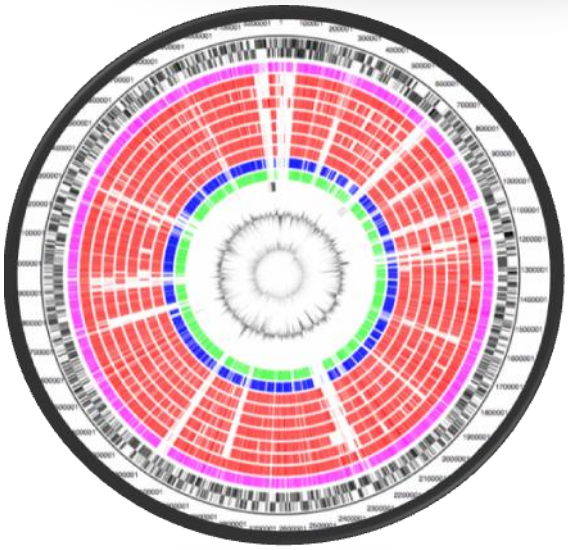
The Application of Molecular Epidemiology to Investigations of Foodborne Disease Outbreaks: Current Status and Future Plans

Report of a Workshop held 17th January 2012, London, UK



Introduction

Following the *E. coli* O104 outbreak in Germany and other incidents it has become clear that outbreaks of foodborne disease can cause serious public health consequences as well as intense political and trade issues. From a policy and regulatory perspective, it is imperative to bring any advances in knowledge and understanding of molecular biology to assist investigations of outbreaks of foodborne disease in the UK. A particular aim should be to utilise advances in molecular biology to assist in identification of the source of foodborne disease outbreaks in order to take the appropriate action to protect consumers.

To help achieve this aim a workshop was organised jointly by the Food Standards Agency (FSA), the Health Protection Agency (HPA), the Biotechnology and Biological Sciences Research Council (BBSRC) and the Advisory Committee on the Microbiological Safety of Food (ACMSF) to gain an understanding of how molecular biology could assist traditional microbiological and epidemiological approaches to outbreak investigations. Experts from key stakeholder organisations were brought together to discuss the current state of the science, how it is likely to develop over the next few years and to assess whether there are specific gaps in knowledge and capacity that need to be addressed.

This report contains a summary of the presentations, break-out sessions, major points of discussion, conclusions and action points from the day.

- The agenda for the workshop is given in appendix 1
- The participant list is given in appendix 2
- Slides from the presentations are given in appendix 3
- Pre-workshop feedback provided by participants is given in appendix 4

Summary and key recommendations

This was a lively workshop with 32 experts from a range of stakeholder organisations (see appendix 2) discussing the latest developments in molecular technologies and the role such approaches could play in improving the prediction, detection and management of foodborne disease outbreaks.

There was great enthusiasm for the potential of the newer generations of sequencing tools, which were widely believed to be capable of providing improved information over currently used approaches at a comparable or lower cost. Clear examples of where such approaches could immediately add significant value if used in an outbreak situation were described.

Discussions around the practical steps required to facilitate efficient, appropriate and timely transfer of the technologies from high-throughput academic centres of excellence to reference laboratories and eventually to front-line laboratories occurred during the afternoon break-out sessions. Issues such as staff training, data handling, storage and interpretation, availability of reference databases, backwards comparability to historical datasets and quality systems likely to be required were covered.

The plenary discussions generated the following key conclusions and recommendations:

1. High-throughput sequencing is currently capable of providing a significant benefit to outbreak investigations and should be used from the next outbreak onwards
2. Whilst technical, quality, logistical and training issues need to be addressed these issues should not delay initiation of the roll-out of the technology
3. Roll-out is expected to be in phases, with transfer from academic centres of excellence where the technologies are already established to key “early-adopter” clinical and reference laboratories occurring immediately, transfer to all reference laboratories within the next 2-5 years and to all front-line laboratories (clinical, epidemiological, food, animal) within the next 5-10 years

4. An audit of current methods should be undertaken to help the Agency understand how soon such techniques could be widely replaced by sequencing
5. Efficient implementation will require co-operation between multiple funding bodies (FSA, DEFRA, DH and funding councils) and funding should be allocated to aid the transition
6. Consideration and implementation of standardisation as sequencing technologies become more widely used will be crucial. Standardisation of methods should be considered where possible, alongside clear method performance assessments to identify sources of variability and uncertainty between the different methodologies
7. It should be possible to train current laboratory staff to generate the data, but analysis of the data will require further specialist training and expertise
8. Interpretation software that is widely accepted, easy to understand and interpret needs to be developed
9. There is a current lack of well-curated, inter-operable and quality controlled databases containing microbial sequences. Examples of where such databases do exist include a DEFRA/FSA supported campylobacter database (pubmlst.org/campylobacter). Funding should be allocated to help build and maintain wider pathogen sequence databases and to sequence historical isolates to help populate the databases
10. Standardisation in the recording of additional information is also needed. It is clear that to be of maximum benefit genome sequence information needs to be integrated with other information from environmental, clinical and animal data and this additional information needs to be collected at the time and not retrospectively
11. The Agency needs to engage with those developing policy at the clinical end (Public Health England) to help reduce the timescales and improve consistency of testing and reporting
12. The gaps in knowledge in animal populations also need to be considered and funding should be allocated to look at endemic levels of zoonotic pathogens

Presentations

Dr Andrew Wadge (FSA)
What is the policy need?

Andrew Wadge welcomed participants to the workshop, outlined its aims and gave an overview of the Agency's priorities in relation to foodborne disease food incident management.

- The second Agency funded study of Infectious Intestinal Disease (IID2 Study) identified up to 17 million cases of IID annually in the UK, with norovirus, and *Campylobacter* being common causes of infection
- Tackling *Campylobacter* in poultry has been identified as a key priority by the Agency, but work is also underway on *Listeria monocytogenes*, verocytotoxin-producing *E.coli* (VTEC) and norovirus
- Over 1500 incidents were dealt with by the Agency in 2010, of which 271 were microbiological in nature
- Some of these incidents involved human illness and the outbreaks of *E. coli* O104 in Germany and France in 2011 illustrate the challenges involved in managing outbreaks
- The Agency believes that advances in molecular epidemiology will have a role to play in furthering knowledge about the sources of infections, and the impact, on them through interventions in the food chain. However, clarity is needed on where and how these techniques can be applied to maximise impact.

Maria Zambon (HPA)
HPA perspectives on foodborne disease outbreaks

Maria Zambon described the six infection programmes at the HPA and outlined where molecular epidemiological approaches are currently being used. Examples included *Mycobacterium tuberculosis* strain typing, influenza serotyping and bloodborne viruses.

Data from foodborne outbreaks (1992 – 2010) was presented to demonstrate the burden of microbiological outbreaks and a discussion of what is needed to reduce the burden followed. This included:

- Effective use of public resources to ensure added value
- Increased focus on quality and standardisation of new technologies
- Rapid and accurate response capability is essential
- International dimension (International Health Regulations (IHR) and European Union regulations) must be considered.

A need for robust data handling packages was highlighted, calling on experience from the virology community on how to use and curate databases. The need to train staff to adopt the necessary informatics skills was also highlighted.

The requirement to understand the limitations of whole genome sequencing (WGS) was discussed and data from a study showing non-concordant data quality assessments from different next generation sequencing (NGS) platforms used to highlight the issue. The take home message was that validation of sequence data is equally as important as validation of traditional typing methods.

Sarah O'Brien (ACMSF) and John Wain (UEA)

What do we mean by molecular epidemiology?

Sarah O'Brien highlighted the fact that most of the definitions for molecular epidemiology involve a combination of molecular microbiology and clinical exposure data. However, in practice it is the study of organisms and organism changes that is often undertaken, without the inclusion of clinical or exposure data.

A discussion of molecular epidemiology challenges followed:

- Using molecular epidemiology in an outbreak situation is usually too late in the process to be of real use. Routine surveillance needs to use these technologies, but is not doing so at the moment. Without understanding diversity, changes cannot be tracked
- Molecular methods need to come first in the routine front-line laboratories with reference laboratories being used for particularly hazardous and difficult samples
- Making sense of the output quickly requires clinically and public health relevant reporting
- Carefully designed clinical and epidemiological studies are essential
- The function of a reference laboratory should be in the interpretation of data.

John Wain discussed whether NGS would have been useful in the Godstone Farm *E. coli* O157 outbreak. Sequencing of the isolates demonstrated that sequence data was more informative than VNTR typing, which obscured the true relationship between the strains.

However, to realise the full potential of sequencing data there is a need to understand how much infection is in the background and to integrate data on strains from livestock, food and human clinical samples.

Additional considerations include the need to agree on which regions of the genome to compare and to what. Tools to take a genome sequence and determine if it is related to another will be critical.

Julian Parkhill (Sanger Centre)

What can NGS approaches contribute that other techniques cannot?

Julian Parkhill presented examples of four areas where NGS can contribute:

1. **Context** – understanding the source of an outbreak. In the 2010 cholera outbreak in Haiti NGS was able to define the relatedness of isolates across countries and continents, and also the timescales for geographical replacement
2. **Resolution** – discriminating between re-currence and re-infection. Sequencing and clustering of *Salmonella Typhimurium* ST313 and MRSA ST239 isolates were used as examples
3. **Comprehensiveness** – understanding global population structures. *Streptococcus pneumoniae* 23F was used as an example with NGS identifying 23,000 SNPs from global isolates. This data was used to map macrolide resistance acquisition, serotype switches and capsular replacements
4. **Speed** – Mapping an MRSA outbreak in Addenbrooks hospital was given as an example. A retrospective analysis showed that full sequence data could have been generated within 48 hours and used to confirm an outbreak, demonstrate a previously undetected transmission event, assemble a resistome (drug resistance) and toxome (toxin gene) database and demonstrate the presence of a mutator.

Lisa Crossman (TGAC)

What are the strengths and limitations of NGS and how do we future-proof this approach?

Lisa Crossman described the current microbial sequencing programmes at TGAC and used the German *E. coli* outbreak as an example of the manner in which next generation sequencing is helping to deal with outbreaks. This was followed by an overview of the current major NGS technologies and the relative strengths and weakness of the systems including read length, accuracy, price, ease of use and ability to cope with homopolymers.

