions | Curate

Genome Comparator

This analysis has been submitted to the job queue.

Please be aware that this job may take a long time depending on the number of comparisons and how busy the server is.

[Follow the progress of this job and view the output](#)

Analysis will take a few minutes (this would be quicker if the rMLST loci had been tagged previously, but as they haven't been they need to be scanned):

PubMLST Query: Search | Browse | Profile/ST | List
Breakdown: Isolate fields | Scheme/alleles | Publications
Links: Contents | Home | Options | Profiles/sequences definitions | Curate

Job status viewer

Status

Job id: BIGSdb_5904_1348129414_81767
Submit time: 2012-09-20 09:23:34
Status: started
Start time: 2012-09-20 09:24:01
Progress: 5%
Stage: Analysing locus: BACT000004
Elapsed time: 1 minute and 21 seconds

Output

Analysis against defined loci

Allele numbers are used where these have been defined, otherwise sequences will be marked as 'New#1', 'New#2' etc. Missing alleles are marked as 'X'. Truncated alleles (located at end of contig) are marked as 'T'.

Locus	1 (FAM18)	2 (L93/4286)	3 (Case 1)	4 (Carrier 1)	5 (Case 3)	6 (Case 6)	7 (Remote case 1)	8 (Remote case 2)	9 (Carrier 2)	10 (Carrier 3)	11 (Carrier 4)	12 (Carrier 5)
BACT000001 (rpsA)	3	2	2	2	2	2	2	2	2	2	2	2
BACT000002 (rpsB)	3	3	3	3	3	3	3	3	11	24	3	3
BACT000003 (rpsC)	3	3	3	3	3	3	3	3	3	3	3	3

This page will reload in 1 minute and 20 seconds. You can refresh it any time, or bookmark it and close your browser if you wish.
Please note that job results will remain on the server for 7 days.

When the analysis completes, a table showing the alleles at each of the rMLST loci will be displayed along with a NeighborNet network. From the network answer the following questions:

Now we will analyse the isolates against a complete annotated genome.

- Repeat the Genome Comparator analysis but this time select the FAM18 genome from the dropdown box (deselect FAM18 from the list of isolates). This will compare every coding sequence in the FAM18 genome against each of the genomes of the comparator isolates to produce a higher resolution network:

PubMLST Query: Search | Browse | Profile/ST | List
Breakdown: Isolate fields | Scheme/alleles | Publications
Links: Contents | Home | Options | Profiles/sequences definitions | Curate

Genome Comparator

Please select the required isolate ids and loci for comparison - use ctrl or shift to make multiple selections. In addition to selecting individual loci, you can choose to include all loci defined in schemes by selecting the appropriate scheme description. Alternatively, you can enter the accession number for an annotated reference genome and compare using the loci defined in that.

Isolates
1) FAM18
2) L93/4286
3) Case 1
4) Carrier 1
5) Case 3
6) Case 6
7) Remote case 1
8) Remote case 2

Reference genome
Enter accession number:
or choose annotated genome:
FAM18 (Nm)
or upload Genbank/EMBL file:

Parameters / options
Min % identity: 70
Min % alignment: 50
BLASTN word size: 15
☐ Use TBLASTX
☐ Produce alignments (Clustal + XMF)
☒ Include ref sequences in alignment
☐ Align all loci (not only variable)
☒ Use tagged designations if available
☐ Disable HTML output

Restrict included sequences by
Sequence method:
Project:
Experiment:

Section 7 Mycobacteria

7.1 Molecular diagnostics for identification and monitoring of mycobacterial infections

The major challenge for diagnosis of tuberculosis is the slow rate of growth of the organism in conventional and automated culture. There are numerous examples of transmission of the infection among vulnerable patients as a result of failure to detect tuberculosis rapidly. Stained smears of infected material (Ziehl-Neelsen or auramine) are, in contrast, rapid but lack both sensitivity and specificity. For many years, culture has been considered as the gold standard i.e., reference method for tuberculosis diagnosis. Positive culture results may be available in as few as five days in patients with a high bacterial load, but the negative results (from liquid culture) can only be finally reported more than six weeks after sample receipt. Drug susceptibility requires an additional week at a minimum and may require significantly longer time if the organism is in low concentration. Despite the disadvantage of being slow the limit of detection for culture is low translating into optimal sensitivity. Consequently, it remains the cardinal diagnostic method.

There has been a longstanding endeavour to develop rapid molecular based assays.

7.1.1 Detection of Nucleic Acid Species

DNA

Multiple methods of detecting mycobacterial DNA in clinical samples have been reported. The recent release of the GeneXpert® system for tuberculosis (Xpert MTB/RIF; Cepheid, Sunnyvale, CA) has raised considerable interest with regard to its potential use as a biomarker of treatment response. The results of recent clinical trials using this method for tuberculosis diagnosis are discussed in detail below. However, it should be noted that older studies suggest that mycobacterial DNA remains in the sputum for a considerable period of time after initiation of therapy and this may compromise the value of this technique.

RNA

RNA can be used for detection and quantification of *M. tuberculosis* in clinical samples (Honeyborne *et al.*, 2011). Further details are described in section 7.1.4.

7.1.2 Cepheid GeneXpert

The Cepheid GeneXpert® System is a fully-automated and integrated system for Polymerase Chain Reaction (PCR)-based nucleic acid (DNA) testing. There are 11 FDA-cleared assays and 12 CE-IVD marked assays that use the Xpert platform, including tests for enteroviral meningitis, methicillin-resistant *Staphylococcus aureus* (MRSA), Influenza A and B, group B *Streptococcus*, and anthrax, in addition to the assay for *Mycobacterium tuberculosis* and rifampin resistance detection. The tuberculosis assay consists of a microfluidic cartridge with multiple chambers that contains all necessary reagents and diluents for sample processing, nucleic acid amplification, and detection of products. The respiratory sample, after treatment with sample buffer, is added with a pipette into one chamber of the cartridge, the top is closed, and the

cartridge is placed into the assay module in which all subsequent testing steps occur (see Figure 1). The reaction tube, unlike the cylindrical shape of most PCR reaction tubes, is a flat, window-like shape, allowing efficient and rapid heating and cooling by the solid-state heating element present in the assay module. Liquids move between chambers through the action of a rotary valve and a central syringe-like plunger. The PCR reactants are contained in freeze-dried beads, which are dissolved only when the liquid enters the appropriate chamber. Internal software controls monitor pressure, reagent presence, inhibition, and bead activity. The ease of use allows non-technical workers to perform the assay with minimal training. This concept has been validated in a study conducted by Boehme and others.

Prior to initiating the Xpert MTb/RIF assay, the respiratory sample is mixed with a sample reagent at twice the volume of the sputum and shaken vigorously; this liquefies mucus and is critical for the mycobacteria. Two mls of the suspension is added to the cartridge; this large sample volume increases the sensitivity of the assay. A second factor in enhanced sensitivity is the action of a glass-bead enhanced sonication of the mycobacteria, which are fixed on a filter at the base of the cartridge, efficiently releasing DNA. The DNA passes through the filter into the reaction buffer chamber. A third factor that enhances assay sensitivity is the use of a hemi-nested PCR amplification protocol. The first reaction amplifies the specific *rpoB* genetic sequence and the second PCR reaction amplifies each of 5 smaller internal fragment. The 6-color molecular beacon assay was designed by Dr. David Alland and Dr. David Persing. From start to finish, results are available within 2 hours of sample receipt, which includes approximately 5 minutes of hands-on manipulation. The PCR amplifies an 81 base-pair region of the ribosomal polymerase B gene, which is unique to *M. tuberculosis* complex. The majority of mutations that confer rifampin resistance are in this region. Five molecular beacons, each with a different fluorophore, bind to independent regions of the amplicon. A sixth target identifies the internal control, a *Bacillus globigii* spore which indicates any potential inhibition in the reaction. Similar cycle threshold amplification of all five target sequences signifies a wild-type, rifampin-susceptible *M. tuberculosis* complex organism, whereas delayed or missing signals of one to three targets signifies one or more mutations conferring rifampin resistance. The reaction is semi-quantitative, so that higher cycle threshold levels suggest lower numbers of organisms in the sample. For the CE-IVD version, the results for *M. tuberculosis* detection are reported as “high, moderate, low, or very low.” The quantitative estimation ability of the assay has been reported anecdotally in at least 6 publications, and Blakemore et al. have suggested an algorithm for incorporating the amount of inhibition into the result.

Several publications have described the development of the assay, sensitivity of the assay in samples from children, and the biosafety features of the sample preparation and handling procedures. The first large-scale assessment of the assay was conducted by the Foundation for Innovative New Diagnostics (FIND) in five sites with relatively high numbers of HIV-positive patients and multi-drug resistant TB (MDR-TB) patients: Peru, Azerbaijan, India, and South Africa. Results from specimens collected from more than 1700 patients were evaluated and compared with the results of acid fast smear, liquid culture, and mycobacterial drug susceptibility tests. A subsequent study by FIND evaluated the performance of the assay in a more rural setting with inexperienced laboratory workers. Posters presented at several international meetings have also reported on use of the assay in large-scale testing situations (Naidoo, others). At least one study showed that the assay failed to detect some positive samples that a laboratory-developed molecular assay was able to detect. A number of studies have evaluated the use of the Xpert assay for samples other than respiratory secretions, with varying levels of success.

Several features of the Xpert® Mtb/RIF assay differ from existing molecular diagnostic methods. The most commonly used nucleic acid amplification test (NAAT) in the United States is the GenProbe MTD assay, which targets and amplifies ribosomal RNA rather than DNA, but does not have a rifampin resistance detection component. The GenProbe assay has been reported to be 97% sensitive and 96% specific for smear-positive samples, and 76% sensitive and 97% specific for smear-negative samples. This test is highly complex, takes about 2 hours of hands-on time, and delivers results in 6 hours. Roche also has two NAATs for *M. tuberculosis*, which are not currently available in the United States. Published results show 96% sensitivity and 74-83% specificity in smear-positive patient samples and 76% sensitivity and 97% specificity for smear-negative samples. BD Probe-Tec shows similar results as the other two NAAT products in a meta-analysis based on 12 studies. The ligase-chain reaction product from Abbott (10 studies) actually yielded slightly worse performance, with 96% sensitivity and 71% specificity for smear-positive samples and 57% sensitivity and 98% specificity for smear-negative samples and has since been withdrawn from the market.

In comparison with acid fast smear results, the Xpert Mtb/RIF assay is more sensitive and the results are delivered within the same time frame, i.e., approximately two hours. All other types of methods require longer time periods to produce results, and in rural settings, many final results never reach the intended user. Training to perform the Xpert assay is more successful than training to read TB smears. Because extensive infrastructure, biosafety equipment, and special environments are not needed for performance of GeneXpert, the startup costs and continuing resources (including maintaining proficiency of workers) are lower than for development of liquid culture capability for a new testing setting. With regard to clinical endpoints, the implications of positive PCR results after or even during therapy have not yet been determined. Studies have shown that even if samples yield smear-negative results, it does not mean that the patient cannot transmit tuberculosis to others. Another recent study using a guinea pig sentinel system has confirmed that patients who are smear-negative may still be infectious, whereas those patients on effective therapy, despite having positive smears, are not likely to be infectious. Moreover, patients with MDR TB being treated presumptively with conventional front-line TB regimens have been shown to be infectious in this model.

The Cepheid GeneXpert System (Xpert) is a fully automated and integrated system for Polymerase Chain Reaction (PCR)-based nucleic acid (DNA) testing. Xpert targets the *rpob* hotspot region of *M. tuberculosis* DNA providing simultaneous speciation and identification of rifampicin resistance. It requires minimal hands on time. The ease of use allows non-technical workers to perform the assay with minimal training. This makes it an attractive approach in countries where it is difficult to recruit the experienced laboratory staff needed to perform smear and culture examination effectively. It has already been shown to be more sensitive than conventional stained smear examination in clinical studies in high burden countries and has been implemented widely in South Africa replacing smear in the national tuberculosis programme. The full test characteristics are still to be evaluated in communities where tuberculosis is relatively uncommon but here there is an advantage for laboratories who have a low throughput of tuberculosis tests in which there is a risk that laboratory skills will have declined.

