Next-generation sequencing applied to molecular virology in a transfusion medicine laboratory

Virginie Sauvage, PhD
Département des Agents Transmissibles par le Sang,
Centre National de Référence pour les Hépatites B et C et du VIH en Transfusion,
Institut National de la Transfusion Sanguine, Paris

HEMO 2016
Florianopolis, Brazil
10-13 November 2016
Conflict of Interest Disclosure

I hereby declare the following potential conflicts of interest concerning my presentation:

- Consultancy: **None**
- Research Funding: **None**
- Honoraria: **None**
- Patents and Royalties: **None**
- Membership on an Entity’s Board of Directors or Advisory Committees: **None**
- Discussion of off-label drug use: **None**
Basic Principle

Sanger = First generation (ABI: 1986)

NGS = Second generation (454 Roche, Ion Torrent, Illumina)

NGS = Third generation (PacBio, Oxford Nanopore, …)
Second generation sequencing

From www.biorigami.com
## The most popular sequencers of the second generation

<table>
<thead>
<tr>
<th>Reads length</th>
<th>2x300 pb (v3)</th>
<th>200 et 400pb (500 pb expected)</th>
<th>200 pb (300 pb expected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single reads</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flow cell standard</td>
<td>Ion 314 Chip v2</td>
<td>Ion 316 Chip v2</td>
<td>Ion 318 Chip v2</td>
</tr>
<tr>
<td>25 millions</td>
<td>400-550 mille</td>
<td>2-3 millions</td>
<td>4-5,5 millions</td>
</tr>
<tr>
<td>Output data/run</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 Gb (paired end)</td>
<td>200 bases = 30-50 Mb</td>
<td>= 300 -600 Mb</td>
<td>= 600 Mb-1Gb</td>
</tr>
<tr>
<td>400 bases = 60 -100 Mb</td>
<td>= 4,4 hr</td>
<td>= 7,3 hr</td>
<td></td>
</tr>
<tr>
<td>Run Time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 heures</td>
<td>200 bases = 2,3 hr</td>
<td>= 3,0 hr</td>
<td>= 4,4 hr</td>
</tr>
<tr>
<td>400 bases = 3,7 hr</td>
<td>= 4,9 hr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applications</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Séquençage petit génome</td>
<td>V (+++)</td>
<td>V (+++)</td>
<td>V (+++)</td>
</tr>
<tr>
<td>Séquençage d’amplicons et régions génomiques ciblées</td>
<td>V (+++)</td>
<td>V (+)</td>
<td>V (+)</td>
</tr>
<tr>
<td>Transcriptome</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>ChIP-Seq (détermination zone intéraction protéine-ADN)</td>
<td>V</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Exome</td>
<td>V</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Génome Humain</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Métagénomique 16S</td>
<td>V</td>
<td>X</td>
<td>V</td>
</tr>
<tr>
<td>Métagénomique sans a priori</td>
<td>V (+)</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Third generation sequencing: single molecule

- **Characteristics:**
  - No clonal amplification
  - Single molecule DNA sequencing, long reads (>10kb)
  - Real-time sequencing

The Sequel System:
The Scalable Platform for SMRT Sequencing

MinION mkI
(~10 x 4 x 2 cm, 75g)
Interest of deep sequencing in Virology

High throughput sequencing allows the acquisition of large amounts of sequence data that can:

- Capture low-frequency drug resistance mutations
- Analyse the genetic diversity of viral populations intra- and inter-host and mixed infections
- Encompass the whole genome

- Characterization of viral biodiversity in different organs, organisms or environments (= virome)
- Virus discovery

→ metagenomics

Chevaliez, Gastroenterology, 2012

FouloungneV et al., Plos One, 2012
The term «metagenomics» refers to the collection of all microbial nucleic acids present in a given sample by passing the need for isolation and lab culture of individual species (Plasma, serum, stool, cerebrospinal fluid, urine...and environmental samples).

