What is Viromics and why does it matter?

Oxford Viromics Overview
HiDi Day, 1 October 2018

Rory Bowden – WHG-OGC, OxSingleCell, Oxford Viromics
What is Viromics?

Viromics is just the study of viruses at genome and population scale. The study of viruses becomes vastly more tractable once we get complete genome information. Modern genomics techniques give us a chance to access whole virus genomes cheaply, completely and at scale. Host factors in virus infection only relevant once we fully define the virus. Viromics is intimately interconnected with Immunology. Viromics can be used in diagnosis, tracing, monitoring, treatment, epidemiology, pathogenesis, …
What is Oxford Viromics?

Oxford Viromics is...

• An Idea

I. Make available sequencing and analysis for human virus-containing samples for small and large studies.

II. Proof of concept of the value of virus detection and genomics in clinical contexts.

III. Enable focused funding applications.

IV. A pipeline for virus detection and genetic characterization to enable clinical trials that define the role of state-of-the-art sequencing technology in clinical virology.
What is Oxford Viromics?

Oxford Viromics is...

- An Idea
- Some People

Paul Kleereman (TGU/Medawar)
David Bonsall (Medawar/WHG/BDI)
Rory Bowden (WHG)
Tanya Golubchik (WHG/BDI)
Mariateresa de Cesare (WHG)
Azim Ansari (WHG/Medawar)
What is Oxford Viromics?

Oxford Viromics is...

• An Idea
• Some People
• Collaborations

Barnes – Medawar/StopHCV
Matthews – Medawar/BRC
Pybus – Zoology
Simmonds – Medawar/BRC
Crook – MMM/BRC
Fraser – BDI
Maiden – Zoology
Lythgoe - BDI
What is Oxford Viromics?

Oxford Viromics is...

• An Idea
• Some People
• Collaborations
• Some Funding

Oxford Wellcome Trust ISSF + WHG
Oct 2016 – Sept 2018
2 half-time posts for 2 years
What is Oxford Viromics?

Oxford Viromics is...

• An Idea
• Some People
• Collaborations
• Some Funding

**Some Technology**
- Peter Medawar Building:
  - extraction platform
- Oxford Genomics Centre:
  - liquid handling
  - quantitation and QC
  - sequencing: Illumina (x4), Nanopore
What is Oxford Viromics?

Oxford Viromics is…

• An Idea
• Some People
• Collaborations
• Some Funding
• Some Technology
• Some Lab Methods

Optimising available methods:
- Extraction platform
- RNA-seq kits
- Sequence enrichment kits
- RNA-and-DNA libraries
- Streamlining, miniaturization, automation
- Nanopore
What is Oxford Viromics?

Oxford Viromics is...

• An Idea
• Some People
• Collaborations
• Some Funding
• Some Technology
• Some Lab Methods
• **Some Analytical Methods**

An initial pipeline ...

- snork
- metagenomics pipeline(s)
- shiver and phyloscanner
- nanopore-specific tools
- probe-design tools
What is Oxford Viromics?

Oxford Viromics is...

• An Idea
• Some People
• Collaborations
• Some Funding
• Some Technology
• Some Lab Methods
• Some Analytical Methods
• A Symposium

Save the Date:

11 January 2019

James Martin School
Proof-of-concept studies

How we knew Viromics was tractable and worthwhile

- RNA-seq has advantages over PCR-seq
- Probe enrichment is robust and predictable
- Can design probe sets to any arbitrary group of pathogens: 10kb – 5Mb++
RNA-seq vs PCR

RNA-seq:
• is less sensitive than PCR
• doesn’t depend on matching primers
• can work for degraded samples
• needs fewer tubes
Sample

Plasma

CSF

... →

lysed sample

RNA + DNA →

magnetised silica beads

- Reverse Transcribe
- Second Strand
- Ligate adapters →

sample adapters

index

index

96 or 384 sets of index tags

1. Isolate Nucleic Acids
2. Make Library
3. Index + Amplify
4. Pool 96+

5. Hybridize
6. Pull Down + Wash
7. Re-Amplify

(10^3-10^6 fold enrichment)

magnetic streptavidin beads

120nt biotinylated probes

Tree of target sequences (database)