The role of NGS in molecular epidemiology was then discussed:

- NGS is rapidly becoming faster and cheaper than traditional approaches
- Lots of data is generated:
    - Whole organisms can be defined
    - Expression profiling and short RNA analysis can be undertaken
- Limitations include sequence accuracy and discriminating sequence errors from SNPs:
    - Requirement to validate with PCR and small scale sequencing at high cost, or ensure sufficient coverage such that the depth of sequence information can enable bioinformatics to identify sequence errors
- Data storage was highlighted as a major issue as the cost of hard disk storage is not reducing as quickly as the cost of NGS technology
- Other considerations include:
    - Value for money
    - Balance between depth of coverage and accuracy
    - Identifying and dealing with bottlenecks – downstream assembly and bioinformatics.

The presentation ended with a discussion of requirements for future–proofing, which included the need for collaboration, communication, cloud computing, database formation and multidisciplinary approaches. The need to integrate microbial data with human sequencing data was discussed alongside the emergence of new longer-read platforms, single cell analysis and metagenomics.

John Cowden (HPS) and John Coia (ACMSF)

What is the difference between steady-state management and outbreak management?

John Cowden gave an epidemiologist's perspective on subtyping and described the need to understand its aims in terms of:

- Outbreak detection
- Trend analysis
- Outbreak investigation

- Research
- Individual patient management.

The strategy selected should depend on the aim and should be evaluated in relation to all relevant criteria (including sampling and logistics etc.) as opposed to just the typing technique.

The level of information needed also depends on the aims and may consist of studying all cases identified, representative subsets, linked cases or numbers dictated by research requirements. The appropriate technologies to use will also depend on the aims and should be sufficiently discriminatory for the purpose and cheap/cost effective.

Barriers to uptake by clinicians were discussed and included the use of techniques that were too detailed, too general, too volatile or too expensive/time consuming.

John Coia gave a laboratory's perspective on subtyping. For outbreak detection and trend analysis the same approaches may often be used, but different populations may be sampled. For outbreak investigation a technique that has sufficient resolution to discriminate outbreak isolates from non-outbreak isolates is needed.

A discussion on the variability in current practice followed which highlighted that not all samples are routinely forwarded by clinicians to the front line laboratories and once the sample arrives at the laboratory there is variability in what is tested for. The need for standardised approaches and technologies was highlighted, by way of examples given.

Determining whether strains are related (same) or not will require the generation and use of data from background populations.

The presentation ended with a discussion of the pros and cons of molecular approaches and included the objectiveness of molecular methods and the need to standardise to enable comparability. The importance of bioinformatics tools was again stressed.

Break-out sessions

Participants were divided into two break-out groups to discuss issues in more detail.

Group 1 – Applications – How can we deploy NGS?
Group 2 – Resources – What is the current capacity and facilities?

An outline of the more detailed questions considered by the break-out groups is provided in the workshop agenda in appendix 1.

Group 1

- Priority organisms include *Campylobacter*, *Salmonella*, Verocytotoxin-producing *E. coli* (VTEC) and *Listeria monocytogenes*
- Useful current collections of samples were highlighted within various organisations and considerations for future collections given. These include collections that are well-structured, available, accessible, documented and international
- The need for representative samples and the need to integrate with clinical, zoological and environmental data in an easily accessible way was highlighted as a key issue
- Data collected should be shared widely and appropriately once anonymity has been assured
- A large number of databases were not considered to be necessary, but it is essential that database structures are compatible, quality is assured and inter-operability and linkage is preserved
- Technology should be rolled out immediately in a consistent and staged approach.

Group 2

- The consensus was that there are currently enough trained people capable of generating NGS data but not enough capable of analyzing it appropriately. However, this should not delay the roll-out of technologies and training should occur in parallel
- Whether facilities currently have the appropriate accreditation and quality standards was not considered a critical issue as the field is changing so quickly. It is more important to consider how things will look in a few years
- Over the next two years it is likely that all the main clinical and research laboratories will have NGS capability and that within five years all reference laboratories will have the capability
- Sequencing approaches are currently cheaper to use than older techniques in some high throughput laboratories and as the cost is reduced further it will become cost effective for more laboratories to roll out the technology
- Delaying the roll-out will cost more
- All outbreaks from today onwards should involve full genome sequencing and sequencing of relevant historical collections to inform decision making.

Plenary discussion and conclusions from the day

It was felt that most participants in the room were converts to sequencing approaches and that the community needed to "just do it" whilst seeking to resolve highlighted issues in parallel. In the next foodborne outbreak we anticipate NGS being applied to human, food and environmental samples as part of the investigation process.

Sequencing of historical samples does not need to occur before implementation of the technologies, and sequence databases will become self-populating over time. However, funding should be allocated to sequence historical isolates or samples where these are likely to add value. The gaps in knowledge in animal populations also need to be considered.

The transition to sequencing is expected to occur in phases, with high-throughput laboratories and universities already using the technologies, reference laboratories expected to transition within two to five years and routine testing laboratories in the next five to ten years. One or two early adopter trail-blazers will help to incentivise physicians by demonstrating the utility of the technologies.

Consideration and implementation of standardisation as sequencing technologies become more widely used will be crucial. Standardisation of methods should be considered where possible, alongside clear method performance assessments to identify sources of variability and uncertainty between the different methodologies.

Standardisation in the recording of additional information is also needed. It is clear that to be of maximum benefit genome sequence information needs to be integrated with other information from environmental, clinical and animal data and this additional information needs to be collected at the time and not retrospectively.

Metadata, data sharing and interpretation will be key issues and the curve of implementation from early adopters to routine laboratories will require a funding commitment to aid the transition. The development of interpretation software that is widely accepted and everyone can understand could be an example of where funding would be needed.

A recommendation for an audit of current methods was made to help the Agency understand how soon such techniques could be replaced by sequencing.

The Agency also needs to engage with those developing policy at the clinical end (Public Health England) to help reduce the timescales and improve consistency of testing and reporting.

Appendix 1 – Workshop Agenda

THE APPLICATION OF MOLECULAR EPIDEMIOLOGY TO INVESTIGATIONS OF FOODBORNE DISEASE OUTBREAKS: CURRENT STATUS AND FUTURE PLANS

17 January 2012

Grand Connaught Rooms, 61-65 Great Queen Street, London WC2B 5DA

Agenda

Morning Session Chair – Dr Andrew Wadge

Item	Lead	Time
1. Welcome and introductions	Andrew Wadge	10:00- 10:05
2. What is the policy need? - citing German *E. coli* outbreak and lessons learned as an example	Andrew Wadge	10:05 – 10:15
3. HPA perspectives on foodborne disease outbreaks	Maria Zambon	10:15 -10:25
4. What do we mean by molecular epidemiology?	Sarah O'Brien & John Wain	10:25 – 10:50
5. What can next-generation sequencing approaches contribute that other techniques cannot?	Julian Parkhill	10:50 – 11:15
Tea/coffee		11:15 – 11:30
6. What are the strengths and limitations of next-generation sequencing and how do we future-proof this approach?	Lisa Crossman	11:30 – 12:00
7. What is the difference between steady-state management and outbreak management?	John Cowden and John Coia	12:00 – 12:30
8. Introduction to afternoon breakout group aims etc	Paul Cook	12:30 – 12:45
Lunch		12:45 – 13:30

9. **Breakout Group 1 (Parallel session with 10 below)** How can we deploy next generation sequencing in foodborne outbreak investigation to best effect? Which may *inter alia* address the following - - which microorganisms should be prioritised?	Chair – Paul Cook Rapporteur – John Cowden .	13:30 – 14:30

(Campylobacter, VTEC, Salmonella, L. monocytogenes) - what collections are available that could potentially be used? - what type and quality of samples would be needed? - how much and what type of data would be needed? - what should we do with the data? - how should we make the strains and data available? - how many databases would be needed?		
10. **Breakout group 2 (Parallel session with 9 above)** What is the current availability and capacity of relevant facilities? Which may *inter alia* address the following - are there enough appropriately trained staff available to analyse as well as generate the data? - do the facilities available have appropriate accreditation and quality standards? - what are the likely costs? - given current available resources how would we do this?	Chair – Gerry Hoad Rapporteur – Adam Staines.	13:30 – 14:30
Tea/coffee		14:30 – 14:45

Session Chair – Prof Sarah O'Brien

11. Feedback from parallel sessions	Group rapporteurs	14:45 – 15:15
12. Plenary Discussion – including • Could lessons learned by other organisations inform the approach • What do we need to do next to make this happen?	Sarah O'Brien	15:15 – 16:00
13. Close		16:00

Output: Workshop report.