All microbial DNA sequences recovered from a sample (= metagenome) are randomly shear and can be:
- cloned in an expression system and screen for functional activities
- or sequenced

The application of High Throughput Sequencing (HTS) for de novo sequencing in viral metagenomics has demonstrated the power of this “without a priori” knowledge approach to identify and discover novel viruses.

In 2008, first example of a human viral pathogen discovered using metagenomics and HTS

Identification of a novel human polyomavirus (MCPyV) as a causative agent of a skin cancer (Merkel cell carcinoma)

High Throughput sequencing and metagenomics: a novel approach for Viral Discovery

- Since, this approach was particularly used:
  - to investigate infectious diseases of unknown etiology

- to describe viral or bacterial communities and their diversity in environmental and clinical samples of human
In the framework of a Pathogen Discovery program launched in 2009 at the Institut Pasteur, we identified several novel human and animal viruses.

**Identification of 2 novel DNA viruses during a study of human skin microbiome:**

- **the first human gyrovirus (HGyV)**
  (Maggi F et al., Emerg Infect Dis, 2012 ; Biagini P et al., Emerg Infect Dis, 2013)

- **a ninth human polyomavirus (HPyV9)**

**Description of the high diversity of DNA viruses identified on the human skin**
High Throughput sequencing and metagenomics: a novel approach for Viral Discovery

- Identification of the first member of a novel *Picornaviridae* genus related to parechovirus (RNA viruses) during a study of the fecal microbiome from healthy piglets


<table>
<thead>
<tr>
<th>Taxonomy</th>
<th>% homology</th>
<th>match length</th>
<th>config length</th>
<th>blast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Parechovirus; Human parechovirus</td>
<td>30.3</td>
<td>198</td>
<td>206</td>
<td>blastx nrprot</td>
</tr>
<tr>
<td>Viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Parechovirus; Human parechovirus</td>
<td>32</td>
<td>225</td>
<td>227</td>
<td>blastx nrprot</td>
</tr>
<tr>
<td>Viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Parechovirus; Human parechovirus 1</td>
<td>62</td>
<td>150</td>
<td>1172</td>
<td>blastn nt</td>
</tr>
<tr>
<td>Viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Parechovirus; Human parechovirus 1</td>
<td>78.57</td>
<td>140</td>
<td>384</td>
<td>blastn nt</td>
</tr>
<tr>
<td>Viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Parechovirus; Human parechovirus 1</td>
<td>21.86</td>
<td>364</td>
<td>442</td>
<td>blastx nrprot</td>
</tr>
<tr>
<td>Viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Parechovirus; Human parechovirus 1</td>
<td>31.08</td>
<td>975</td>
<td>860</td>
<td>blastx nrprot</td>
</tr>
<tr>
<td>Viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Parechovirus; Human parechovirus 1</td>
<td>41.16</td>
<td>265</td>
<td>254</td>
<td>blastx nrprot</td>
</tr>
<tr>
<td>Viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Parechovirus; Human parechovirus 3</td>
<td>61.54</td>
<td>195</td>
<td>1518</td>
<td>blastn nt</td>
</tr>
<tr>
<td>Viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Parechovirus; Human parechovirus 5</td>
<td>42.42</td>
<td>99</td>
<td>116</td>
<td>blastx nrprot</td>
</tr>
<tr>
<td>Viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Parechovirus; Human parechovirus 5</td>
<td>30.12</td>
<td>249</td>
<td>252</td>
<td>blastx nrprot</td>
</tr>
<tr>
<td>Viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Parechovirus; Human parechovirus 5</td>
<td>72.6</td>
<td>73</td>
<td>155</td>
<td>blastn nt</td>
</tr>
</tbody>
</table>

6896nt

→ Blastx analysis of the complete genome = 31% aa identity and 50% similarity to Ljungan Virus
• The risk of blood-borne agent transmission related to well-known viruses (HIV, HBV, HCV, and HTLV) can be considered as controlled, especially in developed countries.