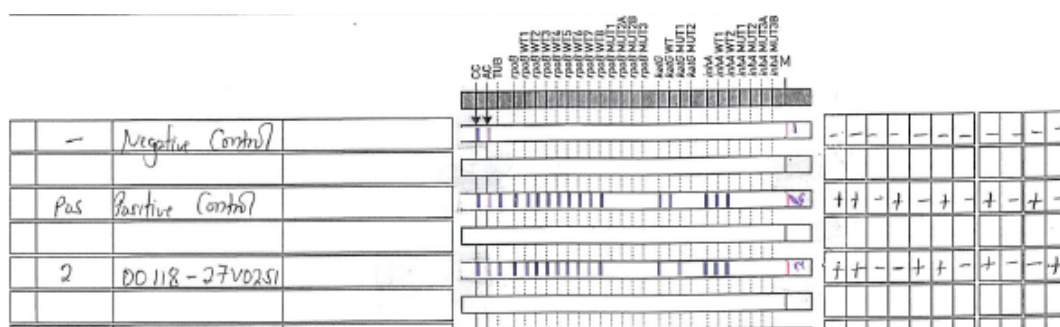
One limitation of DNA tests like GeneXpert® is that they potentially detect both dead and viable organisms at the same level of sensitivity. It should be noted that older studies suggest that

mycobacterial DNA remains in the sputum for a considerable period of time after initiation of therapy and this may compromise the value of this technique as a way of monitoring therapy.

The GeneXpert® Mtb/RIF assay could be used to determine the infection status of patients rapidly for entry into a clinical trial. The semi-quantitative nature of the results suggests the potential to use the assay to monitor a patient's response to treatment (see below). A current clinical trial comparing patients treated during the intensive phase of therapy with either rifampin or rifapentine (CDC-TB Trials Consortium Study 29X trial) has been expanded to include evaluation of the Xpert Mtb/RIF assay results at several time points to determine the feasibility of this approach. Poor performance of GeneXpert® for monitoring sequential data has been confirmed in a recent publication from the REMoxTB consortium that that GeneXpert® ct values responded less rapidly than smear.

7.1.3 Hain system

An alternative to GeneXpert® is the line probe assay such as the Hain assay which detects amplicons specific to *M. tuberculosis* and some of the drug resistance genes by PCR and hybridisation of the amplicons to a nitrocellulose strip. Although technically straightforward to perform care must for the test steps to be performed consistently if the results are to be reproducible. The test strips, compared with control pattern, can be photographed to maintain a record as the bands fade on storage. Isoniazid rifampicin resistance are determined using the standard strips, and extended resistance testing with Hain S⁺ is available giving it an advantage of coverage over the geneXpert system. This system has proved valuable in the REMoxTB clinical trial in areas of high MDRTB prevalence.



Follow up visit with the patient complaining of symptoms returning

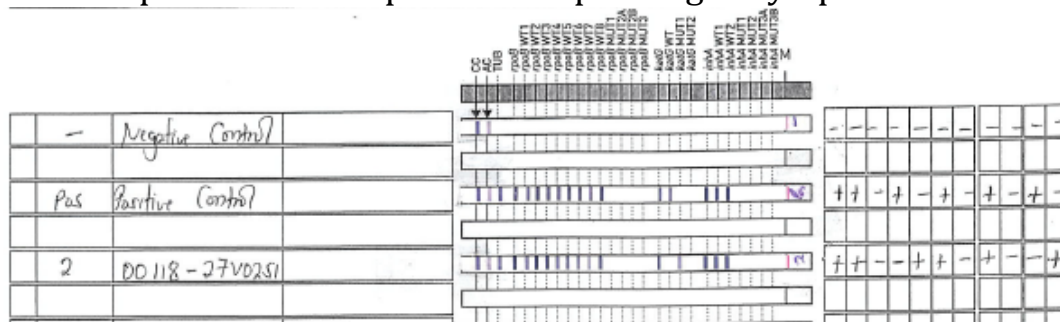


Figure 7.1.1 An example of Hain test results in a patient being treated for tuberculosis. Screening visit

7.1.4 Molecular Bacterial Load Assay

The Molecular Bacterial Load (MBL) assay monitors the molecular load of TB cells in clinical or model samples and hence provides accurate information on bacterial response (decline) to antimicrobial treatment.

Introduction

The MBL assay detects *M. tuberculosis* 16S ribosomal RNA, which is a relatively stable RNA species and after its extraction from sputum can provide the information on bacterial viable counts. The assay involves internal control that monitors for the RNA extraction efficiency and amplification performance. Internal control (IC) is an exogenous RNA template added to each specimen before the RNA extraction. The IC can be prepared as described in Honeyborne I. *et al.*, J Clin Microbiol, 2011. The TB and IC detection is based on a quantitative duplex real-time PCR and the identification of a specific fluorescent signal.

Materials and Methods

General consumables

- sterile DNase and RNase-free microtubes, 1.5 ml
- pipettes and sterile pipette tips, DNase and RNase-free, range: P1000, P200, P10, P2
- sterile disposable universal tubes
- PCR reaction tubes – use accordingly with the PCR instrument and the total number of samples to be analysed. For RotorGene Q (Qiagen) use:
 - Single 0.2 ml PCR optical thin wall flat cap microtubes (any supplier), for 32-well rotor
 - Strip tubes and caps, 0.1 ml (Qiagen), for 72-well rotor
 - Rotor-Disc 100 with heat-seal film (Qiagen), for 100-well rotor

Instruments

- Centrifuge and benchtop centrifuge
- Thermoshaker
- Homogenizer (Fastprep, MP Biomedical or Precellys, Peqlab)
- Real-time PCR thermocycler (RotorGene Q, Qiagen)

7.1.4.1 RNA Preservation

The integrity of TB RNA in sputum samples can be maintained by preservation of the sputum in GTC solution immediately after expectoration, and storing at -80°C until used for the RNA extraction.

RNA preservation solution: 4 M GTC (Promega), 0.1 M HCl (pH 7.0) and 1% β -mercaptoethanol

7.1.4.2 RNA extraction

We recommend using of either of two simple and rapid extraction procedures for the total RNA extraction. By this procedure, samples are lysed and homogenized in the presence of guanidinium isothiocyanate, a chaotropic salt capable of protecting the RNA from endogenous RNases. After lysis and homogenization and alcohol precipitation, the sample is processed through a spin column with silica-based membrane to which the RNA binds. Any impurities are removed by subsequent washing and the resulting RNA is eluted into water or a buffer.

For FastRNA Pro Blue kit (MP Biomedicals follow the manufacturer's procedure.

For Purelink RNA (Ambion) extract the RNA according to following procedure:

Reagents

RNA PureLink kit (Ambion)

TURBO DNase I digest kit (Ambion)

Glass beads, 0.1 mm (e.g. PEQlab or Sigma), 0.65 g per 2 ml homogenisation tube (weighed in homogenisation 2 ml tubes).

β -mercaptoethanol

70% ethanol prepared fresh before the extraction

RNase free water

IC RNA, 50 ng per RNA preparation

Procedure

Before starting, prepare fresh Lysis Buffer containing Purelink Lysis Buffer, 1% β -mercaptoethanol and internal control (IC) RNA for each purification procedure (the internal control should be approximately constant for all the RNA readings). I.e. for 600 μ l of Lysis Buffer per each sample, add into 594 ml of Purelink Lysis Buffer a volume of 6 μ l β -mercaptoethanol.

1. Thaw the sputum samples in GTC on ICE and keep the tubes on ice up to the point of adding the lysis buffer.
2. Following thawing, add into each sputum sample in GTC 50 ng of internal control.
3. Spin the tubes at 2,000 $\times g$ for 30 min at room temperature.
4. Discard the supernatant and add 600 μ l Lysis Buffer (supplemented with β -mercaptoethanol) to the pellet and agitate to detach the pellet from the tube. Work at room temperature further on.
5. Transfer the suspension into homogenisation tube and process in a homogeniser. FastPrep setting 6, 40 s or Precellys 30 s at 6000 rpm.
6. Spin the homogenisation tubes at 12000 $\times g$ for 5 min at room temperature.
7. Remove 400 μ l of the homogenate without touching or transferring any glass beads, and add the homogenate to a 1.5 ml RNase free microtube with one volume (400 μ l) of 70% ethanol. Mix well by vortexing. Note: If part of the sample was lost during homogenization, adjust the volume of ethanol accordingly. Briefly spin to remove any drops from the lid.
8. Transfer up to 500 μ l of the sample (including any remaining precipitate) to the Spin Cartridge inserted in a Collection Tube.
9. Centrifuge at 12,000 $\times g$ for 15 seconds at room temperature. Discard the flow-through, and reinsert the Spin Cartridge in the same Collection Tube.

10. Repeat until the entire sample is processed.
11. Add 700µl Wash Buffer I to the Spin Cartridge containing the bound RNA. Centrifuge at $12,000 \times g$ for 15 seconds at room temperature. Discard the flow-through and the Collection Tube. Insert the Spin Cartridge into a new Collection Tube.
12. Add 500 µl Wash Buffer II (with ethanol added) to the Spin Cartridge. Centrifuge at $12,000 \times g$ for 15 seconds at room temperature. Discard flow-through and reinsert the Spin cartridge into the same collection tube.
13. Repeat Step the previous washing step once to remove any residual ethanol.
14. Centrifuge the Spin Cartridge at $12,000 \times g$ for 1 minute to dry the membrane with bound RNA. Discard Collection Tube and insert the Spin Cartridge into a Recovery Tube.
15. Add 50µl (or 30µl–100µl) RNase-Free Water to the centre of the Spin Cartridge.
16. Incubate at room temperature for 1 minute.
17. Centrifuge Spin Cartridge and Recovery Tube for 1 minute at $\geq 12,000 \times g$ at room temperature.
18. Store purified RNA at -80°C or proceed directly to DNase treatment.

DNase treatment

1. Add 0.1 volume of 10× Turbo DNase buffer (5 µl) and 1 µl (2U) Turbo DNase to each RNA sample. Mix gently.
2. Incubate at 37°C , 30 min, 1000 rpm.
3. Add again 1 □l (2U) Turbo DNase and incubate for another 30 min (can be longer).
4. Add 2□l homogenized (thawed & vortexed) DNase inactivation reagent. Incubate at room temperature for 5 min and up to 1000 rpm.
5. Centrifuge at $10,000 \times g$ for 1.5 min and transfer the RNA (leaving the pellet in the tube) into a new RNase free tube.
6. Store the RNA at -80°C .

7.1.4.3 Quantitative RT-PCR reaction.

A one step RT real-time PCR is used for the RNA quantification. In one step and one reaction tube, the RNA is reverse transcribed to cDNA and this serves as template in the real-time PCR.

Considerations before start

Use RT-PCR mix without any Rox dye (e.g. QuantiTect Multiplex RT-PCR NR Kit, cat. no. 204843, Qiagen) for the use with RotorGeneQ, ViiA7 and CFX96 instruments, respectively. The ABI7300 instrument requires the use of a mastermix containing the Rox dye (e.g. QuantiTect Multiplex RT-PCR Kit, cat. no. 204643, Qiagen).

Aliquots of 500 µl multiplex NR RT-PCR mix (Qiagen) are prepared and stored at -20°C . Paired forward and reverse primers are mixed in a single tube to the final concentration of 20 µM and stored at -20°C . Probes are diluted to 20 µM and aliquots are stored in -20°C . Aliquots of RT-PCR mix and probe are for single use only.

All oligonucleotides are of HPLC quality.

Water is of PCR grade and RNase-free.

Work on ice when preparing the PCR mastermix and individual reactions. Thaw and keep the RNA on ice.

Use the RNA template neat and as 1/10 dilution

Every sample, including the neat and diluted RNA and the dilutions for the standard curve, is analyzed in duplicate.