Appendix 2 - List of Participants

Molecular Epidemiology Workshop – Tuesday 17 January 2012

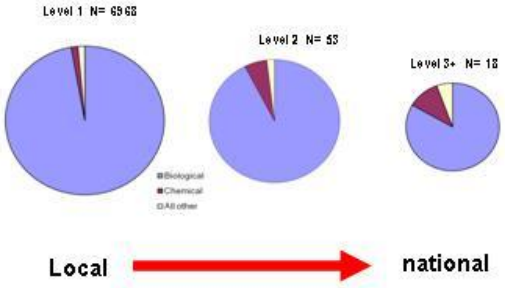
Participants	(break out group)		Organisation
Dr Bob Adak	(1)	-	Health Protection Agency (HPA)
Dr Jo Aish	(2)	-	Food Standards Agency (FSA)
Dr Roy Betts	(2)	-	Advisory Committee on Microbiological Safety of Food (ACMSF)
Prof Mark Blaxter	(1)	-	University of Edinburgh
Dr Derek Brown	(2)	-	Scottish Salmonella Reference Lab
Prof John Coia	(1)	-	ACMSF
Dr Paul Cook	(1)	-	FSA
Dr John Cowden	(1)	-	Health Protection Scotland (HPS)
Dr Lisa Crossman	(2)	-	The Genome Analysis Centre (TGAC)
Dr Richard Ellis	(1)	-	Animal Health and Veterinary Laboratories Agency (AHVLA)
Dr Ken Forbes	(2)	-	University of Aberdeen
Dr Carole Foy	(1)	-	LGC
Dr Vanya Gant	(1)	-	University College London Hospitals (UCLH)
Dr Kathie Grant	(1)	-	HPA
Dr Jonathan Green	(2)	-	HPA
Dr Geraldine Hoad	(2)	-	FSA
Dr Rebecca Hodges	(1)	-	Medical Research Council (MRC)
Dr Jane Ince	(2)	-	FSA
Prof Rowland Kao	(1)	-	University of Glasgow
Prof Doug Kell	(2)	-	Biotechnology and Biological Sciences Research Council (BBSRC)
Asst Prof Mette Voldby Larsen	(2)	-	The Technical University of Denmark (DTU)
Prof Martin Maiden	(1)	-	University of Oxford
Prof Duncan Maskell	(2)	-	University of Cambridge
Prof Sarah O'Brien	(1)	-	ACMSF
Dr Julian Parkhill	(1)	-	The Sanger Centre
Dr Norval Strachan	(1)	-	University of Aberdeen
Dr Adam Staines	(2)	-	BBSRC
Dr Andrew Wadge	(1)	-	FSA
Dr John Wain	(1)	-	University of East Anglia (UEA)
Dr Alan Walker	(2)	-	The Sanger Centre
Prof Brendan Wren	(1)	-	London School of Hygiene and Tropical Medicine (LSHTM)
Prof Maria Zambon	(1)	-	HPA

Appendix 3 – Presentation materials

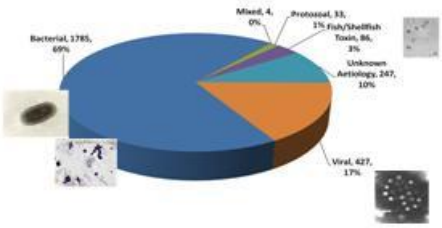
HPA perspectives on foodborne disease outbreaks

Key Health Protection Infection Programmes

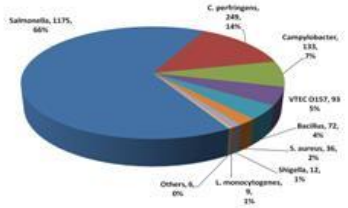
Outbreaks & Incidents 2009-2011
Local to national



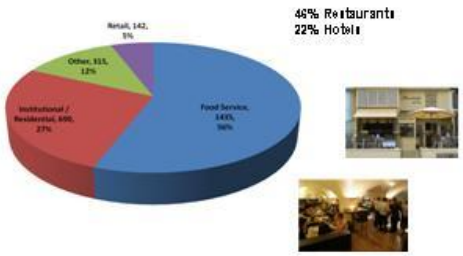
Food-borne Outbreaks in England & Wales, 1992-2010 – Aetiology (n=2582)



Food-borne Outbreaks in England & Wales, 1992-2010 Pathogens – Bacteria (n=1785)

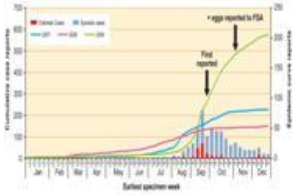


Food-borne Outbreak settings, 1992-2010



Outbreak vs Sporadic

Salmonella Enteritidis PT 14b Epidemic Curve 2009

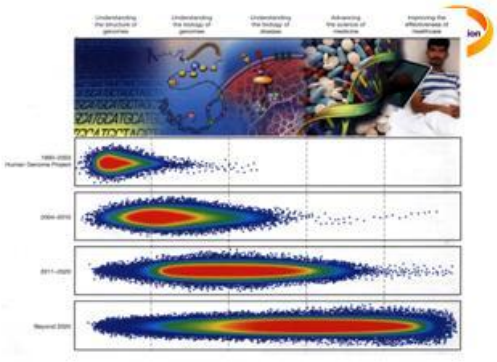


E.Coli O157

New technologies & Public Health

What do we need to solve the problem
Effective use of public resources. Added Value.
Focus on quality and standardisation
Response capability (fast & accurate)
International dimension IHR & EU

Achieving effective disease surveillance systems



Service Developments and Service Delivery

- Clear strategic objectives
- Partnership working
- Mutual respect
- Learn from Virology experience
- Robust Information handling strategy
- Staff Skills & workforce planning

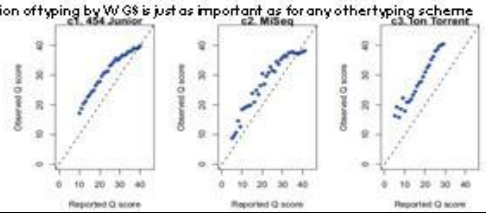
WGS Understanding the limitations

Ecoli 104 Direct comparisons

Different technologies: Ion Torrent in Birmingham, Roche 454 Junior at Colindale and Illumina Mi-seq at Great Chesterford, gave different quality data.

The analysis of the data is key – quality control, quality assurance will use the data rationally –

validation of typing by WGS is just as important as for any other typing scheme



UNIVERSITY OF LIVERPOOL

What do we mean by molecular epidemiology?

Sarah J O'Brien

Definition

"The contribution of potential genetic and environmental risk factors, identified at the molecular level, to the aetiology, distribution and prevention of disease within families and across populations."

Model analogous to that of traditional and clinical epidemiology - investigate levels of disease prevalence and incidence with respect to exposure to various risk factors.

Molecular epidemiology?

- Epidemiology of bacterial pathogens and evolution
  - To explain how virulence and other phenotypic traits evolve in bacterial species over time

Baker *et al*, Curr Opin Microbiol 2010: 13: 640-5.

- Identify genes and genetic elements that encode resistance

O'Brien & Stelling, Clin Microbiol Rev 2011: 24: 281-295.

Challenges

- Real-time technology or tool for post-hoc rationalisation?
  - How often does molecular epidemiology incorporate clinical or epidemiological exposure data?
  - How often does molecular epidemiology influence case definitions in foodborne disease outbreak investigations?
- Use in outbreaks is (usually) too late
  - Routine surveillance also needs to incorporate use of these technologies

Challenges

- Making sense of the output quickly
  - Clinically relevant reporting
  - Public health relevant reporting
- To realise benefits fully will need carefully designed clinical and epidemiological studies

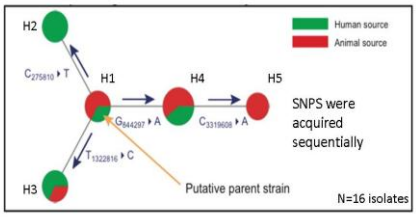
What do we mean by molecular epidemiology?

John Wain

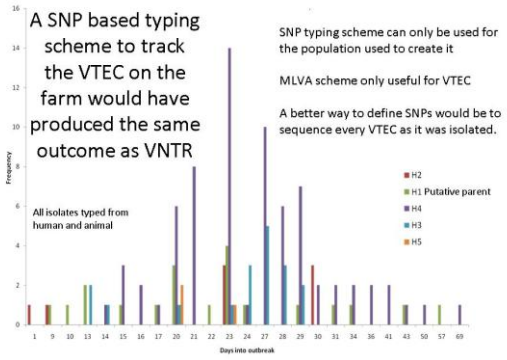
HPA and The University of East Anglia

Would NGS have been useful during the Godstone farm outbreak?

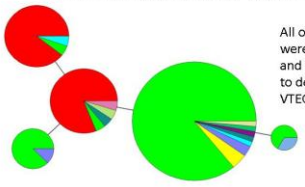
From the Godstone outbreak 16 isolates were sequenced and analyzed at the Sanger. Analysis of the genomes at the HPA revealed four SNPs (of high quality after manual inspection) across the whole chromosome and we then designed a SNP assay to test all isolates. Seven SNPs were initially called but 3/7 rejected. An error of three in 5million was enough to change the outcome of the analysis.



A SNP based typing scheme to track the VTEC on the farm would have produced the same outcome as VTNR



SNPs have higher information content than Multi Locus VNTR Analysis



All of the VNTR profiles were single locus variants and VNTR typing was used to detect this outbreak of VTEC O157:H7....

.....but within the outbreak the same VNTR type was associated with multiple haplotypes.

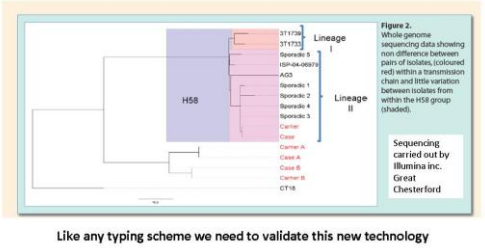
- Rapid repeat number changes occur in VNTRs which obscures the true relationship between strains
- Single locus changes occur during outbreaks

Interpreting genome sequence data

.....PFGE showed that the case and carrier were different — HPU reluctant to act but we now know that genomic rearrangement can change PFGE patterns and....

.....cases and carriers have the same genome sequence – red highlights

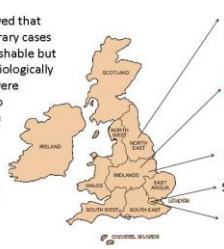
.....most of our imported Typhi have very similar genome sequences - H58 globally dominant



Like any typing scheme we need to validate this new technology

Example 2: Tracking typhoid carriers

Sub-typing showed that most contemporary cases were indistinguishable but that the epidemiologically linked isolates were different (by two typing methods)



Sporadic5 - PT E1, MDR

Case - PT E9var, MDR – PFGE 2
Carrier- PT Untyp Vi, MDR – PFGE 1
Linked epidemiologically within UK

Sporadic4 - PT E9var, MDR – PFGE 2

Sporadic2 - PT E9var, MDR – PFGE 2

Sporadic3 - PT E9var, MDR – PFGE 2

Case and carrier had identical genomes

Conclusion: no transmission of typhoid in the UK – sporadic cases were linked to travel

Concerns – the current typing schemes cannot differentiate between the *S.* Typhi isolates being imported into the UK from the ISC

Challenges

Making sense of the output quickly for clinical and public health reporting

Which region of the genome do we compare to what?