• However, transmission by transfusion of new emerging infectious agents remains a continuing threat.

• This is particularly relevant considering that a significant proportion of transfused subjects is immunocompromised.

• Several measures to prevent transmission of viruses have been implemented:
  - the exclusion of at-risk donors,
  - leukocyte reduction of labile blood products,
  - and physical or chemical treatments aimed to inactivate infectious agents.

• However, there are currently no universal methods to inactive pathogens:
  - applicable for all blood components (not currently applicable on red blood cell concentrate)
  - effective for all classes of agents (Sobral et al., 2012)
Blood safety

- Some inactivation procedures are less effective on non-enveloped viruses (as human parvovirus B19, Hepatitis A (HAV) and Hepatitis E (HEV) viruses).

- Inadequate capacity of some viral reduction methods to inactivate very high viremic concentration of such viruses in plasma pools and plasma derivates (Stramer et al., 2009; Zhang et al., 2012).

- In this context, unknown or unsuspected viruses, responsible for persistent and asymptomatic viremia in blood donor, remain a threat for blood safety.

There is a need to improve the safety of blood products by engaging a surveillance of unknown infectious agents that may affect humans.

Viral metagenomics is a powerful tool to provide information on new viral pathogens that could impair blood safety.
Study protocol for plasma samples

1- Sample Preparation

Pool of 4 plasmas

Pre-treatment
(e.g. centrifugation and DNase treatment)

Roche Nucleic acids extraction
(DNA and RNA)

WTA (Qiagen)
= Reverse transcription and random amplification

HTS on WTA product (= DNA molecule)

2- Library construction
Study protocol for plasma samples

3- High Throughput Sequencing

Ion Chef™ Instrument (supports sequencing preparation)

Ion Proton™ Sequencer

70-80 millions raw sequences (reads)
(length of 200 pb with PI chip)

Cleaning
(e.g. low quality reads, short reads, duplicated reads)

Host decontamination
(Mapping of reads on human genome reference hg38)

Taxonomic classification using Kaiju
/Search by Blastx in microbial proteins database/

Identification of known viruses and NEW ONES
(and also bacterial sequences)

Refining
(Blastn in nucleotide database)

Plasma virome
Study protocol for plasma samples

• **What is the sensitivity of our pipeline?** (sample preparation and bioinformatics analysis)

The pipeline’s performances had been assessed on clinical datasets from blood samples artificially spiked with 7 viral pathogens at various titers.

<table>
<thead>
<tr>
<th>Spike viruses’ copy numbers introduced in sample</th>
<th>ssDNA</th>
<th>ssDNA</th>
<th>ssRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome type</td>
<td>aDNA</td>
<td>ssDNA</td>
<td>ssRNA</td>
</tr>
<tr>
<td>Spikes</td>
<td>Hepatitis B virus</td>
<td>Cowpox virus</td>
<td>Parvovirus B19</td>
</tr>
<tr>
<td>Copies/mL</td>
<td>$10^7$</td>
<td>$10^3$</td>
<td>$2.5 \times 10^3$</td>
</tr>
</tbody>
</table>

Sensitivity around 500 cp/mL in individual plasma sample
Study protocol for plasma samples

• **Evaluation of the pipeline on a Viral Mutiplex Reference (NIBSC code: 11-242-xxx)**
  
  - The preparation:
    - contains **25 viruses** representing a range of common hazard group 2 human viruses with a variety of genome and envelope types
    - is intended for use as a reference in adventitious virus detection assays employing deep sequencing technology (*Mee Et al.*, *Vaccine*, 2016:2035-43)

  
  18 Viruses detected by real-time PCR (NIBSC)

  7 Viruses NOT detected by real-time PCR (NIBSC)

  → **23/25 viruses** identified by viral metagenomics deep sequencing
Projects prospects

