‘The transformation of viral epidemiology and clinical studies by next-generation sequencing’
Utility of virus genomes

• Sequencing virus genomes
  – The individual
  – National epidemic tracking
  – Zoonotic chatter
  – The beginning of an outbreak

• Endemic disease

Holmes and Grenfell PLoS Comp Biol
Genome sequencing 2016

1. Samples & phenotype

2. DNA/RNA

3. Genome Amplification or enrichment

4. MinIon
   - MiSeq
   - Ion Torrent

5. Assembly + QA/QC

6. Analysis + data integrations
• Zoonotic chatter – Rotaviruses in Vietnam
• The beginnings of human transmission – MERS CoV
• Major outbreaks – Ebola virus
• Infection control in the UK – Influenza virus
Sampling frameworks - zoonosis
The work of My Phan and Matt Cotten in collaboration with Stephen Baker

High Risk Cohorts
Disease surveillance

Pilot work of Wellcome Trust VIZIONS
Agnostic Virus Genome Detection

Cell Free Fluid/Infected cells

DNAse/RNase digestion of high speed pellet

Nucleic acid prep
Library prep
Next Gen Sequence

Readset

Read1
Use BLASTN to identify read.

Retrieve full sequence from GenBank

Map all other reads aligning to the full sequence

Specific virus reads
De novo assembly

Remaining reads Subjected to next round

Bin reads according to identity
Virus Non-Virus Mystery
ViSeq method Sensitivity - Norovirus

The graph shows the relationship between norovirus reads by ViSeq and Ct values. The x-axis represents Ct values, while the y-axis shows the number of norovirus reads. The data points are plotted on a logarithmic scale, indicating a linear relationship between the two variables. The trend line suggests a decrease in norovirus reads as Ct values increase, which is typical for qPCR assays.
Enteric virus content of human, porcine fecal samples

Criteria to be a virus:
A contiguous assembled sequence of length x (i.e. <1000bp)
A minimum read depth per base for the contig. (i.e. 100)
**Reoviridae – Rotavirus**

For Group A rotaviruses

Major classification based on G (VP7, seg 9) and P (VP4, seg 4) - targets of neutralizing antibodies

**Extended classification**

Gx-P[x]-Ix-Rx-Cx-Mx-Ax-Nx-Tx-Ex-Hx

27 37 17 9 9 8 18 10 12 15 11

**Major Human Group A**

G1P[8], G2P[4], G3P[8], G4P[8], G9P[8]

on a genotype constellation

I1-R1-C1-M1-A1-N1-T1-E1-H1

or

I2-R2-C2-M2-A2-N2-T2-E2-H2
## Rotavirus genome segment reassortment

<table>
<thead>
<tr>
<th>Genotype constellation</th>
<th>Host count</th>
<th>Human</th>
<th>Pig</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1 P[8] I1 R1 C1 M1 A1 N1 T1 E1 H1</td>
<td>33</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G2 P[4] I2 R2 C2 M2 A2 N2 T2 E2 H2</td>
<td>12</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G3 P[8] I1 R1 C2 M1 A1 N1 T1 E1 H1</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G4 P[6] I1 R1 C1 M1 A8 N1 T7 H1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>G4 P[6] I1 R1 C1 M1 A8 N1 T7 H1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>G5 P[13] I5 R1 C1 M1 A8 N1 T7 x H1</td>
<td>0</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>G5 P[13] I5 R1 C1 M1 A8 N1 T7 x H1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>G6 P[13] I5 R1 C1 M1 A8 N1 T7 E1 H1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>G9 P[23] I5 R1 C1 M1 A8 N1 T7 E1 H1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>G11 P[13] I5 R1 C1 M1 A8 N1 T7 E1 H1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>G1 P[8] I2 R2 C2 M2 A2 N2 T2 E2 H2</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G2 P[8] I2 R2 C2 M2 A2 N2 T2 E2 H2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G2 P[8] I2 R2 C2 M2 A2 N2 T2 E2 H2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G1/G2 P[8][4] I1/I2 R1/C2 M1/M2 A1/A2 N1/N2 T1/T2 E1/E2 H1/H2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G1/G2 P[8][4] I1/I2 R1/R2 C1/M1 A1/A2 N1/N2 T1/T2 E1/E2 H1/H2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G1/G2 P[8][4] I1/I2 R1/R2 C1/M1 A1/A2 N1/N2 T1/T2 E2 H2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G1/G2 P[8][4] I1/I2 R1/R2 C1/M1 A1/A2 N1/N2 T1/T2 E2 H2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G1/G2 P[8][4] I1/I2 R1/R2 C1/M1 A1/A2 N1/N2 T1/T2 E1/E2 H1/H2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G1/G2 P[8][4] I1/I2 R1/R2 C1/M1 A1/A2 N1/N2 T1/T2 E1/E2 H1/H2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G1 P[6] I1 R1 C1 M1 A1 N1 N1 T1 E1 H1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G1/G4 P[8] I1 R1 C1 M1 A1/N1 T1/T1 E1/E1 H1/H1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>G6/G11 P[13][23] I5/I5 R1/R1 C1 M1 A8/A8 N1/N1 T1/T7 E1/E1 H1/H1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