To realise benefits fully will need carefully designed clinical and epidemiological studies

What can next-generation sequencing approaches contribute that other techniques cannot?

Julian Parkhill

FSA workshop Jan 2012

What can next-generation sequencing approaches contribute that other techniques cannot?

- Context
- Resolution
- Comprehensiveness
- Speed

What can next-generation sequencing approaches contribute that other techniques cannot?

- Context
- Resolution
- Comprehensiveness
- Speed

The Haitian cholera outbreak

-April 2004 United Nations Stabilization Mission

-Jan' 2010 earthquake.

-Sept' 2010 Additional Nepalese UN troops

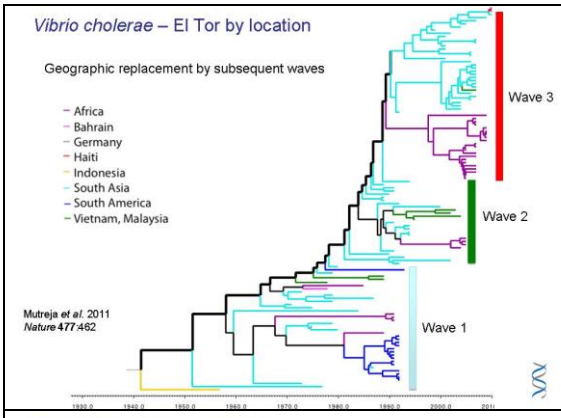
-Oct' 2010 -cholera appeared in Haiti
(1st first time in a century).

-February 2011
>4,500 deaths & 300,000 infected.

Vibrio cholerae – El Tor by location

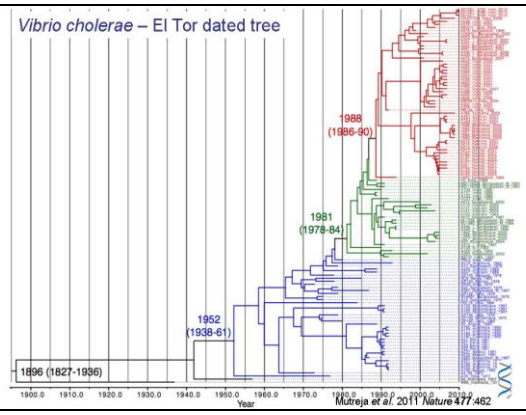
Geographic replacement by subsequent waves

- Africa
- Bahrain
- Germany
- Haiti
- Indonesia
- South Asia
- South America
- Vietnam, Malaysia



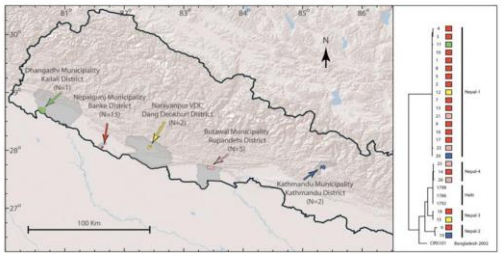
Mutreja *et al.* 2011
Nature 477:462

Vibrio cholerae – El Tor dated tree



Mutreja *et al.* 2011 *Nature* 477:462

The Haitian cholera outbreak



Rene S. Hendriksen 2011 mBio 2(4):e00157-11

What can next-generation sequencing approaches contribute that other techniques cannot?

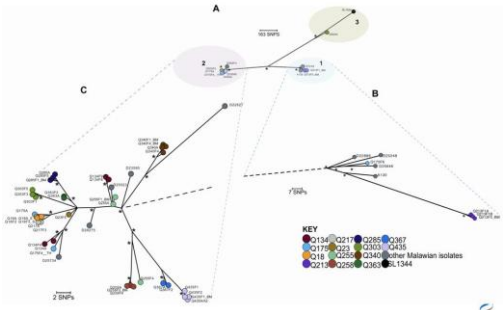
- Context
- **Resolution**
- Comprehensiveness
- Speed

Salmonella Typhimurium ST313 recrudescence

Index isolate	Date of isolation	Recurrence isolate 1	Date of isolation	Recurrence isolate 2	Date of isolation	Recurrence isolate 3	Date of isolation	Recurrence isolate 4	Date of isolation
Q18A	28/03/2002	Q18A_S	28/03/2002	Q18F2	24/05/2002	Q18F3_S	22/06/2002		
Q23A	26/02/2002	Q23F1	25/03/2002	Q23F2_BM	26/04/2002	Q23F5	26/03/2003		
Q134A	30/05/2002	Q134F4	16/10/2002	Q134F8	28/04/2003	Q134F9	06/09/2003		
Q175A	01/07/2002	Q175F4_TH	27/01/2003	Q175F6	07/04/2003				
Q213F1a	02/09/2002	Q213F1b	05/09/2002	Q213F3_BM	21/10/2002				
Q217A	18/03/2002	Q217F3	12/11/2002						
Q255A	17/09/2002	Q255F1_BM	10/10/2002	Q255F2_U	13/11/2002	Q255F4	26/03/2003		
Q258A	21/10/2002	Q258F3_BM	07/01/2003	Q258F4	02/03/2003				
Q285A	22/08/2002	Q285F1_BM	11/11/2002	Q285F5	22/04/2003				
Q303A	26/11/2002	Q363F1	20/12/2002	Q303F3	02/03/2003	Q303F5	22/04/2003		
Q340F1	10/01/2002	Q340F1_BM	10/01/2003	Q340F4	11/05/2003	Q340_F4BM	11/05/2003		
Q363A	03/12/2002	Q363F3	27/03/2003						
Q367A	31/12/2002	Q367F3	17/03/2003						
Q435A	02/03/2003	Q435A2	02/03/2003	Q435F1	07/04/2003	Q435F1_BM	14/04/2003	Q435F2	18/06/2003

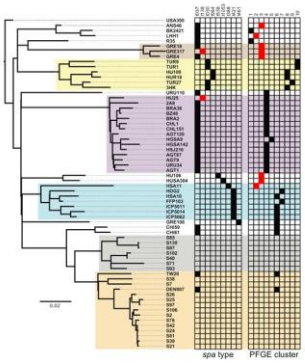
Okoro *et al. CID in press*

Salmonella Typhimurium ST313 recrudescence



Okoro *et al. CID in press*

Comparison with other typing methods: MRSA

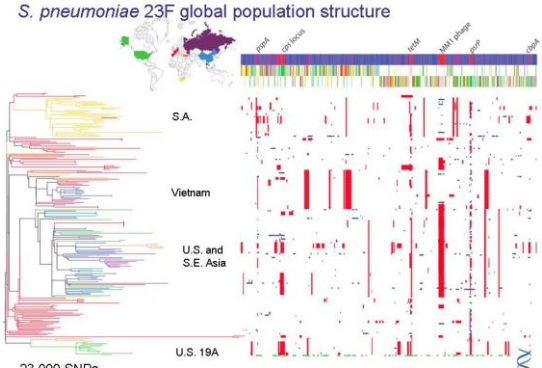


Simon Harris

What can next-generation sequencing approaches contribute that other techniques cannot?

- Context
- Resolution
- **Comprehensiveness**
- Speed

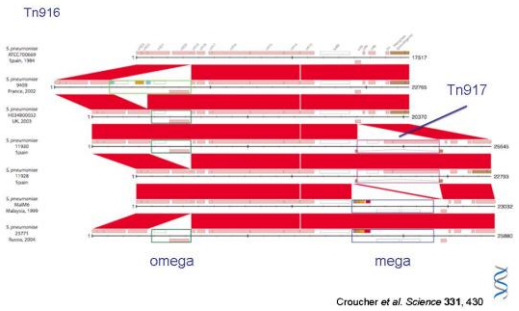
S. pneumoniae 23F global population structure



23,000 SNPs
88% of variants are due to recombination

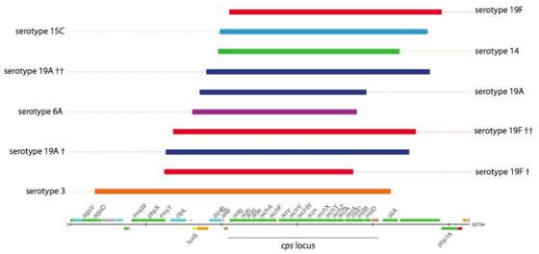
Croucher *et al. Science* **331**, 430

Macrolide resistance acquisition in *S. pneumoniae* 23F



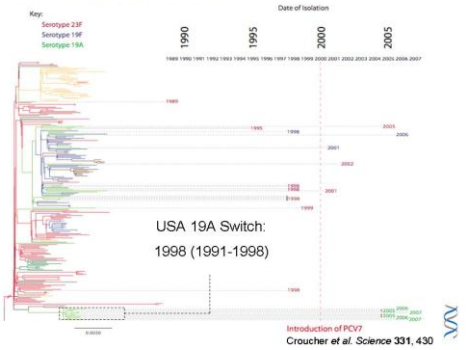
Croucher *et al. Science* **331**, 430

Serotype switches *S. pneumoniae* 23F



Croucher *et al. Science* **331**, 430

Capsular replacement in *S. pneumoniae* 23F

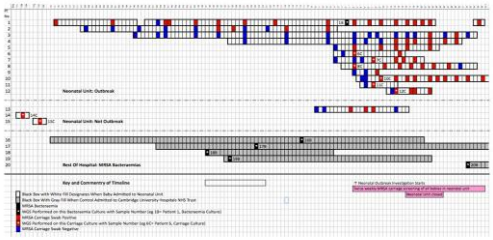


Croucher *et al. Science* **331**, 430

What can next-generation sequencing approaches contribute that other techniques cannot?