Reagents

QuantiTect Multiplex RT-PCR NR Kit

RNase free water

Primers for TB and IC

Fluorescently labeled TaqMan probes for TB (FAM labelled) and IC (HEX labelled)

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all times during the practical

Procedure

1. Prepare the PCR mastermix for all samples to be analysed according to Table 1 and calculate for 10% extra volume in order to prevent pipetting errors. Vortex the mastermix and briefly spin down.
2. Distribute 8 µl of the mastermix into each reaction tube.
3. Add 2 µl of RNA template (neat and diluted) into each tube. Add 2 µl TB DNA (positive control) and 2 µl nuclease-free water (negative control). Analyze each sample in duplicate.
4. Run the PCR in the real-time PCR instrument, set the thermal cycling conditions as detailed in Table 2.

Standard curve construction.

Use the RNA extracted from a standard material - *M. tuberculosis* culture with estimated concentration of 10⁸ CFU/ml or higher. In detail, dilute the extracted RNA decimally to create a series of standards. Prepare 7 decimal dilutions and use them together with the neat RNA to construct the standard curve for PCR quantification.

The standard curve can be prepared in a separate run for the use with RotorGene Q and it can be further incorporated in data analysis of samples with unknown bacterial load.

The PCR efficiency can be evaluated by the parameters of standard curve. The equation for an ideal standard curve and 100% amplification efficiency (E=1) is:

$Ct = \text{slope} \times \text{Log}(\text{concentration}) - \text{intercept}$

or

$Ct = -3.32 \times \text{Log}(\text{concentration}) - \text{intercept}$

Aim for the efficiency of 90%-100%, i.e. $E=0.9$ to 1.0 . The efficiency can be calculated from the slope of the standard curve using the equation:

$$E=10^{-1/-3.32 \cdot \text{slope}} - 1$$

Table 7.1.4.1 PCR reaction composition.

Reagent	Stock concentration	Reaction concentration	Volume per reaction (μl)	Volume per n reactions (μl)
RNA template			2	
IC Forward + Reverse primer	20 μM	200 nM	0.1	
TB 16S Forward + Reverse primer	20 μM	200 nM	0.1	
IC probe (HEX-BHQ1)	20 μM	200 nM	0.1	
TB 16S probe (FAM-BHQ1)	20 μM	200 nM	0.1	
Quantitect RT-mix	2×	1×	5	
Quantitect RT enzyme			0.1	
RNase free water			2.5	
Total			10	

Table 7.1.4.2 Thermal cycling programme.

Hold	50°C	20 min	
Hold	95°C	15 min	
40 cycles	94°C	45 s	
	60°C	45 s	Acquiring fluorescent signal for FAM and HEX

Data analysis

The TB-specific amplification is detected by the fluorescent signal in the FAM channel. The internal control amplification signal is detected in VIC (or HEX) channel.

Set the fluorescence threshold.

Import the standard curve for TB and IC. Using the instrument's software, determine the CT values and sample concentrations.

Calculate the TB concentration in sputum.

7.1.4.4 Additional reading related to molecular diagnosis of *M. tuberculosis*

Ahmad S, Al-Mutairi NM, Mokaddas E. Comparison of performance of two DNA line probe assays for rapid detection of multidrug-resistant isolates of *Mycobacterium tuberculosis*. *Indian J Exp Biol* 2009;47:454-62.

Armand S, Vanhuls P, Delcroix G, Courcol R, Lemaitre N. Comparison of the Xpert MTB/RIF test with an IS6110-TaqMan real-time PCR assay for direct detection of *Mycobacterium tuberculosis* in respiratory and nonrespiratory specimens. *J Clin Microbiol* 2011;49:1772-6.

Banada PP, Sivasubramani SK, Blakemore R, et al. Containment of bioaerosol infection risk by the Xpert MTB/RIF assay and its applicability to point-of-care settings. *J Clin Microbiol* 2010;48:3551-7.

Blakemore R, Nabeta P, Davidow AL, et al. A Multi-Site Assessment of the Quantitative Capabilities of the Xpert(R) MTB/RIF Assay. *Am J Respir Crit Care Med* 2011.

Blakemore R, Story E, Helb D, et al. Evaluation of the analytical performance of the Xpert MTB/RIF assay. *J Clin Microbiol* 2010;48:2495-501.

Boehme CC, Nicol MP, Nabeta P, et al. Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study. *Lancet* 2011;377:1495-505.

Boehme CC, Nabeta P, Hillemann D, et al. Rapid molecular detection of tuberculosis and rifampin resistance. *N Engl J Med*;363:1005-15.

Ciftci IH, Aslan MH, Asik G. [Evaluation of Xpert MTB/RIF results for the detection of *Mycobacterium tuberculosis* in clinical samples]. *Mikrobiyol Bul* 2011;45:43-7.

El-Hajj HH, Marras SA, Tyagi S, Kramer FR, Alland D. Detection of rifampin resistance in *Mycobacterium tuberculosis* in a single tube with molecular beacons. *J Clin Microbiol* 2001;39:4131-7.

Friedrich SO, Venter A, Kayigire XA, Dawson R, Donald PR, Diacon AH. Suitability of Xpert MTB/RIF and Genotype MTBDRplus for Patient Selection for a Tuberculosis Clinical Trial. *J Clin Microbiol* 2011;49:2827-31.

Sven O Friedrich, Andrea Rachow, Elmar Saathoff, Kasha Singh, Chacha D Mangu, Rodney Dawson, Patrick P J Phillips, Amour Venter, Anna Bateson, Catharina C Boehme, Norbert Heinrich, Robert D Hunt, Martin J Boeree, Alimuddin Zumla, Timothy D McHugh, Stephen H Gillespie, Andreas H Diacon, Michael Hoelscher, on behalf of the Pan African Consortium for the Evaluation of Anti-tuberculosis Antibiotics (PanACEA) Assessment of the sensitivity and specificity of Xpert MTB/RIF assay as an early sputum biomarker of response to tuberculosis treatment *Lancet Respiratory Medicine* 2013; 1: 462–70

Helb D, Jones M, Story E, et al. Rapid detection of *Mycobacterium tuberculosis* and rifampin resistance by use of on-demand, near-patient technology. *J Clin Microbiol*;48:229-37.

Helb D, Jones M, Story E, et al. Rapid detection of *Mycobacterium tuberculosis* and rifampin resistance by use of on-demand, near-patient technology. *J Clin Microbiol* 2009;48:229-37.

Hillemann D, Rusch-Gerdes S, Boehme C, Richter E. Rapid molecular detection of extrapulmonary tuberculosis by the automated GeneXpert MTB/RIF system. *J Clin Microbiol* 2011;49:1202-5.

Honeyborne I, McHugh TD, Phillips PP, Bannoo S, Bateson A, Carroll N, Perrin FM, Ronacher K, Wright L, van Helden PD, Walzl G, Gillespie SH. Molecular bacterial load assay, a culture-free biomarker for rapid and accurate quantification of sputum *Mycobacterium tuberculosis* bacillary load during treatment. *J Clin Microbiol*. 2011 Nov;49(11):3905-11.

Ioannidis P, Papaventsis D, Karabela S, et al. Cepheid GeneXpert MTB/RIF Assay for *Mycobacterium tuberculosis* Detection and Rifampin Resistance Identification in Patients with Substantial Clinical Indications of Tuberculosis and Smear-Negative Microscopy Results. *J Clin Microbiol* 2011;49:3068-70.

Kennedy N, Gillespie SH, Saruni AO, et al. Polymerase chain reaction for assessing treatment response in patients with pulmonary tuberculosis. *J Infect Dis* 1994;170:713-6.

Ling DI, Flores LL, Riley LW, Pai M. Commercial nucleic-acid amplification tests for diagnosis of pulmonary tuberculosis in respiratory specimens: meta-analysis and meta-regression. *PLoS One* 2008;3:e1536.

Lawn SD, Mwaba P, Bates M, et al. Advances in tuberculosis diagnostics: the Xpert MTB/RIF assay and future prospects for a point-of-care test. *The Lancet Infectious Diseases* 2013; **13**: 349–61.

Marlowe EM, Novak-Weekley SM, Cumpio J, et al. Evaluation of the Cepheid Xpert MTB/RIF assay for direct detection of *Mycobacterium tuberculosis* complex in respiratory specimens. *J Clin Microbiol* 2011;49:1621-3.

Miller MB, Popowitch EB, Backlund MG, Ager EP. Performance of Xpert MTB/RIF RUO Assay and IS6110 Real-Time PCR for *Mycobacterium tuberculosis* Detection in Clinical Samples. *J Clin Microbiol* 2011.

Nicol MP, Workman L, Isaacs W, et al. Accuracy of the Xpert MTB/RIF test for the diagnosis of pulmonary tuberculosis in children admitted to hospital in Cape Town, South Africa: a descriptive study. *Lancet Infect Dis* 2011.

Picard FJ, Gagnon M, Bernier MR, et al. Internal control for nucleic acid testing based on the use of purified *Bacillus atrophaeus* subsp. *globigii* spores. *J Clin Microbiol* 2009;47:751-7.

Rachow A, Zumla A, Heinrich N, et al. Rapid and accurate detection of *Mycobacterium tuberculosis* in sputum samples by Cepheid Xpert MTB/RIF assay--a clinical validation study. *PLoS One*;6:e20458.

Vadwai V, Boehme C, Nabeta P, Shetty A, Alland D, Rodrigues C. Xpert MTB/RIF: a new pillar in diagnosis of extrapulmonary tuberculosis? J Clin Microbiol 2011;49:2540-5.

7.2 IS 6110 DNA fingerprinting overview

DNA fingerprinting of *Mycobacterium tuberculosis* has been shown to be a powerful epidemiological tool. Standardisation of this technique was proposed in 1993 by Van Embden *et al* to exploit variability in both the number and genomic position of the insertion site (IS) 6110 (IS6110) to generate strain-specific patterns. The insertion sequence IS6110 is found in almost all isolates of *M. tuberculosis* and is present in one to approximately 26 copies. The variability in the position of these insertion sequences provides the basis of a Restriction Fragment Length Polymorphism (RFLP) typing method. A standardised method permits results to be compared between different laboratories and therefore allows investigation of the international transmission of tuberculosis and identification of specific strains with unique properties such as high infectivity, virulence, or drug resistance, as well as monitoring cross contamination (Ruddy *et al.*, 2002).

7.3 Repeat based typing: Variable Number Tandem Repeats (VNTR) overview

The IS6110 fingerprinting technique is based on electrophoretic separation followed by probing of largely intact whole genomic DNA, it is a relatively cumbersome method requiring a high concentration of high quality DNA that must be isolated from pure cultures. It is also cumbersome for large numbers of strains and requires experienced staff both to ensure DNA quality and to perform the electrophoresis. Standardisation between gels can be very challenging when large numbers of samples are to be compared. Importantly, a significant number of strains of *M. tuberculosis* (5-40% depending on location) carry few or no copies of IS6110, which renders this method of little value.