(a) selection of targets

(b) tiling

120nt probes, overlapping by 60nt on each strand

DNA or RNA

sparser designs ... cheaper
Enrichment Sequencing

- Like exome sequencing
- Biotinylated DNA or RNA baits
- Enrich pooled libraries
- Probe-based enrichment is robust and predictable, even for HCV
- Typically $10^3$ – $10^4$-fold enrichment
- Tolerant of substantial sequence divergence
- Can design probe sets to any arbitrary group of pathogens: 10kb–5Mb+
Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus

M Ariz Ansari1,2,3, Vincent Pedergnana1,2,3, Camilla L Cipriani1,2, Andrea Magri1, Annette Von Delft1, David Bondal1, Nimita Chatravedi1, Istvan Barth1, David Smith2, George Nicholson3, Gideon McVernon4, Amy Trebro2, Paolo Piazza1, Jacques Fellay1, Graham Cook2, Graham R Foster5, STOP HCV Consortium5, Emma Hudson5, John McL anchan5, Peter Simmonds6, Rory Bowden1, Paul Kleringer3, Eleanor Barnes3, Chris C A Spencer1

Outcomes of hepatitis C virus (HCV) infection and treatment depend on viral and host genetic factors. Here we use human genome-wide genotyping arrays and new whole-genome HCV viral sequencing technologies to perform a systematic genome-to-genome study of 342 individuals who were chronically infected with HCV, predominantly genotype 3. We show that both alleles of genes encoding human leukocyte antigen (HLA) molecules and genes encoding components of the interferon-induced innate immune system drive viral polymorphism. Additionally, we show that D3A genotypes determine HCV viral load through a mechanism dependent on a specific amino acid residue in the HCV NS5A protein. These findings highlight the interplay between the innate immune system and the viral genome in HCV control.

HCV infection presents a major health burden, with more than 185 million people being infected worldwide, which can lead to liver failure and hepatocellular cancer in infected individuals. Genetic variations in both the host and the virus are associated with important clinical outcomes. Genetic polymorphisms in the host, most notably in the interferon (IFN) lambda 3 (IFNL3) and IFN alpha, are associated with spontaneous clearance of the virus, response to treatment, viral load and progression of liver disease. Viral genotypes and distinct viral genetic motifs have been associated with the response to interferon-based therapies, whereas resistance-associated substitutions (RASs) have been identified for most of the new oral direct-acting antiviral (DAA) drugs. HCV can be divided into seven major genotypes, and most of the genetic data accumulated to date has focused on HCV genotype 1, with a lack of data for other genotypes. HCV genotype 1 comprises a large number of viruses found worldwide, and is associated with a higher failure rate to DAA therapy9,10.

Previous work, including candidate gene studies of the association between the human leukocyte antigen (HLA) type 1 proteins and the HCV genotype11, has shown that within-host virus diversity evolves in response to the host adaptive immune system. HLA molecules are expressed on most cell types, and they present viral peptides (epitopes) to cytotoxic T lymphocytes (CTLs), which kill infected cells. CTL-mediated killing of virus-infected cells drives the selection of viral polymorphisms (escape mutations) that can escape host restriction12,13. Understanding how host HLA molecules affect viral selection has important implications for the development of HCV-specific T cell vaccines that aim to prevent infection14,15. A comprehensive host genotype to viral genotype association scale will assess the relative contribution of host HLA molecules in driving changes in the HCV genome, and it may also identify other host genes that have a key role in shaping the HCV genome.

We generated data from a cohort of 401 HCV-infected patients from the ROSON16 clinical trial to systematically look for associations between host and virus genomes, exploring the effect that while the host genotype remains fixed the virus mutates, allowing it to evolve during infection. For this, we developed a targeted viral enrichment methodology17,18 to obtain whole HCV genomes, and we used high-throughput genotyping arrays in association with statistical imputation to obtain data for nucleotide polymorphisms across the human genome and the alleles of genes encoding HLA molecules (hereafter referred to as HLA genes). We provide evidence that polymorphisms relevant to the innate (IFN) and adaptive immune systems (HLA genes) are associated with HCV sequence polymorphisms. We show that an interaction between host 371A genotypes and an amino acid residue in the HCV NS5A protein determines HCV viral load, by assessing viral evolution in individuals with different IFN α


Figure 1 Human-to-HCV genome-wide association study in 542 patients. The lower arc shows the human autosomes from chromosomes 1 to 22, and the upper arc shows the HCV proteome core protein (C) to NS5B. The red line represents the most significant association (P < 0.2 x 10⁻¹¹). The four blue lines represent suggestive associations (P < 4 x 10⁻⁹). The thin gray lines represent associations with P < 10⁻³. The outer mini-panels represent, on the upper arc, the viral diversity as measured by Shannon entropy and, on the lower arc, the density of human SNPs in bins of 1 Mb, with higher values further away from the center for both the upper and lower arcs.