- Context
- Resolution
- Comprehensiveness
- Speed

Sequence mapping of an outbreak



Black – MRSA bacteraemia
Red – MRSA colonization
Blue – MRSA negative

Sequencing technology and approach

Illumina MiSeq
Read length 150 nt paired end
Multiplex 12 genomes in 27 h

7 'outbreak' MRSA and 7 'non-outbreak' MRSA sequenced

NB: Other rapid sequencing technologies are available

Mapping to a reference genome

NICU outbreak

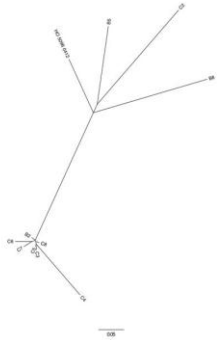
NICU, control

Not NICU, control

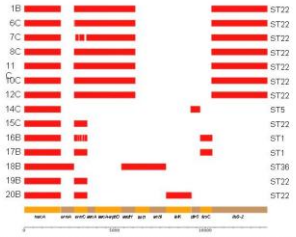
Sample	ST	SCCmec	CXT	ERY	CLIN	CIP	GEN	KAN	TOB	TET	TMP	RIF	FUS	MUP	LIN
1B	22	IV	R	R	R	R	R	R	R					R	
6C	22	IV	R	R	R	R	R	R	R					R	
8C	22	IV	R	R	R	R	R	R	R					R	
7C	22	IV	R	R	R	R	R	R	R					R	
10C	22	IV	R	R	R	R	R	R	R					R	
11C	22	IV	R	R	R	R	R	R	R					R	
12C	22	IV	R	R	R	R	R	R	R					R	
14C	5	IV	R								R	R			
15C	22	IV	R	R	R	R									
16B	1	IV	R	R	R								R		
17B	1	IV	R	R	R								R		
18B	36	II	R	R	R	R		R	R						
19B	22	IV	R	R	R	R									
20B	22	IV	R	R	R	R					R				

- ST22 mapped to EMRSA-15 reference
- Mapping coverage 33x to 111x
- % mapped 96.9% to 97.8%

Unrooted Maximum Likelihood tree of ST22 isolates

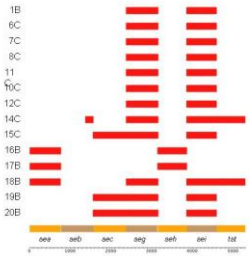


Resistome



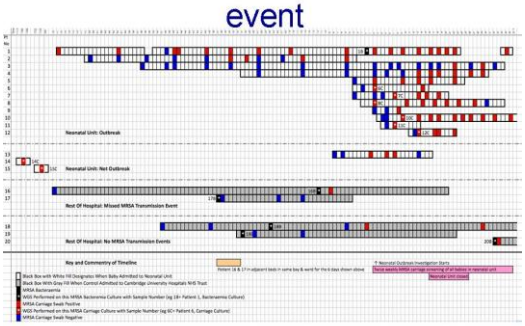
tBlastx of *de novo* assemblies of each isolate against a panel antibiotic resistance genes. Blast matches with a score >30 and a %identity >90% are displayed.

Toxome

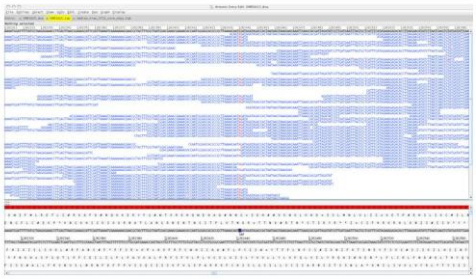


tBlastx of *de novo* assemblies of each isolate against a panel toxin genes. Blast matches with a score >75 and a %identity >90% are displayed. In 14C there is a truncated match to *entB*, examination of the assembly shows that this is a match to a entB gene fragment (66 amino acid fragment).

ST1 – a missed transmission event



*mut*S mutation



Achievements within 48 hours

- Confirmed an outbreak
- Demonstrated a previously undetected transmission event
- Assembled a resistome
- Assembled a toxome
- Demonstrated the presence of a mutator
- Demonstrated the basis for a SCV

WTSI:

Gordon Dougan
Nick Thomson
Stephen Bentley
Matthew Holden
Simon Harris
Nick Croucher
Ankur Mutreja
Thomas Connor
Mike Quail
Carol Churcher

Cambridge:

Sharon Peacock
Estee Torok
Claudio Köser
Matthew Ellington
Ed Cartwright

MRSA:

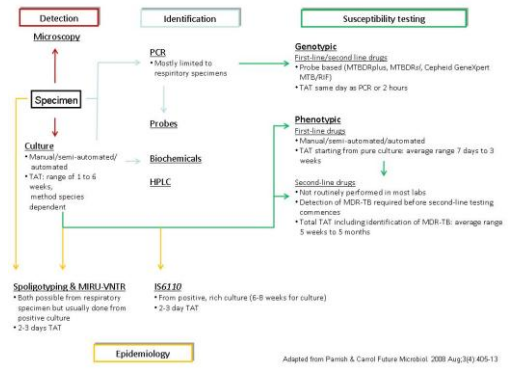
Bath
Ed Feil
Mahidol/Oxford
Sharon Peacock
Emma Nickerson
Narisara Chantratita
Nick Day
St George's
Jodi Lindsay
Jonathan Edgeworth
Universidade Nova de Lisboa
Herminia de Lencastre
Susana Garadete
Ana Tavares

S. pneumo

Emory University, USA
Keith Klugman
NICD, South Africa
Anne von Gottburg
CDC, USA
Lesley McGee
ARFID, South Korea
Kwan Soo Ko
OUCRU, Vietnam
Steve Baker
SSI, Denmark
Lotte Lambertsen
NRCS, Germany
Mark van der Linden
Glasgow University, UK
Tim Mitchell
HPA, UK
Bruno Pichon
Imperial College, UK
Bill Hanage

V. cholerae

IVI, South Korea
Dong Wook Kim
Je Hee Lee
Seon Young Choi
Eun Jim Kim
John D Clemens
Cecil Czerkinsky
Seoul National U.
Jongsik Chun
KEMRI, Kenya
Samuel Kariuki
U. Gothenburg
Jan Holmgren
Michael Lebens
NICED, India
G Balakrish Nair
Swapan Kumar Niyogi
T. Ramamurthy
U. Cambridge
James Wood



Adapted from Parrish & Carrol Future Microbiol. 2008 Aug;3(4):405-13

TGAC
The Genome Analysis Centre™

Strengths and limitations of next generation sequencing and how do we future-proof this approach?

Dr. Lisa Crossman
Microbial Genomes Project Leader, TGAC

The Genome Analysis Centre

TGAC - Microbial projects

Single species – microbial ecology projects
– food microbiology projects
– human health microbiology projects
– plant pathogenic bacterial projects
– fungal projects

Resequencing projects
– *Pseudomonas aeruginosa* - U. Strathclyde
– *Listeria monocytogenes* - HPA
– polyploid yeast project

Metagenomics projects

Summary

Strengths and limitations of the next generation sequencing platforms

Strengths of next generation sequencing for molecular epidemiology

Limitations of next generation sequencing for molecular epidemiology

Future-proofing

Next generation sequencing technologies

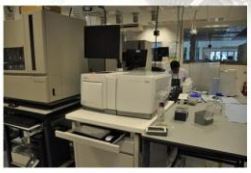
Second generation sequencing technologies

454 - Roche
Illumina
SOLiD – Life Technologies

454 (Roche)

Basic approach

Bead – PCR – light detection



Advantages

Comparatively long read length

Major disadvantage

Homopolymers poorly resolved

Illumina

Basic approach

Bridge amplification

Advantages

Price

Disadvantage

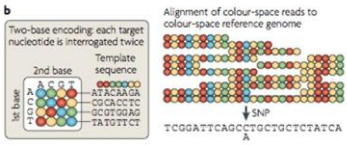
Short read length

SOLiD (Life Tech)

Sequencing by ligation

Advantage
High accuracy

Disadvantage
Colour space difficult to work with



Next Generation Sequencing
in Molecular epidemiology

Molecular epidemiology for outbreaks

- German outbreak of *E.coli* 0104:H4 associated with beansprouts
- Slower health procedures could potentially lead to more deaths
- Have genomics data in place for rapid response

Spanish cucumbers

Speed and price

- NGS is fast and cheap
- Provides large amounts of data
- Define the whole organism

Transcriptomics/RNA-Seq

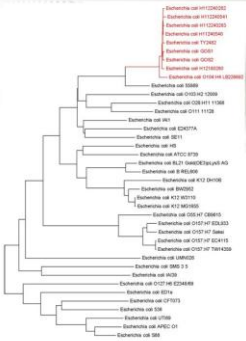
- We can look at the expression of all the genes under specific conditions and compare them
- Can look at short RNA

Tracing organisms by SNP data

SNPs are single nucleotide polymorphisms
SNVs are single nucleotide variants

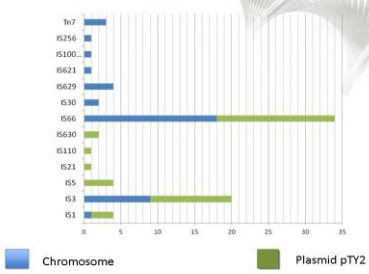
Both are single nucleotide changes
Technically a SNP is a change expected throughout a species
& a SNV is a change expected to be private to a given individual

Alignment free method



Kwan lab, crowdsourcing

Insertion sequences/Mobile elements



Limitations

Sequence Accuracy - 'RIRO'

sequencing errors/SNPs

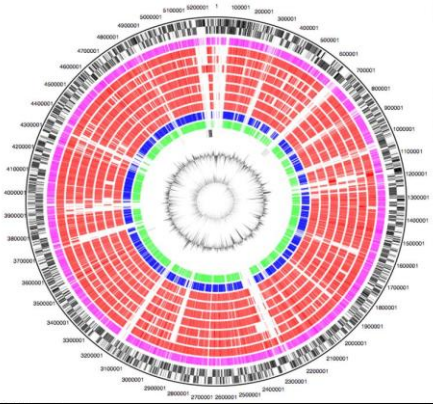
To check SNPs involves PCR and small-scale sequencing – high cost

Sufficient coverage to be clear about SNPs

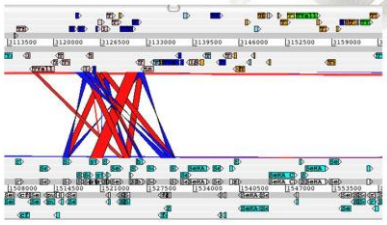
We could correct 'false SNPs' bioinformatically

Variable regions

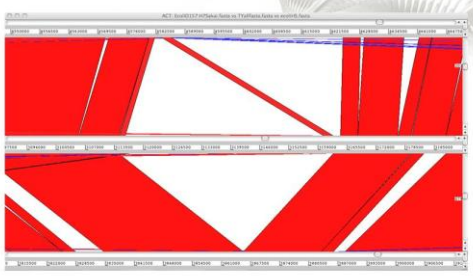
- Analysis of variable regions – mobile elements
- Plasmid analysis
- Phage analysis



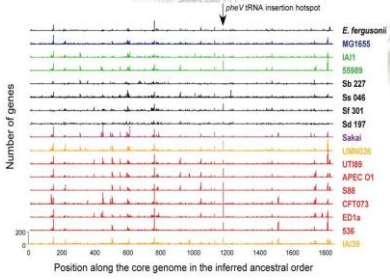
Variable Region VR11



Insertion hotspots



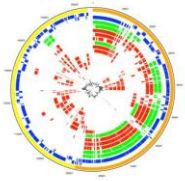
E. coli Insertion hotspots



Touchon *et al.*, 2009 PLOS Genet.