### Color codes
- DS-1-like genotype (G2/G8-P[4]-I2-R2-C2-M2-A2-N2-T2-E2-H2)
- G4-P[6]
- A8-T7
- G9-P[23]
- G5
- G11
- I5
- P[13]
- x Sequence not determined
Figure 3A. Maximum-likelihood tree of RVA VP7
Virus genomes and zoonotic chatter

- Human and animal contact areas
  - Random sequencing
  - Risk groups
  - Random sampling of human clinical samples
  - Random sampling of veterinary samples

- Linked to serology surveys and risk maps
  - Virus genetic diversity and genome movement risk maps

- Requires simple sample preparation, sequencing and assembly methods (commercial?)
## 12 years of Virus Outbreaks

<table>
<thead>
<tr>
<th>Year</th>
<th>Virus Type</th>
<th>Cases</th>
<th>Deaths</th>
<th>CFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002-3</td>
<td>SARS Coronavirus</td>
<td>8273</td>
<td>775</td>
<td>~10%</td>
</tr>
<tr>
<td>2003-4</td>
<td>H5N1 Influenza A</td>
<td>630</td>
<td>375</td>
<td>~60%</td>
</tr>
<tr>
<td>2009-10+</td>
<td>H1N1 Influenza A</td>
<td>global</td>
<td>~579,000</td>
<td>~0.01%</td>
</tr>
<tr>
<td>2013</td>
<td>H7N9 Influenza A</td>
<td>134</td>
<td>44</td>
<td>~33%</td>
</tr>
<tr>
<td>2012-5</td>
<td>MERS Coronavirus</td>
<td>1266</td>
<td>388</td>
<td>~30%</td>
</tr>
<tr>
<td>2014-5</td>
<td>Ebola virus</td>
<td>28,103</td>
<td>11,290</td>
<td>~40%</td>
</tr>
</tbody>
</table>

In an outbreak there should be a commitment to turn a portion of RESIDUAL diagnostic nucleic acid into a publically available pathogen genomes at NO additional cost to the country(s) experiencing the outbreak & that the data leads to SHARED, ACTIONABLE & INTERPRETABLE INFORMATION

$Case fatality rate
Virus genomes & molecular epidemiology

Phylodynamics, a term coined to denote the interplay between evolution and epidemiology when occurring on the same timescale.
Clinically actionable/useful virus genome

Cases by date

MERS-CoV – 30Kb RNA genome

Fraction of Genome Obtained

Ct value

MERS-CoV – 30Kb RNA genome
Sharing data allows new insights

http://mers.nextflu.org

Cotten et al (2014), mBio 5(1), e01062-13
Shared data allows real-time updates
Reduction in cluster size as a proxy for control of infection or limit of an infection cycle?
Ebola virus genome sustained local sequencing

Move the data not the samples

Ian Goodfellow, Armando Arias, Jia (Luca) Lu, Lucy Thorne, Matt Cotten, My Vu Tra Phan, Simon Watson, Andrew Rambaut, Dhamari Naidoo

Yozwiak et al Nature 2015; 518(7540): 477-9
Minority variants and transmission chains
Real time virus genomes

Suspected Case

Case taken to nearest ETC & sample taken for diagnostics

Positive test

Epidemiology starts & data recording

Contact tracing

Residual Material Sequenced

Phylogenetics

WHO report

855 sequencing samples were processed between 16 April 2015 and 15 September 2015 yielding 614 EBOV genomes (72%).
Utility of virus genomes

Opportunities exist if virus genome sequences are routinely obtained from diagnostic samples during outbreaks or from routine diagnostic services.

In an integrated healthcare system pathogen genome sequence will allow evidence-based infection control at different health care levels, will inform national epidemiology and will allow stratified patient management for treatments.