More rapid alternatives to IS6110 fingerprinting have been developed. Spoligotyping has been used for a number of years (Goyal *et al*, 1997). In 1998 Frothingham *et al* introduced Variable Number Tandem Repeat (VNTR) analysis and Philip Supply introduced a newer version called Mycobacterial Interspersed Repeat Unit (MIRU) analysis in 2000. Both of these are methods are PCR-based which means they can be performed from broth cultures, such as BACTEC/MGIT, as soon as they are positive. It is important to remember that both VNTR and MIRU analysis are based on the same or very similar elements (figure 2.). Indeed, some of the elements overlap in the different schemes. They are highly reproducible and produce digital results that are easier to compare than the IS6110 patterns (figure 3). The *M. tuberculosis* genome contains 41 loci with direct tandem repeats of 50-70 base pairs (figure 4). The number of repeats per locus varies between strains and strains can be typed based on the number of repeats per locus. Variation in the number of repeats is thought to be due to strand slippage during replication (figure 5). Variation in the IS6110/MIRU pattern occur at different rates depending on the differing molecular clock (McHugh *et al.*, 2005) VNTR variation is associated with function in various bacterial species and includes surface protein variation/regulation and phase variation or 'switching':

Surface protein variation/regulation

- **Opa genes**
 - *Neisseria gonorrhoeae*
- **Porin gene regulation**
 - *Neisseria meningitidis*
- **Emm gene**

- Group A streptococci
- **Sheath protein**
 - *Bacillus anthracis*

Phase variation

- **Lipopolysaccharide**
 - *Haemophilus influenzae*
 - *Neisseria gonorrhoeae*
- **Fimbriae**
 - *Haemophilus*

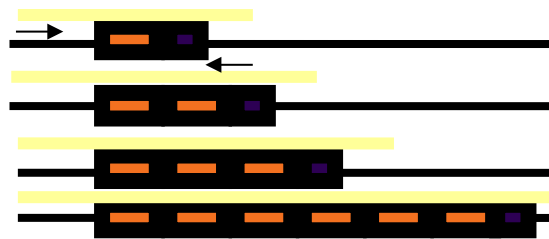


Figure 7.3.1 VNTR structure and variation in repeat copy number between strains. This region can be amplified by PCR using primers (arrows) outside of the repeat region and the size of the PCR product, shown above each example, corresponds to the number of repeats. So, for MIRU 4 (the number refers to the specific locus, not to the number of alleles) the allele numbers would be 1, 2, 3, and 6 for these examples. Combining the results for all 12 MIRU loci yields a 12-digit MIRU designation.

MIRU	02	04	10	16	20	23	24	26	27	31	39	40
Size (bp)	283	332	201	168	365	287	414	303	226	264	226	220
No. of copies	2	3	2	2	3	4	2	5	3	3	2	2

MIRU pattern: 232234253322

Figure 7.3.2 Example of digital result for MIRUs.

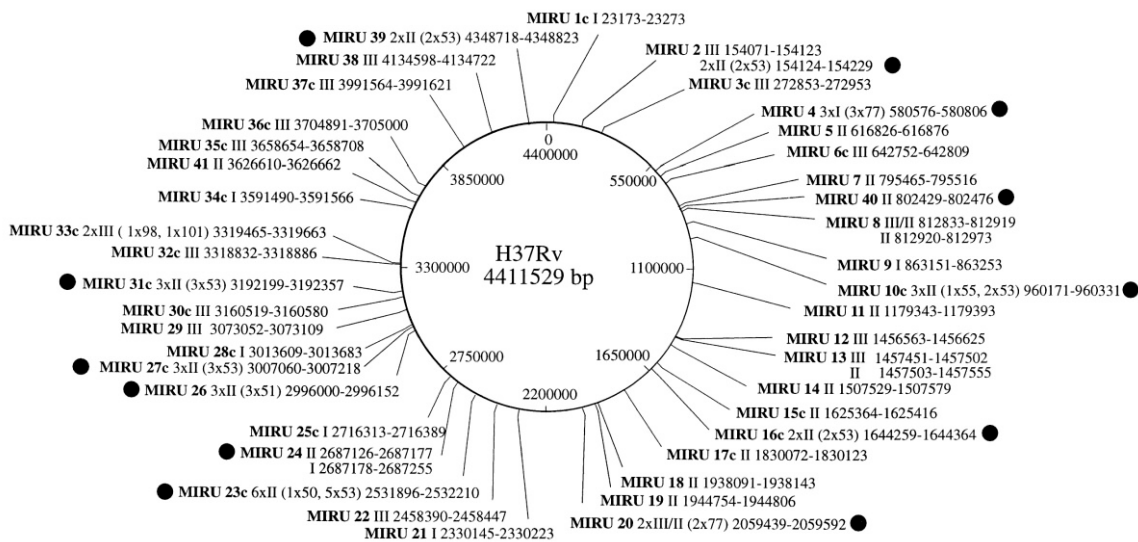


Figure 7.3.3 MIRU Locations in the Mtb genome (taken from Supply *et al.*, 2000)

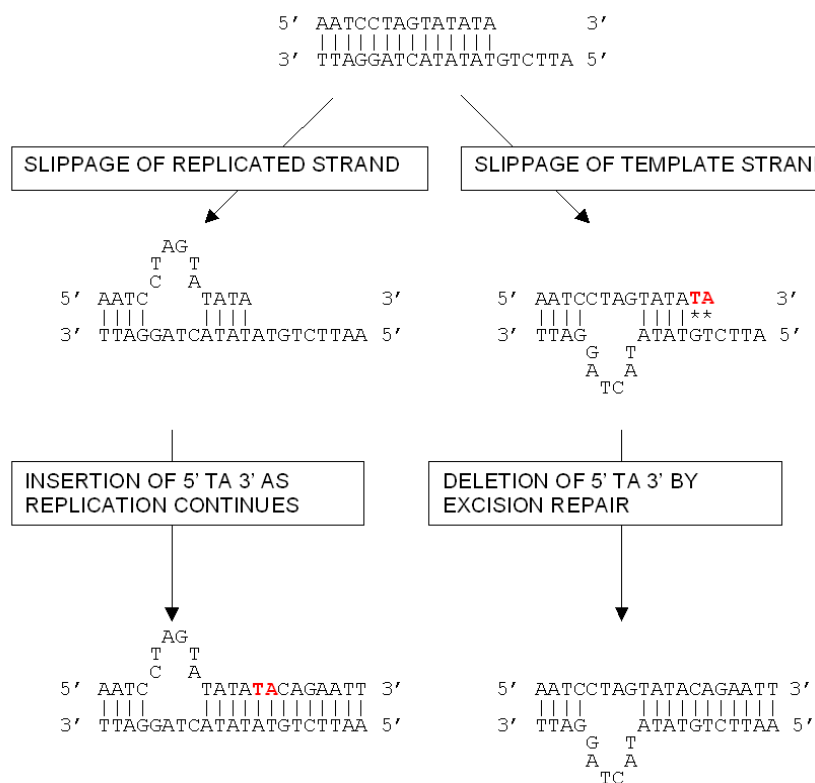


Figure 7.3.4 Genesis of Repeats in VNTRs. The peculiar tertiary structure of repetitive DNA allows mis-matching of neighbouring repeats, and, depending on the strand orientation, repeats can be inserted or deleted during replication.

7.3.1 Sizing of VNTR/MIRU loci

Overview

As mentioned in the previous section, all VNTR methods are based on PCR amplification of the repeats and the flanking DNA regions, followed by size separation of those products using various methods. Deducing the number of repeats in VNTR PCR (essentially by sizing the PCR products generated) can be executed in different ways. The method used in many laboratories that can be performed without specialised instrumentation is agarose gel electrophoresis (included in workshop practical). This method does not require PCR primers to be labelled and does not require specialised software/instrumentation expertise although it can be relatively time consuming. The second method is fragment sizing using capillary electrophoresis (automated sequencer) instrumentation (figure 7.3.5). In this case, several VNTR loci can be amplified at once in the same PCR, as the capillary electrophoresis instrument is capable of recognising different fluorophores labelling the different loci. These are incorporated into the product during PCR by the use of labelled specific primers. The labelled products are then separated by capillary electrophoresis and compared against a known size standard and product size automatically called following the generation of a size curve (see figure 7.3.5 below). Products are usually sized within 1-2 bp using this method although this level of precision is not always necessary.

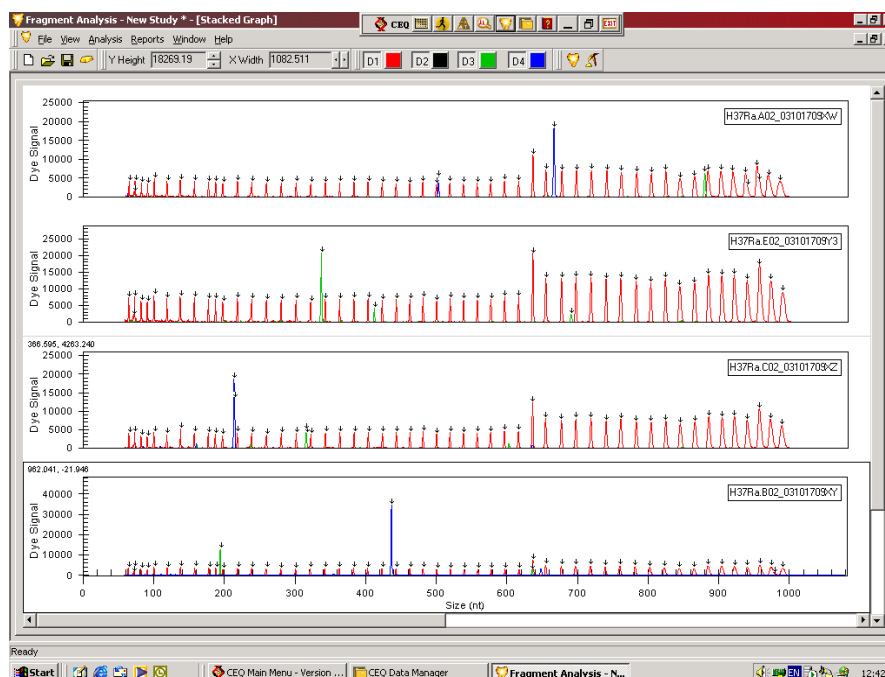


Figure 7.3.5 Multiplexed fragment sizing using products labelled with different fluorophores ('blue' and 'green') against a known size standard (in 'red').

7.3.2 VNTR/MIRU PCR Practical

7.3.2.1 DNA extraction

Principle

For the purposes of safety, we will extract DNA from *M. smegmatis*, a non-tuberculosis *Mycobacterium* species. We will use QIAamp, DNA Mini Kit, Qiagen and a modified extraction protocol. You will use *M. tuberculosis* DNA for the MIRU-VNTR.

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all times during the practical

Proteinase K – Irritant, skin and respiratory sensitizer.

Buffer AW1 – Harmful, Irritant, Respiratory sensitizer.

Buffer AL-Skin and Eye, Irritant, Skin sensitizer

Ethanol – Highly flammable/Irritant

Protocol:

- Pellet 1 ml of fresh exponential culture at 10000 *g* for 10 min. Discard the supernatant.
- Suspend bacterial pellet in 180 µl of 20 mg/ml lysozyme in 20 mM Tris·HCl, pH 8.0; 2 mM EDTA; 1.2% Triton. Incubate for a minimum of 30 min) at 37°C, 1000 rpm.
- Add 20 µl proteinase K and incubate at 56 °C, 1000 rpm for a minimum of 30 min.
- Add 4 µl RNase A and incubate at room temperature for 2 min.
- Add 200 µl buffer AL, mix by vortexing. Incubate at 70°C, 1000 rpm for 10 min and then for a further 15 min at 95°C.
- Briefly centrifuge the 1.5 ml microcentrifuge tube to remove drops from the inside of the lid.
- Add 200 µl ethanol (96–100%, ice-cold) to the sample, and mix by pulse-vortexing for 15 s. After mixing, briefly centrifuge the 1.5 ml microcentrifuge tube to remove drops from inside the lid. It is essential that the sample, Buffer AL, and ethanol are mixed thoroughly to yield a homogeneous solution. A white precipitate may form on addition of ethanol. It is essential to apply all of the precipitate to the QIAamp Mini spin column. This precipitate does not interfere with the QIAamp procedure or with any subsequent application. Do not use alcohols other than ethanol since this may result in reduced yields.
- Carefully apply the mixture (approximately 600 µl, including the precipitate) to the QIAamp Mini spin column (in a 2 ml collection tube) without wetting the rim. Close the cap, and centrifuge at 6000 × *g* (8000 rpm) for 1 min. Place the QIAamp Mini spin column in a clean 2 ml collection tube (provided), and discard the tube containing the filtrate. Close each spin column to avoid aerosol formation during centrifugation. It is essential to apply all of the precipitate to the QIAamp Mini spin column. Centrifugation is performed at 6000 × *g* (8000 rpm) in order to reduce noise. Centrifugation at full speed will not affect the yield or purity of the DNA. If the solution has not completely passed through the membrane, centrifuge again at a higher speed until all the solution has passed through.
- Carefully open the QIAamp Mini spin column and add 500 µl Buffer AW1 without wetting the rim. Close the cap, and centrifuge at 6000 × *g* (8000 rpm) for 1 min. Place the QIAamp Mini spin column in a clean 2 ml collection tube (provided), and discard the collection tube containing the filtrate.*

- Carefully open the QIAamp Mini spin column and add 500 µl Buffer AW2 without wetting the rim. Close the cap and centrifuge at full speed (20,000 x g; 14,000 rpm) for 3 min.
- Place the QIAamp Mini spin column in a new 2 ml collection tube (not provided) and discard the old collection tube with the filtrate. Centrifuge at full speed for 1 min. This step helps to eliminate the chance of possible Buffer AW2 carryover.
- Place the QIAamp Mini spin column in a clean 1.5 ml microcentrifuge tube (not provided), and discard the collection tube containing the filtrate. Carefully open the QIAamp Mini spin column and add 50 µl Buffer AE (or water). Incubate at room temperature for 1 min, and then centrifuge at 6000 x g (8000 rpm) for 1 min.
- Repeat previous step, i.e. elute the DNA once again with 50 µl Buffer AE (or water). Incubate at room temperature for 1 min, and then centrifuge at 6000 x g (8000 rpm) for 1 min.
- Store eluted DNA at -20°C.
- The quality and quantity of extracted DNA can be checked on NanoDrop and by electrophoresis in agarose gel.