Received 9 August 2016; accepted 22 March 2017; published online 10 April 2017; doi:10.1038/ng.3899
Viromics as a complete solution for management of a chronic virus infection

**Measurement:**
- Viral load
- Genotype
- Transmission network
- Drug resistance levels

**Sequencing design:**
- Quantitative standards
- Unbiased probe capture
- Minimal PCR, Fragment size selection
- Quantitative sequencing, Optimisation for low viral loads

**Analysis design:**
- PCR duplicate removal
- Accurate mapping, Consensus calling
- Ancestral host-state reconstruction
- Haplotype calling, HIVdb

HIV: BDI – Christophe Fraser
HIV genotyping and phylogenetics in HPTN 071 (PopART)

HIGH-THROUGHPUT SEQUENCING TO ASSESS VIRAL LOAD, GENOTYPE, DRUG RESISTANCE AND TRANSMISSION

U.S. NATIONAL INSTITUTES OF HEALTH:

- National Institute of Allergy and Infectious Diseases
- National Institute of Mental Health
- National Institute on Drug Abuse
Nucleic Acids Extraction
EasyMag (Biomerieux)

384 samples / week
£30 / sample

Liquid Handling
Labcyte Echo 525
Automated
384 - well format

Library prep
SMARTer (TakaraBio) Adapter attachment
without ligation

Minimal PCR
+ Size selection of larger fragments

Enrichment
650 Custom DNA oligos to capture full HIV A-D diversity

Sequencing
Illumina MiSeq / HiSeq 2500 Rapid
300b / 250b paired reads
Sequencing:
- Quantification standards
- Contamination Controls
- Negative controls
- Double indexing
- Replicates

SHIVER:
- Genome assembly
- Read mapping
- Consensus calling
- Remove PCR duplicates

Phyloscanner:
- Data cleaning
- Detection of dual infection
- Transmission

DRM characterization:
- Stanford algorithm
- Minor variant calling
- Link to transmission
- Mutation epistasis (DRM linkage)
HIV Sequencing:
- Viral load
- Genome coverage
- Sensitivity

[Graphs and histograms illustrating viral load and genome recovery from sequencing yield.]

Sequencing depth

Genome position

Recovered genome (nt)

Standard viral load

Viral load from sequencing yield

Recovered genome (nt)

Standard viral load

Viral load from sequencing yield
A comprehensive genomics solution for HIV surveillance and clinical monitoring in a global health setting

HIV Sequencing:
- Clinical management
- Transmissions
- Dynamics of infection
- Drug resistance

https://www.biorxiv.org/content/early/2018/08/28/397083
Viromics with new methods
Not every virus is ssRNA or dsDNA

Hepatitis B virus has a circular, partially double-stranded, DNA, virion genome that is not covalently closed.

→ “completion/ligation” plus “rolling-circle” amplification with HBV primers

= concatemers (chains) of successive full-length genomes

→ HBV-enriched; suitable for Illumina/Nanopore sequencing, further enrichment.

Philippa Matthews  Paolo Piazza
Anna McNaughton  Anthony Brown
David Bonsall  Azim Ansari
Hannah Roberts  Rory Bowden
Mariateresa de Cesare  Eleanor Barnes
HBV Workflows

Nanopore sequencing read error rates remain a challenge for calling within-sample variants.
Viromics as a diagnostic tool

Virus-agnostic Library + Comprehensive Enrichment Panel

- RNA-and-DNA libraries
- Curated list of viruses and bacteria (full-length, partial, rMLST)
- SureSelect RNA baits (~5Mb)
- For unknown samples
- For any included pathogen (e.g. EBV, HCMV, VSV)

Collaboration:
- GAinS
  - Cyndi Goh
  - Julian Knight
  - Eduardo Svoren
  - Charles Hinds
- ChiMES
  - Tanya Golubchik
  - Ivo Elliott
  - Andrew Pollard
  - Manish Sadarangani
  - Martin Maiden Group
  - Ellie Barnes
  - Rory Bowden