• Emerging infectious diseases (EID) may arise anywhere in the world

• EID can move rapidly across country boundaries, territories and continents by means of infected travelers or international trades (animals, plants and food products)

• Then, it is of paramount importance to implement a pro-active system of assessing the threat and risk of emerging infections for their potential impact on blood safety

• Southern countries such as Latin America, the Sub-Saharan Africa and Asia appear as emerging zoonotic diseases «hotspots» (Jones et al., Nature, 2008)

• Ongoing project: Analyzed potential infectious disease threats through viral discovery in human blood donations collected:
  • in different countries of Sub-Saharan Africa in 2012 (ready to order 16,000 blood donations from Burundi, Cameroon, Madagascar, Mali, Mauritania, Niger and the Democratic Republic of Congo)
  • and two different geographic regions of Brazil, the Amazon basin and the Brazilian highlands (São Paulo state)
Viral metagenomics and blood safety

**Sub-saharan Africa:**
42 Pools of 4 plasma donations analyzed
(Republic democratic of Congo, Madagascar, Cameroon and Mali)

<table>
<thead>
<tr>
<th>Manip 28 MAU-4 28/07/2016</th>
<th>MAU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>25303</td>
<td>14/06/2012</td>
</tr>
<tr>
<td>25356</td>
<td>14/06/2012</td>
</tr>
<tr>
<td>25468</td>
<td>19/06/2012</td>
</tr>
<tr>
<td>26607</td>
<td>30/06/1986</td>
</tr>
</tbody>
</table>

**2 reads HIV-1**
27 reads Human papillomavirus
Reads Human Pegivirus/GBV-C

HIV load n°26607: 539 cp/mL
(Cobas TaqMan HIV-1 Test V2.0, Roche)

<table>
<thead>
<tr>
<th>Manip 25 MAU-6 22/07/2016</th>
<th>MAU-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>25352</td>
<td>14/06/2012</td>
</tr>
<tr>
<td>25366</td>
<td>14/06/2012</td>
</tr>
<tr>
<td>25397</td>
<td>17/06/2012</td>
</tr>
<tr>
<td>25527</td>
<td>20/06/2012</td>
</tr>
</tbody>
</table>

**112 reads HIV-1**
Anelloviruses

HIV load n°25352: 11048 cp/mL
(Cobas TaqMan HIV-1 Test V2.0, Roche)

<table>
<thead>
<tr>
<th>Manip 7 RDC 3 15/06/2016</th>
<th>RDC-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD 006 CF</td>
<td>27/01/2012</td>
</tr>
<tr>
<td>CD 19 GALI</td>
<td>03/02/2012</td>
</tr>
<tr>
<td>CD 08 GALI</td>
<td>03/02/2012</td>
</tr>
<tr>
<td>37 TER</td>
<td>23/02/2012</td>
</tr>
</tbody>
</table>

**833 reads HCV**
GBV-C
Anelloviruses
6 reads Gemycircularvirus

HCV load n°CD 006 CF: 3 330 UI/mL
(Cobas TaqMan HCV Test V2.0, Roche)
## Viral metagenomics and blood safety

<table>
<thead>
<tr>
<th>Manip 26</th>
<th>MAU-7</th>
<th>26/07/2016</th>
<th>MAU-7</th>
<th>26/07/2016</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23627</td>
<td>09/04/012</td>
<td>ND</td>
<td>F</td>
<td>35</td>
</tr>
<tr>
<td>25325</td>
<td>14/06/012</td>
<td>DC</td>
<td>F</td>
<td>40</td>
</tr>
<tr>
<td>25364</td>
<td>14/06/012</td>
<td>ND</td>
<td>F</td>
<td>40</td>
</tr>
<tr>
<td>25537</td>
<td>20/06/012</td>
<td>ND</td>
<td>F</td>
<td>38</td>
</tr>
</tbody>
</table>