Ancient clues muddy the water

- Outbreak *E.coli* plasmids may share some similarities with those from avian pathogenic *E.coli*



TCGTCGTTCCGCCAAGTTCGT INC11 - pTY1

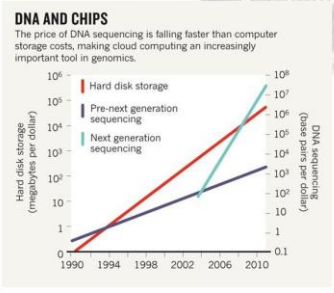
pTY2, has two *rep regions.*
BLAST – IncFII and repFIB.
The two reps bind to the reverse Frep and FIB

GAAGATCAGTCACACCATCC FREP
CTCCCGTCGCTTCAGGGCATT FIB

The Frep also binds to the Frep forward with one mismatch: binding sequence:
TGATCATTTAAGGAATTTTG published primer:
TGATCGTTTAAGGAATTTTG

Appl. Env. Micro. Johnson *et al.* 2007

Limitations – data storage



Other Considerations

- Value for money
- Balancing act between depth of coverage required and accuracy
- Bottlenecks – assembly & downstream bioinformatics analysis

Alternative approaches to compliment NGS

- Traditional epidemiology

- Mathematical approaches

Rossmo's formula

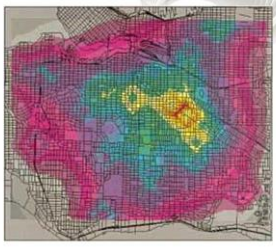
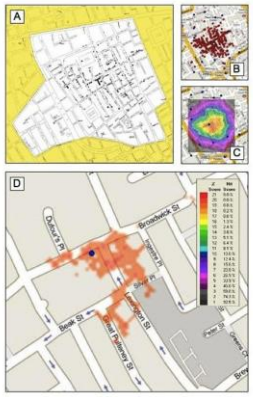


Figure 6.1

A geoprofile of a series of armed robberies in Vancouver, British Columbia, Canada. *Source: Det. Insp. D. Kim Rossmo, Vancouver, British Columbia, Police Department. Reproduced by permission.*



Rossmo's formula

Applied to Snow's cholera study and the broad street pump

Output

- Database formation
- Single or multiple species per database?
- Conflicting demands depending on end user
- Publications

Future-proofing

- Collaboration
- Communication
- Cloud computing
- Multidisciplinary approach

New Data Platforms

Third generation sequencing technologies

PacBio – much longer reads

New data types

- Integrate sequence data with phenomics, transcriptomics and proteomics

(Holistic definitions)

- Integration with human sequence data
- Single bacterial cell sequences

Can we ever have enough data?

Some diseases attributable to more than a single organism

Using metagenomics data?

Summary

NGS data can add much to a speedy health response

Accuracy of sequencing / coverage depth is important

Formation of an appropriate useful database

Typing gastro-intestinal pathogens: how much, how many, and why? - an epidemiologist's perspective.

John M. Cowden
London, 17 January 2012

"If anyone objects to any statement I make, I am quite prepared not only to retract it but also to deny under oath that I ever made it."

Tom Lehrer (1928 -)

How can germs be typed?

- Phenotype
  - e.g. serotype, phage type, AMS, biochemistry, smell
  (for cars: performance, colour, fluffy dice)
- Genotype/molecular type
  - e.g. PFGE, MLST, MLVA, deep sequencing, etc
  (for cars: make, model, VIN)

Epidemiologists don't care, as long as it serves our aims.

Analogy for levels of genotyping

- Road vehicle
- Car
- Austin
- Austin Mini
- Austin Mini Cooper
- Austin Mini Cooper S
- VIN/chassis number.

To identify a traffic jam:
"Road vehicle" is enough

Three Minis in Turin = a cluster, whatever model *(genotype)* or colour *(phenotype).*

To identify a specific car you need:

- **the registration number** ***(phenotype)***
  or (because phenotypes can change)
- **the VIN** ***(genotype).***

What are our aims?

- **Outbreak detection**
  - ***identifying greater than expected numbers of a particular germ/type/subtype***
- **Trend analysis**
  - ***tracking changing numbers or proportion of a particular germ/type/subtype***
- **Outbreak investigation**
  - ***explaining discrepant cases***
  - ***ruling cases in or out***
  - ***linking cases to vehicles/sources***
- **Research**
  - ***linking sporadic cases to vehicles/sources, etc.***
- **Individual patient management**
  - ***AMS.***

What strategy suits our aims?

It depends on **which aim**, and:

- **How sure** you want to be
- The **cost/benefit** balance
- What you can **afford**

"Horses for courses!"

Remember:
You are evaluating a strategy
not just a typing technique

(it's not just "How much?" but also "How many?).

How many isolates do we need to type?

- **Outbreak detection**
  - All *(well, all we get)*
- **Trend analysis**
  - representative sub-set *(or an approximation to)*
- **Outbreak investigation**
  - potentially linked cases
- **Research***
  - as scientifically dictated
- **Individual patient management***
  - as clinically dictated

* Not considered further.

Outbreak detection

The technique should be:

- Sufficiently discriminatory to make microbiological links between epidemiologically unlinked cases, e.g.:
  - *Salmonella serotype*
  - *if rare, might be enough*
  - *if common, supplemented by phage type*
  - *if rare, might be enough*
  - *if common, supplemented by (or replaced with) PFGE*
- Cheap (because universally applied)

Trend analysis

The technique should be:

- Sufficiently discriminatory to group cases microbiologically
  - *similar levels of discrimination as for outbreak detection, but applied to fewer cases*
  - *"any idiot can discriminate between campylobacters: it'll take a genius to group them meaningfully"*
- Fairly cheap (because applied systematically).

Outbreak investigation

The technique should be:

- Sufficiently discriminatory to rule cases in or out of an outbreak, or link them to a source
  - *Varies according to level of evidence required*
- Cost effective (because applied selectively).

What doesn't suit our aims

Techniques that are:

- too detailed
  - *linked cases look different*
- too general
  - *unlinked cases look the same*
- too volatile
  - *characteristics change on the plate*
- too expensive, or time consuming
  - *impractical.*

You can sometimes go too far…..

which is wasteful.

How does a policymaker decide on funding a new technique?

- **B**e clear:
  - about the strategy's aim (OD, TA, OI, R, CM)
- **A**ssess:
  - whether the technique does enough - or too much
- **R**eview:
  - if the strategy does what is needed of it - and how well
- **B**eware:
  - techniques which identify *"Distinctions without differences"*
- **E**valuate:
  - the strategy's cost, benefit, and affordability
- **R**emember:
  - *"Never ask a barber if you need a haircut."*

NHS SCOTLAND

Surveillance & outbreak investigation – the laboratory perspective

John E Coia
Derek Brown
Scottish Microbiology Reference Laboratories, Stobhill Hospital, Glasgow

NHS SCOTLAND

What is required?

- Outbreak Detection & Trend Analysis
  - Same methods may be used
  - Different populations sampled
  - Sensitivity depends on how complete and/or representative the sample is
- Outbreak Investigation
  - Technique requires sufficient resolution to discriminate outbreak isolates from non-outbreak isolates

2

NHS SCOTLAND

Sensitivity of Outbreak Detection – the routine dimension

- **Clinical Disease**
  - Are symptoms present?
  - Is a sample sent?
- **Pathogen Detected**
  - Does the front-line laboratory look for the pathogen?
  - How good is the test?
- **Pathogen characterised**
  - How much does the front-line laboratory do?
  - Do they send to a reference laboratory for further ID & typing?

3

NHS SCOTLAND

Does the routine lab test?

Salmonella	Campylobacter	VTEC
Yes	Yes	Only test for O157 routinely
Isolate referred	Not routinely referred	Isolate referred

4

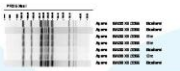
What level of discrimination do you need?

- Depends on the question you want to ask
  - How many salmonella infections are occuring?
    - Salmonella spp. is enough
  - What are the top 10 serotypes of salmonella in the UK?
    - Serotype is enough
  - Are any outbreaks of salmonella occurring?
    - Is it an uncommon serotype?
      - Yes. Serotype may be enough
      - No. Need further subtyping

5

Use of PFGE in multinational outbreak of Sal. Agona infection

- Feb to Aug 2008 increase in cases of Agona infection noted in Scotland, England, Wales, N.Ireland and Eire
- Phage types reported by HPA as PT39 (phage typing not widely available for Agona)
- Electronic exchange of PFGE patterns confirmed indistinguishable clone (SAGOXB.0066)
- Scotland – 34 out of 52 isolates of Agona confirmed as outbreak type by PFGE
- PFGE essential in "case definition" for epidemiological investigation
- Total of 163 cases in Europe linked to distribution of cooked meat from a plant in Eire widely distributed through "sandwich" shops.