7.3.2.2 MIRU-VNTR PCR

You have been provided with DNAs extracted from three *M. tuberculosis* isolates. You have also been provided with PCR reagents and primers for three of the current panel of 24 VNTR/MIRU loci (MIRU10, MIRU24 and MIRU26). Method and primers are according to Allix et al, **Proposal for standardization of optimised mycobacterial interspersed repetitive unit-variable number tandem repeat typing of *Mycobacterium tuberculosis***. 2006 *J Clin Microbiol.* 44:4498-4510.

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all times during the practical
2% E gels (contain ethidium bromide) – Toxic/Mutagen

Protocol

1. Label three sterile 1.5 ml microtubes, one for each MIRU primer pair (one for MIRU10, one for MIRU24 and one for MIRU26), Label 0.2 ml PCR tubes (running each sample in duplicate) for all samples to be analysed three VNTR/MIRU PCR mixtures.
2. Prepare PCR mixtures for each primer pair as in the table below.

Table 7.3.2.2.1 MIRU-VNTR PCR reaction composition.

Reagent	Stock concentration	Reaction concentration	Volume per reaction (µl)	Volume per n reactions (µl)
DNA template			2	
Forward primer	10 µM	200 nM	0.5	
Reverse primer	10 µM	200 nM	0.5	
Sigma PCR mix	2×	1×	12.5	
Sterile distilled water			9.5	
Total			25	

3. Dispense each VNTR/MIRU mastermix by 23 µl into each 0.2 ml PCR tube and add 2 µl of DNA sample into each tube.
4. Carry out the PCR at following cycling conditions:

95°C, 12 min (important for a GC-rich template)

94°C, 30 s]	40 cycles
54°C, 1 min]	
72°C, 2 min]	

72°C, 7 min

7.3.2.3 Agarose gel electrophoresis of MIRU/VNTR PCR products

Principle

PCR products will be separated according to their size and DNA negative charge in agarose gel electrophoresis. For ease of use we are going to size the products generated in the PCR by running them out on a ready-made E-gel (for product information see Section 4.0 Appendix 1) with a 1kb ladder size marker to compare the sizes of the products generated by our test DNAs D1-D3 for the three VNTR loci.

Protocol

1. Plug the PowerBase™ into an electrical outlet.
2. Open the package containing the gel (**with the comb in place**) into the apparatus right edge first. Press firmly at the top and bottom to seat the gel in the base. You should hear a snap when it is in place. The Invitrogen logo should be located at the bottom of the base, close to the positive pole. A steady, **red** light will illuminate when the E-Gel is correctly inserted (Ready Mode).
3. Press and hold either button until the **red** light turns to a **flashing green** light. This indicates that the 2 minute pre-run has started.
4. At the end of the pre-run the current will automatically shut off. The **flashing green** light will change to a **flashing red** light and the PowerBase™ will beep rapidly.
5. Press and release either button to stop the beeping. The light will change from a **flashing red** light to a **steady red** light.
6. Remove the comb from the gel using both hands to lift the comb gently by rolling the comb slowly towards you. Be careful to pull the comb up straight from both sides and do not bend the comb. Remove any excess fluid with a pipette.
7. *All wells in the gel must contain 20ul of sample or water.* Load 5ul of 1kb ladder in the first well. Top up this well with 15ul of water. Load 5ul of each VNTR PCR product and the control PCR products except for one locus control as there are only 12 wells in the gel. Top up these wells with 15ul water.
8. On the PowerBase™ choose a 30 minute run by pressing and releasing the 30 min button. The light will change to a **steady green** light. The end of the run is signalled by a **flashing red** light and a rapid being.
9. Press and release either button to stop the beeping. The light will turn to a **steady red** light.
10. Remove the gel from the power unit and analyze your results on a UV transilluminator.

Data interpretation

The 1kb ladder sizes are as follows:

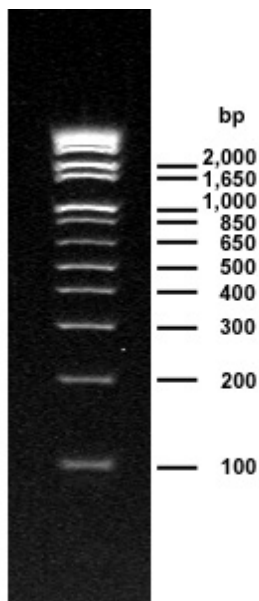


Figure 7.3.2.3.1 VNTR calculation using agarose gel electrophoresis

A simple calculation of fragment size minus the lengths of the forward and reverse primers and regions flanking the VNTR, divided by the length of one individual VNTR unit will give you the number of VNTR units for the isolate. A VNTR/MIRU sizing table is also shown for convenience (Table 7.3.2.3.1). Determine the sizes of the products and then calculate the number of repeats per product. An example is shown below

Is there a variation in copy number at each of the three loci for each DNA?

Example using MIRU4 (aka ETRD):

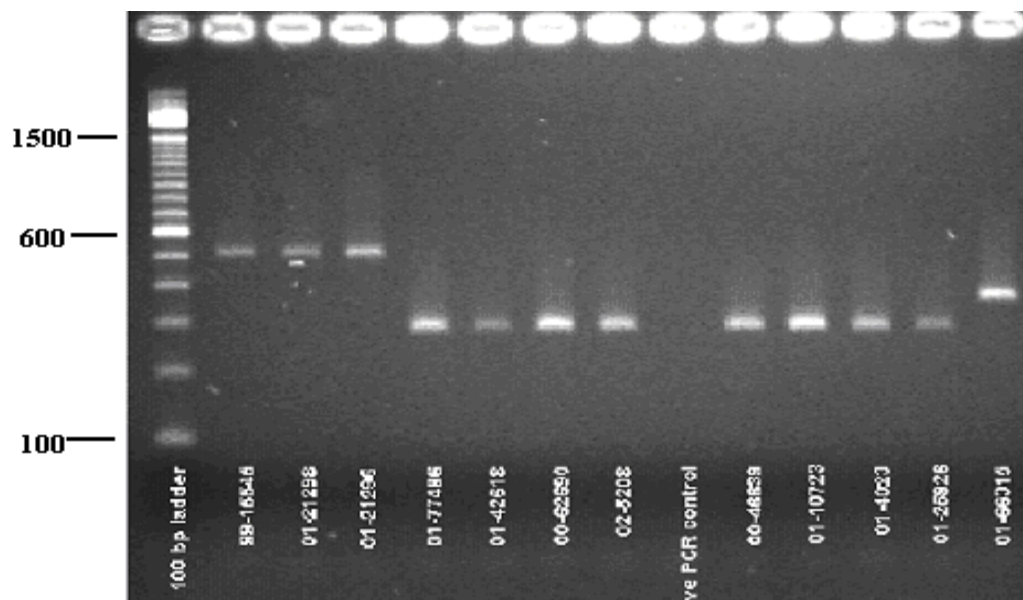


Figure 7.3.2.3.2 Example of MIRU/VNTR sizing. Lane 1 = 100 base pair (bp) marker, increasing in 100 bp increments (NB your marker is different to this picture).

Flanking region of PCR product = 141

Repeat = 77 bp

Table 7.3.2.3.1 VNTR/MIRU sizing table in base pairs

Allele	MIRU 02	MIRU 04	MIRU 10	MIRU 16	MIRU 20	MIRU 23 ¹	MIRU 23 ²	MIRU 24	MIRU 26 ¹	MIRU 26 ²	MIRU 27	MIRU 31	MIRU 39	MIRU 40
0	402	175	482	565	437	558	150	395	461	285	498	492	540	354
1	455	252	537	618	514	608	200	447	512	336	551	545	593	408
2	508	329	590	671	591	661	253	501	563	387	604	598	646	462
3	561	406	643	724	668	714	306	555	614	438	657	651	699	516
4	614	483	696	777	745	767	359	609	665	489	710	704	752	570
5	667	560	749	830	822	820	412	663	716	540	763	757	805	624
6	720	637	802	883	899	873	465	717	767	591	816	810	858	678
7	773	714	855	936	976	926	518	771	818	642	869	863	911	732
8	826	791	908	989	1053	979	571	825	869	693	922	916	964	786
9	879	868	961	1042	1130	1032	624	879	920	744	975	969	1017	840
10	932	945	1014	1095	1207	1085	677	933	971	795	1028	1022	1070	894
11	985	1022	1067	1148	1284	1138	730	987	1022	846	1081	1075	1123	948
12	1038	1099	1120	1201	1361	1191	783	1041	1073	897	1134	1128	1176	1002
13	1091	1176	1173	1254	1438	1244	836	1095	1124	948	1187	1181	1229	1056
14	1144	1253	1226	1307	1515	1297	889	1149	1175	999	1240	1234	1282	1110
15	1197	1330	1279	1360	1592	1350	942	1203	1226	1050	1293	1287	1335	1164

¹with oligonucleotides used in Supply et al., 2000, *Mol. Microbiol.*, 36, 762-771 and described in Mazars et al 2001, *PNAS*, 98, 1901-1906 (see supplemental data on the PNAS Web site <http://www.pnas.org>)

²with oligonucleotides used in Supply et al. 2001, *J. Clin. Microbiol.*, 39, 3563-3571

7.3.3 Background reading and references

7.3.3.1 Genotyping *M. tuberculosis*

The insertion sequence IS6110 has been shown to be an important contributor to the generation of genetic diversity in *Mtb*. IS6110 may insert into genes, disrupting the coding sequence, generate genomic deletions through recombination events and affect gene expression through its intrinsic promoter activity.

For many years, our appreciation of the phylogenetic relationships between strains of *Mtb* has been restricted by the use of limited measures of genetic variation. The standard genotyping instrument for *Mtb* has been IS6110 restriction fragment length polymorphism (RFLP) analysis. This is an excellent tool for epidemiological investigations since it is highly discriminatory, but is not necessarily best suited for phylogenetic analysis. The insertion sequence has irregular rates of transposition and the difficulty in standardizing methodology and interpretation of the fingerprints obtained has lead to a confusing proliferation of nomenclature compounded by a lack of communication between centres.

More recently, typing methods such as spoligotyping and multiple interspersed repetitive unit (MIRU) analysis have offered a more easily standardized approach, albeit not necessarily more suited to phylogenetic analysis.