Both/Other
- Azim Ansari
- Mariateresa de Cesare
- Hubert Slawinski
- David Bonsall
- Amy Trebes
- Paolo Piazza
- Anthony Brown
- Senthil Chinnakannan
- Camilla Ip
- Martin Maiden Group
- Ellie Barnes
- Rory Bowden
<table>
<thead>
<tr>
<th>Family</th>
<th>Species</th>
<th>Adenoviridae</th>
<th>Herpesviridae</th>
<th>Polymaviridae</th>
</tr>
</thead>
<tbody>
<tr>
<td>Streptococcaceae</td>
<td>Streptococcus pneumoniae</td>
<td>Mastadenovirus A</td>
<td>HHV1 / Herpes Simplex Virus Type 1 (HSV-1)</td>
<td>mumps virus - G</td>
</tr>
<tr>
<td></td>
<td>Streptococcus pyogenes</td>
<td>Mastadenovirus B</td>
<td>HHV2 / Herpes Simplex Virus Type 2 (HSV-2)</td>
<td>mumps virus - H</td>
</tr>
<tr>
<td></td>
<td>Streptococcus agalactiae</td>
<td>Mastadenovirus C</td>
<td>HHV3 / Varicella-Zoster Virus (VZV)</td>
<td>mumps virus - I</td>
</tr>
<tr>
<td>Staphylococcaceae</td>
<td>Staphylococcus aureus</td>
<td>Mastadenovirus D</td>
<td>HHV4 / Epstein-Barr Virus (EBV)</td>
<td>mumps virus - J</td>
</tr>
<tr>
<td>Mycoplasmataceae</td>
<td>Mycoplasma pneumoniae</td>
<td>Mastadenovirus E</td>
<td>HHV5 / Human Cytomegalovirus (HCMV)</td>
<td>mumps virus - K</td>
</tr>
<tr>
<td>Legionellaceae</td>
<td>Legionella pneumophila</td>
<td>Mastadenovirus F</td>
<td>HHV6A / Human Herpesvirus 6A</td>
<td>mumps virus - L</td>
</tr>
<tr>
<td>Coxaciaceae</td>
<td>Coxiala burnetii</td>
<td>Mastadenovirus G</td>
<td>HHV6B / Human Herpesvirus 6B</td>
<td>mumps virus - N</td>
</tr>
<tr>
<td>Enterobacteriaceae</td>
<td>Escherichia coli</td>
<td>Lassa mammarenavirus</td>
<td>HHV7 / Human Herpesvirus 7</td>
<td>mumps virus - A</td>
</tr>
<tr>
<td></td>
<td>Klebsiella pneumoniae</td>
<td>Orthomylavirus</td>
<td>HHV8 / Kaposi's Sarcoma Herpesvirus (KSHV)</td>
<td>mumps virus - B1</td>
</tr>
<tr>
<td></td>
<td>Klebsiella oxytoca</td>
<td></td>
<td></td>
<td>mumps virus - B2</td>
</tr>
<tr>
<td></td>
<td>Enterobacter cloaceae</td>
<td>Bunyaviridae</td>
<td></td>
<td>mumps virus - B3</td>
</tr>
<tr>
<td></td>
<td>Enterobacter aerogenes</td>
<td></td>
<td></td>
<td>mumps virus - C1</td>
</tr>
<tr>
<td></td>
<td>Serratia marcescens</td>
<td></td>
<td></td>
<td>mumps virus - C2</td>
</tr>
<tr>
<td>Pasteurellaceae</td>
<td>Haemophilus influenzae</td>
<td></td>
<td></td>
<td>mumps virus - C2</td>
</tr>
<tr>
<td></td>
<td>Haemophilus parasuisne</td>
<td></td>
<td></td>
<td>mumps virus - C2</td>
</tr>
<tr>
<td>Chlamydiaceae</td>
<td>Chlamydia pneumoniae</td>
<td></td>
<td></td>
<td>mumps virus - C2</td>
</tr>
<tr>
<td></td>
<td>Chlamydia psittici</td>
<td></td>
<td></td>
<td>mumps virus - C2</td>
</tr>
<tr>
<td>Pseudomonadaceae</td>
<td>Pseudomonas aeruginosa</td>
<td></td>
<td></td>
<td>mumps virus - C2</td>
</tr>
<tr>
<td>Moraxellaceae</td>
<td>Moraxella catarrhalis</td>
<td></td>
<td></td>
<td>mumps virus - C2</td>
</tr>
<tr>
<td>Moraxellaceae</td>
<td>Acinetobacter baumannii</td>
<td></td>
<td></td>
<td>mumps virus - C2</td>
</tr>
<tr>
<td>Mycobacteriaceae</td>
<td>Mycobacterium tuberculosis</td>
<td></td>
<td></td>
<td>mumps virus - C2</td>
</tr>
<tr>
<td></td>
<td>Stenotrophomonas maltophilia</td>
<td></td>
<td></td>
<td>mumps virus - C2</td>
</tr>
<tr>
<td>Alcaligenaceae</td>
<td>Bordetella pertussis</td>
<td>Dengue Fever Virus Genotype 1</td>
<td>Human Parainfluenza virus 1</td>
<td>Respiratory syncytial virus - A JC polymavirus</td>
</tr>
<tr>
<td>Neisseriaceae</td>
<td>Neisseria meningitidis</td>
<td>Dengue Fever Virus Genotype 2</td>
<td>Human Parainfluenza virus 3</td>
<td>Respiratory syncytial virus - B BK polymavirus</td>
</tr>
<tr>
<td>Listeriaceae</td>
<td>Listeria monocytogenes</td>
<td>Dengue Fever Virus Genotype 3</td>
<td>Human Parainfluenza virus 4</td>
<td>Respiratory syncytial virus - B BK polymavirus</td>
</tr>
<tr>
<td>Spirochaetaceae</td>
<td>Borrelia burgdorferi</td>
<td>Dengue Fever Virus Genotype 4</td>
<td>Human Parainfluenza virus 4a</td>
<td>Respiratory syncytial virus - B BK polymavirus</td>
</tr>
<tr>
<td>Spirochaetaceae</td>
<td>Treponema pallidum</td>
<td>Japanese Encephalitis Virus - All Genotypes</td>
<td>Human Parainfluenza virus 5</td>
<td>Respiratory syncytial virus - B BK polymavirus</td>
</tr>
<tr>
<td></td>
<td>Leptospira (multiple spp)</td>
<td>Murray Valley Encephalitis Virus - All Genotypes</td>
<td>mumps virus - A</td>
<td>Respiratory syncytial virus - B BK polymavirus</td>
</tr>
<tr>
<td></td>
<td>Bartonella henselae</td>
<td>St. Louis Encephalitis Virus - All Genotypes</td>
<td>mumps virus - B</td>
<td>Respiratory syncytial virus - B BK polymavirus</td>
</tr>
<tr>
<td></td>
<td>Brucella (multiple spp)</td>
<td>West Nile Virus - All Genotypes</td>
<td>mumps virus - D</td>
<td>Respiratory syncytial virus - B BK polymavirus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tick-borne Encephalitis Virus - All Genotypes</td>
<td>mumps virus - D</td>
<td>Respiratory syncytial virus - B BK polymavirus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yellow Fever Virus - All Genotypes</td>
<td>mumps virus - F</td>
<td>Respiratory syncytial virus - B BK polymavirus</td>
</tr>
</tbody>
</table>