6

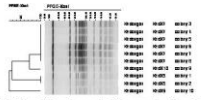
Recent experience - STEC O104 in Germany

- Front line lab test
  - Disease severity
  - Toxin detection
- Outbreak Detection
  - Serogroup sufficient. If you had O104 in your stools in Germany in May, you were part of the outbreak
- Outbreak Investigation
  - Sequencing studies told us a great deal about how this organism evolved BUT
  - Classical epidemiology identified the source

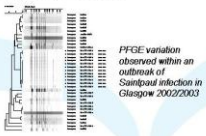
7

How different is different? How different is the same?

- Sufficient discrimination to distinguish outbreak and non-outbreak
- Sufficient similarity to group the outbreak strains together
- How much change can we expect within the time window of our outbreak?



PFGE variation observed within one subculture of a single colony of S. Kedougou



PFGE variation observed within an outbreak of Saintpaul infection in Glasgow 2002/2003

8

Molecular pros

- Level of expertise to perform
- Automation
- No need to transport isolates
  - Electronic interchange
  - DNA transport
- Sequence-based methods should be more objective and easier to standardise
- Speed
- Cost
  - Economies of scale?

9

Molecular issues 1

- We have witnessed a convergence in methodology, but not always in interpretation
- Need to know the background
  - Need to know the genetic plasticity of the organism in question
  - Not enough to look at outbreak isolates
    - How common is that pattern/profile/sequence variant?

10

Molecular issues 2

- Correlation of subtypes/groupings with existing epidemiological associations and/or clinical presentations
- Still need to standardise for comparability (even with sequence-based data)
  - Is data from Lab A/Analyser X correlate with Lab B/Analyser Y
- Bioinformatics issues
  - Genetic methods have developed faster than the tools to analyse the data

11

Molecular Issues 3

- How many centres do we need?
  - Examples of existing networks
- How do we co-ordinate internationally?
  - Existing networks PulseNet, ECDC-Enternet
  - Maintaining international comparability
    - Current issues around different methodologies
  - Need to compare and exchange data in real-time is crucial

12

Breakout Group 1 – how can we deploy next generation sequencing in foodborne *disease relevant surveillance and* outbreak investigation to best effect?

- **Priority organisms?**
  - Campylobacter
  - Salmonella
  - VTEC
  - Listeria

Breakout Group 1 – how can we deploy next generation sequencing in foodborne outbreak investigation to best effect?

- **Potentially useful collections**
  - Currently available? (but ?Accessible)

  Human: Ref lab archives, IID2 samples, ad hoc collections (Oxfordshire 3.5K, Scots 6K campylobacter)

  Animal: Ref labs archives, SAC (e.g campylobacter, VTEC), AHVLA (for animal pathogens – not VTEC!) VTEC

  Food: FSA, industry, HPA

  Environment: CEH, ad hoc

  - Desirable in the future?

  Structured, available, accessible, documented.

  International

Breakout Group 1 – how can we deploy next generation sequencing in foodborne outbreak investigation to best effect?

- **Type and quality of samples?**
  - **Representative, structured, problem specific**
  - **(Specimen) Physical state, not a big issue**

Breakout Group 1 – how can we deploy next generation sequencing in foodborne outbreak investigation to best effect?

- **Type and quality of data?**
  - **Linked epi (PP&T), zoo/enviro, and micro**
  - **Linked phenotypic and genotypic, specimen and other data**
  - **Consistent, repeatable, "confidence"**
  - **Accessible, interrogatable**
  - **Denominators**

Breakout Group 1 – how can we deploy next generation sequencing in foodborne outbreak investigation to best effect?

- **What should we do with the data?**
  - **Collect, analyse, and interpret it in a practically useful way**
  - **Share the consequent information (and raw data) widely and appropriately**

Breakout Group 1 – how can we deploy next generation sequencing in foodborne outbreak investigation to best effect?

- **How should we make the strains and data available?**
  - **Widely!**

Breakout Group 1 – how can we deploy next generation sequencing in foodborne outbreak investigation to best effect?

- **How many databases would be needed?**
  - **Few. Actual number doesn't matter as long as structures are compatible, quality is assured, linkage is preserved – and access is easy and free.**

Breakout Group 1 – how can we deploy next generation sequencing in foodborne outbreak investigation to best effect?

- **What else?**
  - **Nomenclature – simple, and common to all**
  - **"backwards comparability" from WGS to currently used nomenclature/systems (PFGE) – nice for context, but not essential.**
  - **Speed and thresholds for action**
  - **Consistency of approach locally, nationally, and internationally (build up from the bottom – don't wait for laggards) – staged approach**
  - **Link micro methods to data collection, analysis, interpretation, and information sharing (Surveillance!)**
  - **Link micro results to epi/zoo/food/env data**

Breakout Group 2 –what is the current availability and capacity of relevant facilities?

- **Are there enough trained people to generate/analyse data?**
- Generate – yes if we build in whole genome sequencing in routine clinical analysis
- Analyse- no and they are needed
  - But we should not delay role out to wait for them and train them in parallel

Breakout Group 2 –what is the current availability and capacity of relevant facilities?

- **Do facilities have appropriate accreditation and quality standards?**
- Not the issue – cost of machines will change the capability of the community rapidly in the next couple of years, we need to think about what the world will look like
- In next couple of years – main clinical/research labs will have capability
- 5 years: all reference labs will have capability
- 10 years: diagnostics will involve full genome analysis
- Did not discuss quality, but meta data standards important

Breakout Group 2 –what is the current availability and capacity of relevant facilities?

- **What are the likely costs?**
- Not expensive,
- Even now cheaper to use that older techniques in some high throughput labs
- As cost is reduced it is more cost effective for more labs to role it out
- It will cost us more to delay…..

Breakout Group 2 –what is the current availability and capacity of relevant facilities?

- **Given current resources how would we use such techniques in foodborne disease outbreak investigation ?**
- All outbreaks from today onwards should involve full genome sequencing and sequencing of relevant historical collections to inform decision making

Breakout Group 2 –what is the current availability and capacity of relevant facilities?

- **Anything else?**
- **We just need to get on and do it**
  - **We don't need to delay implementation of this as a standard technique while some of the issues are ironed out**
- **Transition will be in phases**
  - **Now- high throughput labs and universities**
  - **Soon all reference labs**
- **It will not be cost effective for every lab to have on now,**
  - **especially if not high throughput**
- **Metadata and data sharing important**
  - **But they are existing issues, and need to be sorted in parallel**
- **A sequence is a sequence – data sharing will be easier**
- **We don't need to sequence all historical data straight away**
  - **Reference samples will be self populating**

Molecular Epidemiology Workshop
17 January 2012

Participants' pre-workshop feedback

The following feedback has been provided by participants for consideration before the workshop.

Dr Jørgen Schlundt - Danmarks Tekniske Universitet

I would like to inform the participants about an important meeting in Bruxelles 1-2 Sep 2011 on this topic – I also attach a Science News story on the subject[*1]

The Bruxelles meeting will be followed by a 2nd meeting in Washington – presently planned for 1-2 March 2012

[*1]Please note that a consensues report of the meeting in Brussels and the Science news story are attached separately.

Dr Vanya Gant - UCLH

Current and recent activities and interests

a) First *Lancet* publication on viable and robust front line array-based diagnostics for human bacteremia 2010
b) Co-PI on second stage FP7 application for €6 million: rapid multiplex detection of Respiratory Tract Infection
c) Co-PI on second stage FP7 application for €6 million: potential for NGS for routine clinical laboratory implementation
d) Chaired and spoke at several International Meetings addressing microfluidics/NGS technology implementation in clinical medicine
e) Assessor for open TSB call for the Detection and Identification of Infectious Agents (DIIA)
f) Particularly interested in the societal barriers to implementation of new diagnostic technologies

---

Prof Martin Maiden – University of Oxford

We are currently working on a number of project areas relevant to this discussion:

(i) We are part of the international Patho-NGen-Trace consortium (led by Stephan Niemann and Dag Harmsen) funded by the EU for four and a half years to develop the used of Next Gen sequencing in clinical microbiology (this includes *Campylobacter*, *Mycobacterium tuberculosis* and Staphylococcus aureus as exemplar organisms). A particular focus of this application is the involvement of industry, particularly SMEs;
(ii) Funded by DEFRA and the FSA, we are undertaking ongoing surveillance of all *Campylobacter* isolates from Oxfordshire (With Kate Dingle at the John Radcliffe Hospital and Noel McCarthy of the HPA). This continues surveillance since 2003 which had been done since 2003 (more recently in near real time) and in collaboration with the Sanger Institute (Stephen Bentley and Julian Parkhill) this is now being done by whole genome sequencing rather than PCR-based sequencing of

individual loci (conventional MLST). The aim is to make the Whole genome sequence data available as assembled sequences in near real time.

(iii) We are undertaking a similar approach to all Meningococcal isolates from the epidemiological year 2010-2011 funded by the Meningitis Research Foundation and in collaboration with the HPA (Ray Borrow) and the Sanger Institute (Juilan Parkhill). This also aims to deposit data in a useable format on the web as it is generated (i.e. as assembled annotated data). Whilst this is not directly relevant to food borne outbreaks the techniques being used are generic.

(iv) Funded by the Wellcome Trust and DEFRA We have for many years operated the PubMLST.org website for the molecular epidemiology of organisms including *Campylobacter*. This is now running our recently developed Bacterial Isolate Genome Sequence database (BIGSdb) database and analysis platform (Jolley & Maiden, 2010, BMC Bioinformatics 11:595) which is fully capable of serving whole genome sequence data. This platform is being used to publish the data generated by the three projects outlined above.

(v) Since 2005 together with Stephen Gillespie (now at St Andrews) and Cath Arnold (HPA) we have been running the Wellcome Trust Advance course in Genomics and Clinical Microbiology, which aims to train Clinical Microbiologists in the application of sequencing technologies.

To comment on some of your specific points:

Breakout group 1
In general terms for *Campylobacter*, surveillance at the detailed genetic level is more important for controlling disease burden generally; the number of point source outbreaks is small. However in the former role sequence based typing methods are essential.
Routine accurate collection of isolates is essential, but is at risk from the lack of local incentives to store isolates locally long term and the need for nationally collected isolates and data from them to be made freely available not held as private collections by those with a responsibility to collect them.