Spoligotyping is based on polymorphism in the direct repeat region of the *Mtb* chromosome. At this locus, there is a series of repeat 36 base pair sequences which are interspersed by non-repetitive spacer elements each 35 to 41 base pairs in length. The presence or absence of each of 43 of these spacer regions is detected using a reverse dot blot technique. Briefly, the spacer regions are amplified using primers directed at the common flanking direct repeat elements. The amplified fragments are then hybridized to a membrane that has probes for each of the spacer elements. The membrane is washed and the presence of bound fragments detected by chemiluminescence. This technique is technically simple, reproducible and easily coded and has, as a result, allowed the development of an international database and introduced a degree of consistency to the naming of the major lineages of *Mtb*. Moreover, there is a high degree of consistency in the identification of the major lineages of *Mtb* between spoligotyping, large sequence polymorphism and single nucleotide polymorphism analysis. However, since major changes in spoligotype may occur due to single deletion events, it is not always useful for defining the relationship between strains, nor is it sufficiently discriminatory to differentiate closely related strains. It is therefore less useful for epidemiological investigations where it is important to differentiate between closely related but different strains.

The genome of *Mtb* contains repetitive 40-100 base pair sequence elements known as 'mycobacterial interspersed repetitive units' (MIRU) which are found as tandem repeats at multiple intergenic loci throughout the chromosome. These loci demonstrate hypervariability in the number of repeats, leading to the use of the term "variable number of tandem repeats" (VNTR).

The typing method MIRU-VNTR is based on the detection of the number of tandem repeats present at several different loci. The most commonly used methods examine 12-20 of these loci. MIRU-VNTR is performed by amplifying each locus and determining the size of the amplicon and hence the number of repeats at each locus. These are then combined into a multi-digit code. This

technique is more discriminatory than spoligotyping, and, since many independent loci are assessed, may be more appropriate for phylogenetic analysis. However it is less visually intuitive than spoligotyping, more labour intensive and does not always concord well with SNP-based phylogenetic analysis. This may relate to the potential for gain or loss of repeats at each locus, leading to the possibility of convergent evolution (broadly different strains acquiring similar numbers of repeats at a particular locus).

One of the important mechanisms by which genetic diversity arises in *Mtb* is through large chromosomal deletions. Such deletions, termed large sequence polymorphisms (LSP), have been used to classify strains of *Mtb* into distinct lineages. Since recombination is rare, all progeny of a parent strain will carry the same deletion. The presence or absence of these deletions is used to define major lineages within *Mtb*. This approach infers a basic structure to the evolutionary tree by assuming the sequential acquisition of deletions. It is also useful for establishing clonal relationships amongst strains with different spoligotype or RFLP patterns. This technique has been used to show that those strains of *Mtb* which have been sequenced cluster in two related 'modern' lineages and do not represent the global diversity of strains.

The most robust method, however, for defining the global phylogeny of *Mtb*, is the use of SNP analysis and whole genome sequencing. SNP analysis, in a restricted form, was used to classify *Mtb* strains into three principal genetic groups, based on SNPs in codon 463 of the *katG* gene and codon 95 of the *gyrA* gene (Sreevatsan *et al* 1997). This three-branched tree defined the broad phylogeny of *Mtb* for many years. The results of recent more extensive SNP analysis will advance this understanding. Gagneux and colleagues have sequenced 90 genes in 104 globally representative strains to define the relationships between the major lineages of *Mtb*. This is an important step forward in defining the nature of diversity in *Mtb* and offers an ideal starting point for evaluating the clinical implications of such diversity. The application of relatively low cost, high throughput sequencing technology to multiple representative strains of *Mtb* may significantly advance our understanding of the extent of genetic diversity in tuberculosis as well as the relationships between strains. Figure 7 and Table 2 summarize the major methods used for genotyping *Mtb* strains.

Finally, the region of the *Mtb* chromosome containing the so-called PPE/PE gene families has been shown to be an important source of genetic variation. In particular, it has been suggested that this region may be partly responsible for antigenic variation in *Mtb*, which could feasibly be a mechanism for immune evasion.

7.4 Whole genome sequencing. Analysis of antibiotic resistance from whole genome sequence data of *Mycobacterium tuberculosis* clinical isolates

Introduction

Although molecular genotyping techniques have proved useful in confirming the occurrence of re-infection, they provide a surrogate measure of inter strain diversity. Recent applications of whole genome sequencing to *M. tuberculosis* have demonstrated that this technique can provide greater insight into the evolution, epidemiology and mutation rate of this clonal organism. Tuberculosis patients undertake treatment consisting of combination of four antimicrobial drugs (isoniazid, rifampicin, pyrazinamide and ethambutol) and the duration of the treatment is 6 months (2HRZE/4HR). Antimicrobial resistance in *M. tuberculosis* usually arises as a result of point mutation in the coding sequence leading to structural changes in the target proteins.

Principle

This guide will help you identify mutations conferring antibiotic resistance in whole genome sequences (WGS) of *M. tuberculosis* clinical isolates. You will be provided WGS data of TB isolates from patients treated with multiple antibiotics. We will use Artemis, a genome browser tool, to analyze the sequences of four genes *rpoB*, *katG*, *inhA* and *gyrA*, which encode for ribosomal polymerase β subunit, catalase/oxidase, enoyl reductase and DNA gyrase, respectively. The mutations in these genes are responsible for antibiotic resistance to rifampicin, isoniazid and fluoroquinolones.

The total genomic DNA of *M. tuberculosis* isolates was sequenced by next generation sequencing on MiSeq (Illumina) using paired-end reads sequencing approach. Thousands of short reads were acquired for each sample. You can find these sequences in each patient's file under the names SRR974841_1.fastq.gz and SRR974841_2.fastq.gz. There are two files for each sample as the library was constructed by paired-end sequencing. You can unzip these files and explore the contents.

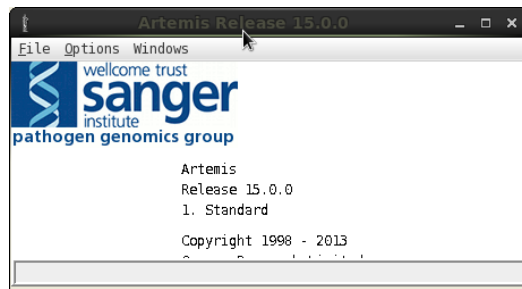
The sequences of clinical isolates are mapped against a reference sequence - *Mycobacterium tuberculosis* H37Rv with Bowtie2 software creating BAM files. The BAM files are located within the folders for each patient. SNP calling was performed with Samtools mpileup with a threshold of 13 for base quality. The result is VCF format files containing all variations that are detected in each patient's isolate. The BAM and VCF are large data files. We will use Artemis software to analyze the resulting sequences.

Artemis is a free genome browser and annotation tool and can be downloaded from Sanger Institute website (<http://www.sanger.ac.uk/resources/software/artemis/>).

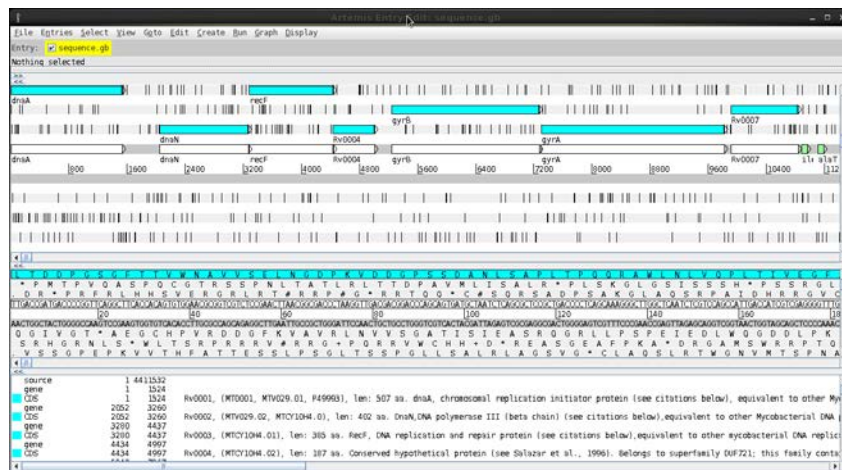
The aim of this practical is to identify any mutations present in any of the four genes *rpoB*, *katG*, *inhA* and *gyrA*.

Procedure

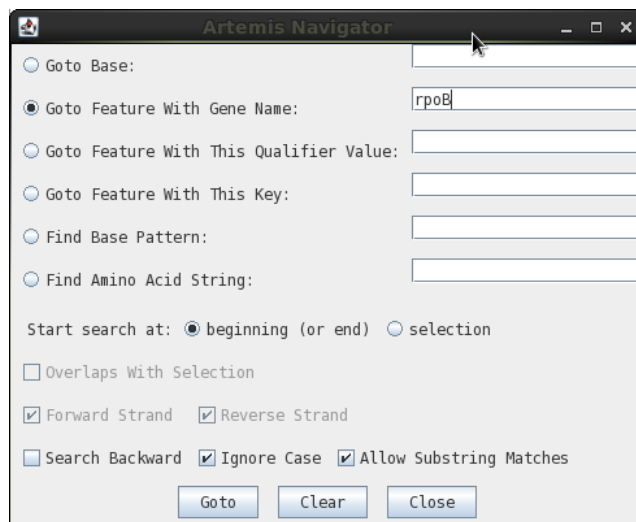
- 1) Open Artemis software.



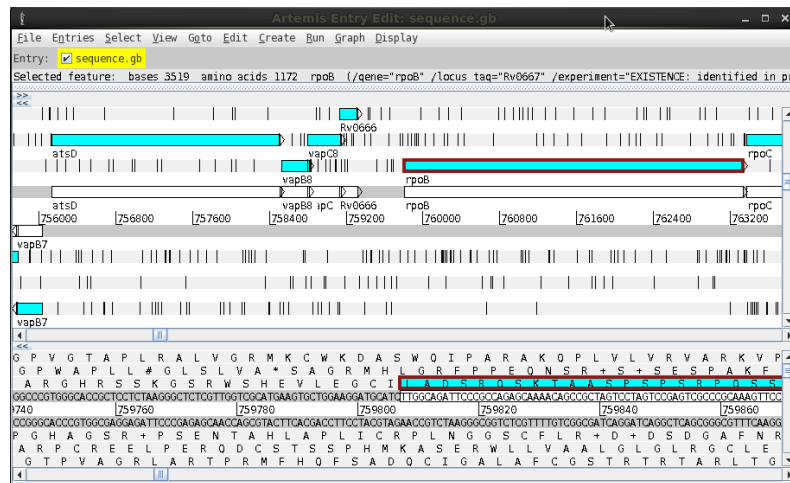
- 2) Open *Mycobacterium tuberculosis* H37Rv genome.
 - a. In the Artemis window "File->Open-> path(../H37RvReference /sequence.gb)"



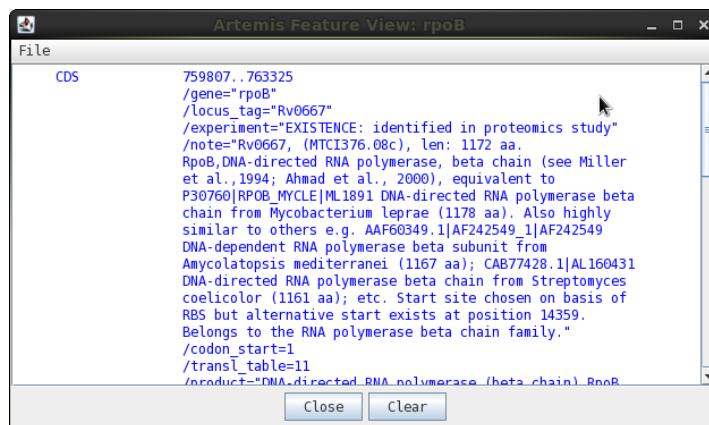
- b. Locate *rpoB* gene “Goto->Navigator” and type “rpoB” in the “Goto Feature with Gene Name”.



- c. You will be directed to the *rpoB* sequence, you can explore the basis and the amino acid sequences in the bottom window.