| Arbovirus Family             | Species                        | Arbovirus | Lympoctic choriomeningitis mammarenavirus | henipavirus - M |
| Rhabdoviridae                | Rhabdovirus 1 - Rabies         | Rhabdovirus A |                                              |               |
| Rhabdoviridae                | Rhabdovirus 4 - Duvenhage       | Rhabdovirus A |                                              |               |
| Rhabdoviridae                | Rhabdovirus 5 - European Bat Lyssavirus 1 (EBLV2) | Rhabdovirus A |                                              |               |
| Rhabdoviridae                | Rhabdovirus 6 - European Bat Lyssavirus 2 (EBLV1) | Rhabdovirus A |                                              |               |
| Rhabdoviridae                | Rhabdovirus 7 - Australian Bat Lyssavirus | Rhabdovirus A |                                              |               |
| Rhabdoviridae                | Rhabdovirus 8 - Lagos Bat | Rhabdovirus A |                                              |               |
| Rhabdoviridae                | Rhabdovirus 9 - Mokola virus  | Rhabdovirus A |                                              |               |
| togaviridae                  | Eastern Equine Encephalitis Virus | Rhabdovirus A |                                              |               |
|                              | Western Equine Encephalitis Virus | Rhabdovirus A |                                              |               |
|                              | Venezuelan Equine Encephalitis Virus | Rhabdovirus A |                                              |               |
|                              | Rubella virus                  | Rhabdovirus A |                                              |               |
There are two types of metagenomics:

(1) Hay classification

(2) Needle detection