For WGS with current technology good quality DNA (i.e. extracted from isolates) is currently required, but this is an ever moving field and this requirement is likely to diminish (if it hasn't already). Collection of isolates should continue to be an aspiration, however, Data and strains should be made available via the internet using suitable databases (such as BIGSdb) as both assembled annotated or partially-annotated sequences and from the short read archive at the EBI (although this is unusable by most epidemiologists and microbiologists).
The public MLST databases are a model for this that provides an effective and efficient means of achieving data distribution (see www.pubmlst.org/campylobacter and www.pubmlst.org/neisseria).

Breakout group 2
Training is a vital step, but a crucial issues is who is trained and in what. Currently there is a disconnect between clinicians and epidemiologists (who understand the clinical and epidemiological context) and statistical geneticists (who are interested in the analysis of the data). It is important that the data are generated and presented in an intelligible and useful way to the users. This does not mean the production on the fly of complex 'SNP' based trees and proper nomenclature schemes are essential. Appropriate quality standards do need to be developed and implemented. In the medium to longer term it is unlikely that cost will be a major issue – it is already cheaper to generate data from more than seven genetic loci using WGS data. Together with our collaborators, we are currently implementing a model (see (i) – (iii) above) that can achieve these aims.

Dr Carole Foy - LGC

One initiative that I would like to share with the other participants relates to a European project that has just been funded and is being led by LGC. One of the main aims of the project is to improve confidence in data from emerging genomics approaches such as next generation sequencing. A short summary of the project is given below:

"The European Metrology Research Programme (EMRP) is a metrology-focused European programme of coordinated R&D (http://www.emrponline.eu/). The EMRP has recently funded a new project (INFECT-MET) which aims to develop novel measurement procedures and validation frameworks to support current and emerging molecular approaches for efficient, harmonized and rapid diagnosis, surveillance and monitoring of infectious diseases. The project's ultimate aim is to establish routes for improving the accuracy, robustness, comparability and traceability of measurements across Europe linked in to international standardisation initiatives in the area through. This project is being led by LGC and includes metrology partners from across Europe as well as numerous collaborators representing 1) the diagnostics/instrument developers, 2) the microbial/clinical/epidemiology communities, 3) the QA and standards communities.
One of INFECT-MET's objectives is to quantitatively and comparatively evaluate new and emerging molecular approaches for surveillance and epidemiological monitoring. Multiparametric, high-through 'omics approaches such as next generation sequencing and high-throughput qPCR will be considered. Another objective is to evaluate new and emerging diagnostic technologies for the rapid (near-patient/on-site) detection of infectious agents."

Therefore, my main interest will lie in breakout group one. As well as sample quality and data availability I would like to explore how we can ensure comparability of data and incorporate (or develop if they are not already available) appropriate reference standards and quality control procedures in the process. Method validation approaches for demonstrating "fitness for purpose" and defining performance criteria of emerging genomic approaches are also of interest to me.

Dr Norval Strachan and Dr Ken Forbes – University of Aberdeen

FSA Workshop on Molecular Epidemiology

Some general points

Sporadic and outbreak cases are two ends of a spectrum where the middle ground may be artifactual due to an absence of knowledge linking together apparently unconnected cases. For this reason it is important to both maintain a level of sampling/ surveying of pathogens of concern and also to identify epidemiological links using state-of-the-art molecular and epidemiological tools. This information will not only clarify sources and routes of infection (informing outbreak investigations) but also give a better understanding of the pathogen's population structure and thus aid in identifying new variants and trends.

Outbreak investigations are typically carried out by a combination of those in Public Health questioning the infected and identifying candidate sources and medical microbiologists in the local lab and subsequently at appropriate Reference Laboratories. This combination works well most of the time, however, expertise outwith the NHS – Universities, SAC, defra - can all offer significant added value for two reasons. Firstly, they facilitate the adoption in the reference laboratories of new techniques and analysis methods that have been developed elsewhere. Secondly, there can be significant gains in the analysis and understanding of the microbiology, pathology and epidemiology of these infectious organisms. Pragmatically, the staff in many of

the reference laboratories do not have the time available to undertake such work. This can best be achieved through external collaboration. There needs to be a culture of cross-fertilization of ideas and information. Significant advances in our understanding of these infectious diseases can be made if there is a recognition that collaboration can be to the advantage of all. This is perhaps most true where the 'bigger picture' is concerned as focus tends to be on individual cases or specific outbreaks. In particular, better links between the microbiologists who are typing the organisms and the epidemiologists with their patient-oriented approach has much potential to increase knowledge.

National, and international, boundaries also need to be considered since both the infected patient and contaminated foodstuffs travel widely nowadays. Outbreak investigations need to be trans-frontier and involve ECDC.

Data pertaining to an isolate has to be standardised across all institutions: molecular typing methods and quality of the data; associated information (source, patient).

Where a new typing technique (NGS) is to be adopted there must be linkage to the molecular epidemiology of the old method and this is best achieved using a strain collection which is representative of the epidemiology characterised to date. This must include isolates from both clinical and food /environmental sources.

Some Specific points

Breakout Group 1

Campylobacter:
- what collections are available that could potential be used?

University of Aberdeen holds several collections of recently isolated *Campylobacter* which have all been typed by MLST:

Clinical isolate datasets.

Period	Region	Total
2005 -06	Scotland	5674
2005 -07	Grampian	1452
2010 -11 (& ongoing)	Grampian	697

Host datasets.

Host		Total
Cattle	2010 -11 & 2005 -06 (& ongoing)	438
Sheep	2010 -11 & 2005 -06 (& ongoing)	247
Chicken	2010 -11 & 2005 -06 (& ongoing)	483
Wild Birds	2005 -06	188
Pigs	2005 -06	40

VTEC:
- what collections are available that could potential be used?

University of Aberdeen holds several collections of recently isolated *E. coli* O157 which have/ are being typed by MLVA:

Clinical isolate datasets.

Period	Region	Total
to 2007	Grampian	100
2009 -11 (& ongoing)	Grampian	69

Host datasets.

Host		Total
Cattle	to 2011 (& ongoing)	130
Sheep	to 2011 (& ongoing)	70

How can we deploy next generation sequencing in foodborne outbreak investigation to best effect?

Need to use molecular attribution models to identify sources of clinical isolates. For this to be operational there is a requirement to develop a database of WGS information from isolates obtained from a representative range of sources. These could be improved beyond simple host identification to encompass the identification of producer or whether all strains are equally likely to pass all the way through the food chain.

Which may inter alia address the following -
- which microorganisms should be prioritised? (Campylobacter, VTEC, Salmonella, L. monocytogenes)

Those that will achieve maximum pay-off in terms of reducing the incidence of human disease for minimum cost. All 4 pathogens are important & FSA has prioritised Campylobacter and Listeria. Molecular sequence based typing data (MLST) are probably most comprehensive for Campylobacter and will require least effort to attain maximum benefit. However, outbreaks are rare for Campylobacter apart from those involving chicken liver. Listeria is relatively rare human disease but with high morbidity/mortality. There will not be many clinical isolates to WGS but there will be a need to sequence large numbers of potential source isolates which are being routinely obtained by industry (but not typed). VTEC – the numbers of clinical isolate are not high but there would be a requirement to isolate from animals (is there any ongoing surveillance) and further what range of organisms should this cover? Certainly O157 but which other serotypes should be included?

It is not always the case that a source or case harbours a single strain of the pathogen. For example, *Campylobacter* cases linked to contaminated chicken liver pate are usually associated with multiple strains. Good microbiological practice is usually to test single colonies and further if different cases have infection caused by different strains in the source, then typing of the isolates in source and in cases may not identify the commonality between them.

- what collections are available that could potentially be used?

Isolates collected by industry during routine monitoring. Very few are typed. There are significant numbers of Listeria and probably also Salmonella and Campylobacter isolated by industry but not VTEC.

- how should we make the strains and data available?

On-line – there will need to be sufficient anonymity for clinical cases (higher level access that enables public health to access these data?)

- how many databases would be needed?

One database that is structured by pathogen. This could be akin to pubMLST but needs to be developed for easy interpretation for outbreak investigation purposes. Would make sense for these databases to be international.

Breakout group 2

What is the current availability and capacity of relevant facilities?

Which may inter alia address the following

- are there enough appropriately trained staff available to analyse as well as generate the data?

Need development of software that makes analysis automatic.

- given current available resources how would we do this?

What are the currently available resources?

Dr Lisa Crossman - TGAC

Currently we have a preliminary investigation into *Listeria* (25 strains) here at TGAC joint with the HPA Colindale. We are definitely intending to scale this investigation up in terms of a deep resequencing project, funding permitted.

Group I

- I would like to suggest *E.coli?* (but this could potentially be split into EAEC, EHEC, ETEC and so on.) and *Clostridium botulinum*
- The IFR hold a *botulinum* strain collection.
- Would like to use Illumina sequencing due to accuracy and price
- Publication in appropriate journals
- We need to formulate appropriate databases to make the data available. There may be conflicting demands between the needs of health professionals and academic researchers. Personally I might prefer a one organism one database basis, there may be some argument for standardising the underlying software across all the organisms.

Group II

TGAC is still growing in terms of capacity and staff

Suggest D. Kell is appropriate person with TGAC director to advise on how we could do this with currently available resources, capacity and availability of facilities.

Prof Mark Blaxter – University of Edinburgh

Firstly, Chris Low and David Gally hosted a FSAS-sponsored workshop on E. coli O157 in November 2011, and that had a somewhat overlapping agenda. The outputs of that meeting might be relevant to this one - do you have access to its report?[*2]

On topic 1:

A “strategic” comment is that the usual practice of collecting a whole lot of samples and frantically sequencing them after an outbreak has started inevitably leaves us doing “catch-up” science. What we need is a much better understanding of the genetic diversity of the strains circulating at a given time and location, which means systematic, structured and sustained surveillance on an appropriate scale. Not sexy, but still very important.

For many outbreaks the source is "environmental" and thus the reservoirs are large and largely unexplored. Surveillance is the key.

On Topic 2:

There are facilities, and we are ready to do the sequencing. With new technologies such as the MiSeq and IonTorrent, this is a rapid and 'real time informative' process.

[*2] It is anticipated that the report will be published at the end of March 2012.