- d. Select the *rpoB* gene and press the “V” key to get more information about this gene.



- 3) Open the data from one patient, for example patient1.
 - a. In the genome window “File->Read BAM/VCF..” and select the “mapping/P1/SRR974841_P1.bam”.
 - b. Add more to the list: open “mapping/P1/SRR974841_P1.vcf.gz”. Press OK.
 - c. VERY IMPORTANT: open the file with “vcf.gz” extension, not just the “.vcf”.



- d. Go to the *rpoB* gene and identify if patient 1 has any mutations in this gene.

The screenshot shows the Artemis genome browser interface. At the top, the 'Entry' is 'sequence.gb'. Below the header, the DNA sequence is displayed with coordinates 761051 to 761171. A red vertical line indicates a specific position. Annotations include: 'Mutation TCG->TTG' pointing to a change in the DNA sequence; 'SNPs' pointing to a red vertical line; 'Reference Amino' pointing to the amino acid 'S' in the protein sequence; 'Reference codon' pointing to the codon 'TCA' in the DNA sequence; and 'Selected base' pointing to the base 'T' in the DNA sequence.

- 4) Open the genome sequences of all four patients together and analyze the mutations in the genes *rpoB*, *katG*, *inhA* and *gyrA*. Create a results table with all mutations that you can find in all four genes for every patient.
- Note: You can select one BAM or VCF file when you have multiple files loaded in Artemis. Right click in the reads or SNP area and select "BAM Files" or "VCF Files" and check/uncheck what file you want or do not want to view. This is important because if you analyze more than one uploaded sequence in the Read Map Area you will not be able to identify a particular mutation. All reads are mixed from all uploaded sequences.

Overview

In a long-term outbreak (1997 to 2010) WGS analyses of the 86 isolates revealed 85 single nucleotide polymorphisms (SNPs), subdividing the outbreak into seven genome clusters (two to 24 isolates each), plus 36 unique SNP profiles. Genome-based clustering patterns mapped contact tracing data and the geographical distribution of the cases than clustering patterns based on classical genotyping. A maximum of three SNPs were identified in eight confirmed human-to-human transmission chains, involving 31 patients. Multiple analyses support an average mutation rate of ~0.3 SNPs per genome per year. Another study sequenced 390 separate isolates from 254 patients, including representatives from all five major lineages of *M. tuberculosis*. The estimated rate of change in DNA sequences was 0.5 single SNPs per genome per year (95% CI 0.3–0.7) in longitudinal isolates from 30 individuals and 25 families. Divergence is rarely higher than five SNPs in three years. 109 (96%) of 114 paired isolates from individuals and households differed by five or fewer SNPs. More than five SNPs separated isolates from none of 69 epidemiologically linked patients, two (15%) of 13 possibly linked patients, and 13 (17%) of 75 epidemiologically unlinked patients (three-way comparison exact $p < 0.0001$). Thus, it is usually assumed that SNP differences of less than 6 are associated with recent transmission.

Whole genome sequencing is of value in investigating the relationship of a relapse strain with the initial culture for a patient failing on or after treatment. A recent study demonstrated that relapse strains had less than 6 SNPs whereas strains that were designated as reinfections had >1300 SNPs and belonged to different lineages making it possible to distinguish these outcomes unequivocally². An important finding of this paper was that mixed infection was found to be common in the patients in South Africa. This may have important implications for future investigations in the evolution of antibiotic resistance and strain transmission.

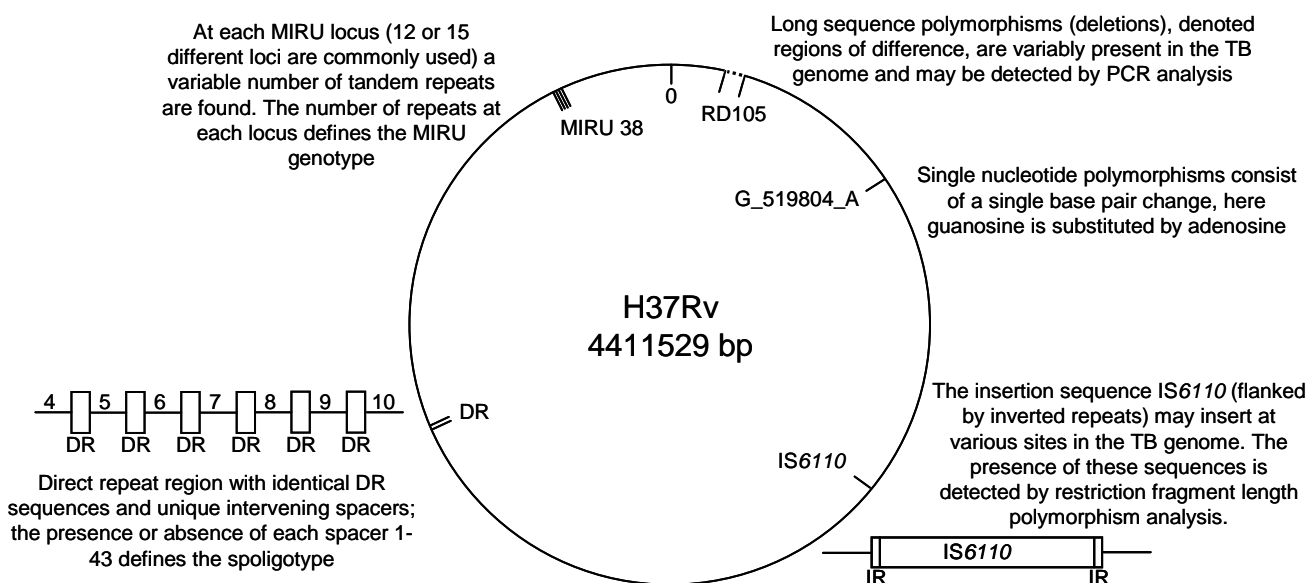


Figure 7.4.1 Schematic representation of the *M. tuberculosis* genome, indicating the genetic basis of genotyping techniques. The circular chromosome of the reference strain H37Rv is shown together with examples of the major genetic elements used for strain genotyping. For clarity only one MIRU (mycobacterial interspersed repetitive unit) locus, one IS6110 (insertion sequence), one region of difference (RD) and one single nucleotide polymorphism are shown.

Table 7.4.1 Comparison of techniques for molecular typing or evaluating genetic diversity within *M. tuberculosis*

	IS6110 RFLP*	Spoligotyping	MIRU-VNTR†	LSP analysis‡	SNP§ analysis	Whole Genome sequencing
Discriminatory power	Excellent	Fair to poor	Good to excellent (newer protocols)	Poor	Poor, likely to improve with increased SNP identification	Very high
Ease of use	Time consuming and technically demanding; requires extracted chromosomal DNA	Rapid and simple; can be performed directly on specimens or heat-killed cultures.	Rapid and fairly simple; automation requires access to sophisticated equipment; can be performed on heat-killed cultures.	Simple and robust	Rapid and fairly simple; high throughput analysis requires access to sophisticated equipment	This method has potential for rapid use but is still too cumbersome for routine clinical use.
Interpretation	Simple, but not easily standardized	Simple visual interpretation	Simple visual interpretation	Straightforward	Straightforward	Requires extensive bio-informatics support
Data sharing	Complex, lack of standardized nomenclature	Straightforward, standardized binary or octal coding	Straightforward, standardized numerical coding	Straightforward, standardized nomenclature	Straightforward, standardized nomenclature	Simple
Utility for epidemiological investigations	Excellent due to high discriminatory power	Useful for rapid cluster identification but requires secondary confirmation	Excellent and rapid with newer protocols	Poor due to low discriminatory power	Poor at present due to low discriminatory power	Very high. Able to dissect transmission series and emergence of resistance simultaneously
Utility for phylogenetic analysis	May be limited by irregular rates of transposition and favoured sites.	Fairly good correlation with SNP-based phylogeny. Large deletions may bias analysis	Relatively poor correlation with SNP-based phylogeny	Useful for evolutionary history (sequential deletions) and for identifying major lineages	Gold-standard; low rate of SNP in <i>Mtb</i> necessitates large-scale sequencing to identify informative SNPs	Gold standard once the practical problems of bioinformatics is addressed

Table footnote:

* restriction fragment length polymorphism

† mycobacterial interspersed repetitive unit-variable number of tandem repeat

‡ large sequence polymorphism

§ single nucleotide polymorphism

Additional reading for IS6110/MIRU /Sequencing

Almeida Da Silva PE, Palomino JC. Molecular basis and mechanisms of drug resistance in *Mycobacterium tuberculosis*: classical and new drugs. *J Antimicrob Chemother.* 2011 Jul;66(7):1417-30.

Arnold C*, Westland L, Mowat G, Underwood A, Magee J, Gharbia S (2005) Single nucleotide polymorphism-based differentiation and drug resistance detection in *Mycobacterium tuberculosis* from isolates or directly from sputum. *Clin Microbiol Infect.*

Borrell S, Thorne N, Español M, Mortimer C, Orcau A, Coll P, Gharbia S, González-Martín J, Arnold C. *Tuberculosis* (Edinb). (2009) Comparison of four-colour IS6110-fAFLP with the classic

IS6110-RFLP on the ability to detect recent transmission in the city of Barcelona, Spain. *May*;89(3):233-7.

Bryant JM, Harris SR, Parkhill J, *et al.* Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *Lancet Respiratory Medicine* doi:10.1016/S2213-2600(13)70231-5.

Comas I, Homolka S, Niemann S and Gagneux S (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PloS ONE* 2009;4:e7815

Frothingham, R. and W. Meeker-O'Connell (1998). "Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats." *Microbiology* 144: 1189-1196.

Goyal, M., N. Saunders, et al. (1997). "Differentiation of *Mycobacterium tuberculosis* isolates by spoligotyping and IS6110 restriction fragment length polymorphism." *J Clin Microbiol* 34: 647-651.

Kanduma E, McHugh TD, Gillespie SH. Molecular methods for *Mycobacterium tuberculosis* strain typing: a users guide. *J Appl Microbiol.* 2003;94(5):781-91

Maguire H, Dale JW, McHugh TD, Butcher PD, Gillespie SH, Costetsos A, Al-Ghusein H, Holland R, Dickens A, Marston L, Wilson P, Pitman R, Strachan D, Drobniewski FA, Banerjee DK. Molecular epidemiology of tuberculosis in London 1995-7 showing low rate of active transmission. *Thorax.* 2002 Jul;57(7):617-22.

McHugh TD, Batt SL, Shorten RJ, Gosling RD, Uiso L, Gillespie SH. *Mycobacterium tuberculosis* lineage: a naming of the parts. *Tuberculosis (Edinb).* 2005 May;85(3):127-36.

Maguire H, Dale JW, McHugh TD, Butcher PD, Gillespie SH, Costetsos A, Al-Ghusein H, Holland R, Dickens A, Marston L, Wilson P, Pitman R, Strachan D, Drobniewski FA, Banerjee DK. Molecular epidemiology of tuberculosis in London 1995-7 showing low rate of active transmission. *Thorax.* 2002 Jul;57(7):617-22.

Otal I, Samper S, Asensio MP, Vitoria MA, Rubio MC, Gomez-Lus R, Martin C (1997). Use of a PCR method based on IS6110 polymorphism for typing *Mycobacterium tuberculosis* strains from BACTEC cultures. *J Clin Microbiol.* Jan;35(1):273-7.

Ramaswamy S, Musser JM (1998) Molecular genetic basis of antimicrobial agent resistance in *Mycobacterium tuberculosis*: 1998 update. *Tuber Lung Dis:* 79(1):3-29.

Reisig F, Kremer K, Amthor B, van Soolingen D, Haas WH (2005). Fast ligation-mediated PCR, a fast and reliable method for IS6110-based typing of *Mycobacterium tuberculosis* complex. *J Clin Microbiol.* Nov;43(11):5622-7.

Roetzer A, Diel R, Kohl TA, *et al.* Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular

epidemiological study. *PLoS Med* 2013; **10**: e1001387.