**12 reads HBV**

**Anelloviruses ++**

- **HBV load n°25364**: 442 UI/mL (Cobas TaqMan HBV Test, Roche)

<table>
<thead>
<tr>
<th>Manip 30</th>
<th>MAL11</th>
<th>03/08/2016</th>
<th>MAL-11</th>
<th>03/08/2016</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15411</td>
<td>09/05/2012</td>
<td>DC</td>
<td>M</td>
<td>60</td>
</tr>
<tr>
<td>14308</td>
<td>02/05/2012</td>
<td>DC</td>
<td>M</td>
<td>46</td>
</tr>
<tr>
<td>14018</td>
<td>27/04/2012</td>
<td>ND</td>
<td>F</td>
<td>58</td>
</tr>
<tr>
<td>13365</td>
<td>24/04/2012</td>
<td>DC</td>
<td>M</td>
<td>56</td>
</tr>
</tbody>
</table>

**Anelloviruses ++**

- **3 reads Parovirus B19** (=Human erythroivirus génotype 1)

- **Sample n°15411**: NS1-VP1 nested PCR pos

**Brazil** (Amazon basin):

- 4 Pools of 4 plasma donations analyzed

- **One pool**: 302 reads of Zika virus (viral load unknown)
Conclusion

- Human Plasma virome is mainly composed of Anelloviruses and the Flavivirus GBV-C, which cause chronic viral infection without any associated disease.

- Viral concentrations detectable with the sample preparation method and the depth of sequencing used: ≥ 500 copies/mL in individual plasma sample.

- No identification of new viral agents.

Viral Metagenomics and NGS:
A tool of surveillance and identification of infectious agents for blood safety.

Metagenomic detection of new viruses in the blood supply (1)

Marseillevirus-Like Virus Recovered From Blood Donated by Asymptomatic Humans

None of the 339 studied samples tested positive for Marseillevirus-like virus DNA.

GBM DNA in 3/30 (10%) of blood donations and in 4% (7/174) of blood donations and in 9.1% (2/22) of patients with thalassemia.

GBM DNA in 2/30 (10%) of blood donations

None of the 339 studied samples tested positive for Marseillevirus-like virus DNA.

Correspondence


Absence of giant blood Marseillevirus-like virus DNA detection by polymerase chain reaction in plasma from healthy US blood donors and serum from multiply transfused patients from Cameroon

Among the 1,364 plasma samples analyzed, no positive PCR signals for HHpgV-1/HPgV-2 were detected (Manuscript in preparation).
Discovery of a highly divergent DNA virus at the interface between the Paroviridae and Circoviridae, designated NIH-CQV

NIH-CQV detected in Chinese patients with seronegative (non-A-E) hepatitis

NIH-CQV is a laboratory reagent contaminant from silica-binding spin columns used for nucleic acid extraction and not bona fide infectious agents of humans
Acknowledgments

- **Syria Laperche**, MD-PhD, Head of the Blood-borne agents department, INTS, Paris
  slaperche@ints.fr

- **Pierre Cappy**, MD-PhD, Assistant of the Head of the blood-borne agents department,
  pcappy@ints.fr

- **Research programs:**
  - **Daniel Candotti**, PhD, dcandotti@ints.fr
  - **Camille Sureau**, PhD, csureau@ints.fr
  - **Virginie Sauvage**, PhD, vsauvage@ints.fr

- **Staff**: Laure Boizeau, Johanna Gomez, Clémence Hamon, Christine Portal, Quentin Lucas

- **Valérie Caro**, PhD, Head of genotyping Pole of pathogens (PGP), Institut Pasteur, Paris
  valerie.caro@pasteur.fr

- **Bioinformaticians**:
  - **Mathias Vandenbogaert**, PhD, mathias.vandenbogaert@pasteur.fr
  - **Anais Barray**, student

- **Staff**: Laure Diancourt, Charlotte Balière