SHARED, ACTIONABLE & INTERPRETABLE INFORMATION

35 recommendations

Infection response through virus genomics
Infection responsive through virus genomics

Samples/Requests

Diagnostics

20,000 HIV, HCV, IAV, NV, MV

Sequence

Patient ID clinical

Clinical reporting

Clinical Virology Workflow

Interpretation
How to share data

Figure 15.3 A simplified vision of a two-tier data sharing strategy

Data generators
- e.g. public health reference laboratories
- e.g. hospital linked microbiology / virology
- e.g. specialist clinics

Data storage / collation
- public domain e.g. ENA
- public health authority - restricted access

Data accessibility
- e.g. researchers, commercial, public
- e.g. authorised users (clinical or public health consultants)

The size of circles (not to scale) are indicative of the relative data storage burden (computational disc space), of the different subsets of data. Raw genomic data will consume the greatest disc space (therefore cost more to store than other data types), and so its longer term storage would be better suited in a consolidated repository build for high volume data storage.
## Mapping Genome sequencing to clinical data

<table>
<thead>
<tr>
<th></th>
<th>Influenza</th>
<th>Norovirus</th>
<th>HIV</th>
<th>HCV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total submitted lab numbers</strong></td>
<td>407</td>
<td>20</td>
<td>360</td>
<td>152</td>
</tr>
<tr>
<td><strong>Total records retrieved from Winpath.Results</strong></td>
<td>406 (99.8%)</td>
<td>20 (100%)</td>
<td>360 (100%)</td>
<td>150 (98.7%)</td>
</tr>
<tr>
<td><strong>Total identifiers found</strong></td>
<td>406 (99.8%)</td>
<td>20 (100%)</td>
<td>240 (66.7%)</td>
<td>148 (97.4%)</td>
</tr>
<tr>
<td><strong>Carecast data found</strong></td>
<td>310 (76.2%)</td>
<td>14 (70%)</td>
<td>76 (21.1%)</td>
<td>30 (19.7%)</td>
</tr>
<tr>
<td><strong>Obvious UCLH inpatient locations</strong></td>
<td>357 (87.7)</td>
<td>14 (70%)</td>
<td>28 (7.8%)</td>
<td>2 (1.3%)</td>
</tr>
<tr>
<td><strong>Patient stay records matches</strong></td>
<td>240 (59%)</td>
<td>11 (55%)</td>
<td>20 (5.6%)</td>
<td>2 (1.3%)</td>
</tr>
</tbody>
</table>
• UCLH outbreaks driven by multiple introductions from community

• Limited patient-patient chains – ~1 in 7 cases initiate a chain of hospital transmission
Early detection of vaccine mismatch
357 influenza genomes from UCLH & Barts Health over 3 consecutive influenza seasons

- January 2013
- January 2014
- January 2015

Graph showing the count of influenza genomes from January 2013 to January 2015, with categories for A/H1N1/09, A/H3N2, B/Vic, Mixed, and Unknown. The graph also indicates the vaccine strain mismatch for Winter 2013/4 & Winter 2014/5.

Legend:
- Red: A/H1N1/09
- Green: A/H3N2
- Blue: B/Vic
- Black: Mixed
- Grey: Unknown

The graph shows a decrease in certain strains over time, indicating potential vaccine mismatch.
Conclusions

• NGS for large scale and rapid virus genome sequencing is almost fit for purpose but need:
  – Commercial sample to multiplex library
  – Accurate minority variant detection required
  – Stable computational pipelines
  – Linking to meta data

• With appropriate sampling framework large scale sequencing can:
  – Characterise a zoonotic reservoir
  – Identify zoonotic virus ‘chatter’
  – Inform outbreak control in the field and in hospitals
Acknowledgements

Sanger Institute
My VT Phan
Simon Watson
Matthew Cotton

University of Edinburgh
Andrew Rambaut
Gytis Dudas

UCL
Zisis Kozlakidis

Cambridge University
Ian Goodfellow
Armando Arias,
Jia (Luca) Lu,
Lucy Thorne

Dhamari Naidoo
WHO Emerging and Dangerous Pathogens Laboratory Network.
WHO Sierra Leone and Laboratory and Technical Working Group from MoHS Centre for Disease Control (CDC), USA
Public Health England (PHE)

Kingdom of Saudi Arabia
Ministry of Health
Ziad Memish
ICONIC Team

WP Leads
Paul Kellam
Andrew Hayward
Eleni Nastouli
Steven Morris

University College London
Laura Shallcross
Dan Frampton
Zisis Kozlakidis
Ellen Fragaszy
Tiziano Gallo Cassarino
Fatima Wurie
Anil Gunesh

BartsHealth
Duncan Clark
Jonathan Hubb
Graham Foster
Anthony Oliver

UCLH
Bridget Ferns
Elizabeth Gyimah
Jade Raffle
Pietro G. Coen

WT Sanger Institute
Simon Watson
Nick Grayson
Spela Binter
Stephanie Edwards
Swee Hoe Ong

PHE
Richard Myers
David James Allen
Richard Harris
David Brown
Tamyo Mbisa

MRC-CTU
David Dunn

Royal Free
Susan Hopkins
Daniel Webster
Tanzina Haque
William Rosenberg

University of Nottingham
William Irving
Jonathan Ball

University of Edinburgh
Andrew Leigh-Brown
Gonzalo Yerba

HCVRUK
John McLauclhan