Ruddy M, McHugh TD, Dale JW, Banerjee D, Maguire H, Wilson P, Drobniewski F, Butcher P, Gillespie SH. Estimation of the rate of unrecognized cross-contamination with mycobacterium tuberculosis in London microbiology laboratories. *J Clin Microbiol*. 2002 Nov;40(11):4100-4.

Shorten RJ, McGregor AC, Platt S, Jenkins C, Lipman MC, Gillespie SH, Charalambous BM, McHugh TD. When is an outbreak not an outbreak? Fit, divergent strains of *Mycobacterium tuberculosis* display independent evolution of drug resistance in a large London outbreak. *J Antimicrob Chemother*. 2013 Mar;68(3):543-9.

Sreevatsan S, Pan X, Stockbauer K, et al. (1997) Restricted gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA*;94:9869-9874

Supply, P., E. Mazars, et al. (2000). Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol Microbiol* 36: 762-771.

Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Loch C, van Soolingen D (2006). Proposal for standardization of optimised mycobacterial interspersed repetitive unit-variable number tandem repeat typing of *Mycobacterium tuberculosis*. 2006 *J Clin Microbiol*. 44:4498-4510.

Thorne N, Evans JT, Smith EG, Hawkey PM, Gharbia S, Arnold C (2007). An IS6110-targeting fluorescent amplified fragment length polymorphism alternative to IS6110 restriction fragment length polymorphism analysis for *Mycobacterium tuberculosis* DNA fingerprinting. *Clin Microbiol Infect*. Oct;13(10):964-70.

Thorne N, Borrell S, Evans J, Magee J, García de Viedma D, Bishop C, Gonzalez-Martin J, Gharbia S, Arnold C (2011) IS6110-based global phylogeny of *Mycobacterium tuberculosis*. *Infect Genet Evol*. Jan;11(1):132-8.

van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R and Shinnick TM (1993). Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol*, , 406-409, Vol 31, No. 2.

Walker TM, Ip CL, Harrell RH, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *The Lancet Infectious Diseases* 2012. doi:10.1016/S1473-3099(12)70277-3.

Wilson SM, Goss S, Drobniewski F (1998). Evaluation of strategies for molecular fingerprinting for use in the routine work of a *Mycobacterium* reference unit. *J Clin Microbiol*. Nov;36(11):3385-8.

Appendix 1: E-Gel product information

E-Gel® Electrophoresis System – taken from Invitrogen E-Gel® Technical Guide

Introduction

The E-Gel® agarose gel electrophoresis system is a complete bufferless system for agarose gel electrophoresis of DNA samples. The major components of the system are:

- E-Gel® pre-cast agarose gels
- Electrophoresis bases

E-Gel® pre-cast agarose gels are self-contained gels that include electrodes packaged inside a dry, disposable, UV-transparent cassette. The E-Gel® agarose gels run in a specially designed device that is a base and power supply combined into one device (two bases are available for running E-Gels, the new iBase™ system and the original, economical E-Gel® Powerbase™).

Advantages of E-Gel® Agarose Gels

Using E-Gel® agarose gels for electrophoresis of DNA samples offer the following advantages:

- Provides fast, safe, consistent, high-resolution electrophoresis
- Eliminates the need to prepare agarose gels and buffers, and to stain gels
- Compatible with most commercially available robotic systems for high-throughput agarose gel electrophoresis
- Available in a variety of agarose percentages, well formats, and throughput capacities to suit your applications
- Offered with a number of different DNA gel stains to accommodate your application
- Includes E-Gel® CloneWell™ and E-Gel® SizeSelect™ gels, to accelerate and simplify DNA gel purification and improve cloning results

Throughput Capacity

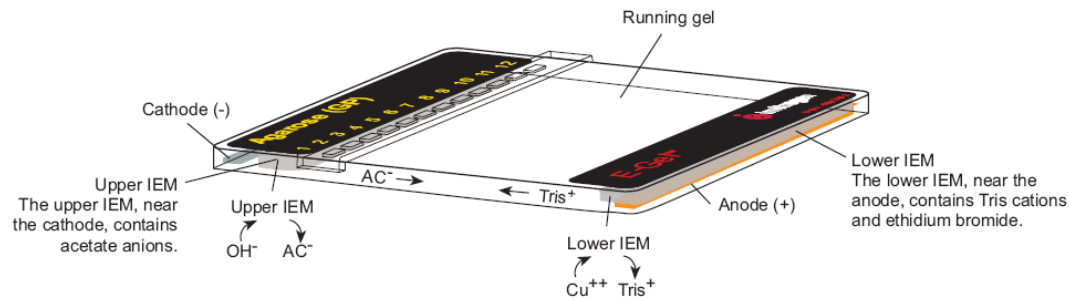
Three categories of E-Gel® agarose electrophoresis systems are available from Invitrogen based on your throughput requirements.

- Low-Throughput E-Gel® Electrophoresis System designed for electrophoresis of 8–16 DNA samples per gel.
- Medium-Throughput E-Gel® Electrophoresis System designed for electrophoresis of 48 DNA samples per gel. This system is compatible for use with multichannel pipettors or automated liquid handling systems.
- High-Throughput E-Gel® Electrophoresis System is designed for electrophoresis of 96 DNA samples per gel. This system is compatible for use with multichannel pipettors or automated liquid handling systems.

E-Gel® Single Comb and DoubleComb Gels

The E-Gel® single comb and double comb gels are bufferless gels containing electrodes embedded in the agarose matrix. Each gel contains an ion generating system (TAE buffer system), a pH balancing system, and ethidium bromide for DNA staining and is packaged inside

an UV-transparent cassette. To create a patented bufferless system, each E-Gel® single comb and double comb cassette contains two ion exchange matrices (IEMs) that are in contact with the gel and electrodes. The IEMs supply a continuous flow of ions through out the gel resulting in a sustained electric field required for running the gel (see figure below).

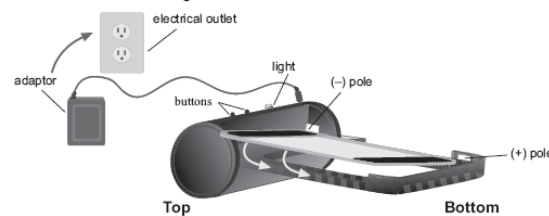


E-Gel® PowerBase™

The E-Gel® PowerBase™ Version 4 (figure below) is an easy-to-use, automated device specifically designed to simplify electrophoresis of single comb or double comb E-Gel® agarose gels from Invitrogen. The E-Gel® PowerBase™ is a base and a power supply all in one device.

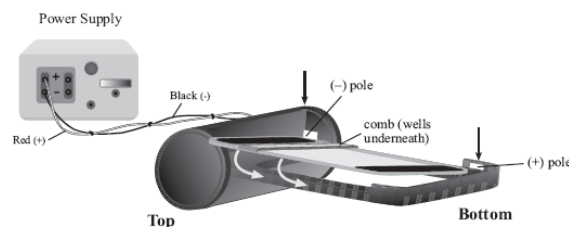
The operation of the E-Gel® PowerBase™ v. 4 is controlled by two buttons on top of the base. The left button is for a double comb run and right button is for a single comb run (see the label on the unit). To select different electrophoresis runs for the PowerBase™, do one of the following (page 30 for details)

- Press and release the button (run) **OR**
- Press and hold the button (pre-run)



E-Gel® Base

The E-Gel® Base (see figure below) previously available from Invitrogen connects to a power supply and is used for electrophoresis of E-Gel® single comb, and double comb agarose gels (page 122 for details).



Appendix 2: Agarose Gel Electrophoresis Sizing of PCR products

Overview of A.G.E.

Electrophoresis is a technique that results in the separation of charged molecules based on their size. DNA is a negatively charged molecule in solution so it will move to the positive pole in an electric field, in this case through a matrix of polymerised agarose molecules. All linear DNA has the same charge per unit length and linear pieces migrate according to size.

The end result is that large pieces of DNA move slower than small pieces of DNA. The position that the DNA migrated to in the gel is observable under ultraviolet light when the gel is stained with ethidium bromide. Sizes of your DNA samples can be determined by running a size standard (or a DNA molecular weight ladder) alongside your samples. A variation of agarose gel electrophoresis, called *pulsed-field gel electrophoresis*, makes it possible to separate even extremely long DNA molecules. Ordinary gel electrophoresis fails to separate such molecules because the steady electric field stretches them out so that they travel end-first through the gel in snakelike configurations at a rate that is independent of their length. In pulsed-field gel electrophoresis, by contrast, the direction of the electric field is changed periodically, which forces the molecules to reorient before continuing to move through the gel. This reorientation takes much more time for larger molecules, so that progressively longer molecules move more and more slowly. As a consequence, even entire bacterial or yeast chromosomes separate into discrete bands in pulsed-field gels and so can be sorted and identified on the basis of their size.

Protocol

Components of AGE:

- Agarose
- Casting tray and comb
- Electrophoresis buffer (usually 1xTBE)
- Power pack
- Electrophoresis tank
- Stain (0.5mg/ml Ethidium bromide)
- Transilluminator

Several different buffers have been recommended for electrophoresis of DNA. The most commonly used for duplex DNA are TAE (Tris-acetate-EDTA) and TBE (Tris-borate-EDTA). DNA fragments will migrate at somewhat different rates in these two buffers due to differences in ionic strength. Buffers not only establish a pH, but provide ions to support conductivity.

DNA Loading Buffer (6x)

- 6mM EDTA
- 300mM NaOH
- 18% Ficoll (in water)
- 0.15% Bromophenol blue
- 0.25% Xylene Cyanol

Making a 2% agarose gel:

- Weigh out 2 g of agarose in a large conical flask and add 100 ml 1xTBE. Stopper the flask with a foam bung.
- Microwave for 2 or 3 minutes until completely dissolved.

- Cool under a cold tap for a short time to 60°C (70°C for concentrations 2% or above).
- Meanwhile, prepare a casting tray by sealing the ends, if necessary, and finding an appropriate comb for the tray you are using and the number of samples you are loading.
- Pour molten agarose into the casting tray, careful not to make any bubbles in the gel.
- Leave gel to set for approx. 30 mins.

Table to show DNA Size Separation by Agarose Concentration

Effective Range of Separation for Linear Fragments (kb)	Agarose (%)
30 to 1	0.5
12 to 0.8	0.7
10 to 0.5	1.0
7 to 0.4	1.2
3 to 0.2	1.5

N.B. Bromophenol blue (BP) generally co-migrates with 0.5 kb fragments of DNA.

Loading samples:

- Once set, remove comb carefully and remove tape from the ends if used.
- Place the casting tray with gel into the electrophoresis tank and make sure there is sufficient TBE buffer in the tank to completely submerge the gel. The orientation of the gel should be so the wells are closest to the negative electrode. Load your PCR products with loading buffer and include a DNA ladder in one of the wells. Bromophenol blue (BP) generally co-migrates with 0.5 kb fragments of DNA. Stain with 0.5 mg/ml ethidium bromide for 10 minutes. Place lid on electrophoresis tank.
- Switch on the power pack. Voltage: Agarose gels can be run at various voltages, depending on the separation desired and the available time. As the voltage applied to a gel is increased, larger fragments migrate proportionally faster than small fragments. For that reason, the best resolution of fragments larger than about 2 kb is attained by applying no more than 5 volts per cm to the gel (the cm value is the distance between the two electrodes, not the length of the gel). For PCR products smaller than 600 bp, separation is better and bands are sharper if gels are run very fast (3-4 hours for a 15-20 cm long 2-3% agarose gel). When the same gel runs at a low voltage overnight (14-16 hours) the products become less separable due to the diffusion in the gel.

Staining and viewing gel:

- When finished, switch off the power pack and remove lid from tank.
- Take out casting tray with gel inside and slide gel out from the tray.
- Carefully place the gel into the 0.5 mg/ml ethidium bromide in 1 X TBE tank to stain (wearing NITRILE gloves!) for 2-15mins, depending on the age of the stain.
- Place stained gel under UV light.