Deep sequencing for virus detection and characterization with special emphasis on *Prunus* and cherry viruses
Plant virus diversity is huge!
Detection & characterization with or without prior knowledge...

- Detection of a **known virus** is now (generally) straightforward
  - ELISA, immuno-printing, lateral flow assays...
  - Molecular hybridization (dot blot, tissue print...)
  - **PCR** in different formats (or RCA, isothermal amplification...)

- Broad detection, **without prior knowledge**, is much more challenging due to viral diversity
  - No typically viral **property** to look for (**no ITS or 16S...**)
  - No **unique test** for detection of all phytoviruses
  - Certification or Quarantine approaches largely virus-specific and not « all inclusive »
  - Etiology studies long and complicated
High throughput sequencing (NGS) changes that...

- **Pathogen detection** can be achieved generating sequences from a sample and by sifting sequences to separate viral ones from host or environmental reads.
- Viruses not accessible in free form hence need to sequence a complex pool of molecules.
- Various possible sequencing templates:
  - total RNAs, mRNAs or DNA (+/- substraction)
  - double-stranded RNAs
  - small interfering RNAs (siRNAs)
- **Bioinformatics challenge** replaces technical or biological experimental challenges.
- The number of novel viruses discovered using these approaches is rapidly increasing...
Analysis of plant EST banks

- EST: « Expressed Sequence Tag »
- cDNA single-pass sequences from mRNAs
- 5’ or 3’ reads, some banks subtracted or normalized
- Large number of banks, variety of plants, WEB-queryable in Genebank

- Complete screening of Genebank plant EST resources using TBlastX and BlastN with all complete reference genomes for RNA plant viruses (+ 1 representative animal virus per genus)

- Identify viruses (% identity) or validate viral nature by reverse BlastX of the EST hit on Genebank nr
Lessons from plant EST analyses

- **Possibility** to determine the infection status of a plant by deep sequencing of cDNAs (from mRNAs)

- An exhaustive search requires significant sequencing depth (≥ 100,000 reads)
  - Confirmed by Al Rwahnih et al. *(Virology, 2009)*
    
    Grapevine infected by 4 viruses and 3 viroids, only 3 viruses detected with 65,000 reads

- **Possibility to reduce sequencing depth if enrichment of viral sequences**
  - Subtracted or normalized libraries : **x5-15**
  - dsRNA analysis : **x28** [Al Rwahnih et al., Virology, 2009]

- Novel ssRNA viruses ≈ known viruses
- **Novel dsRNA viruses >> known viruses**
Virus-enriched targets

• Double-stranded RNAs (dsRNAs)
  – Produced during RNA virus replication
  – Normally absent from healthy hosts (but cryptic viruses...)
  – Can be readily purified using cellulose CF11 chromatography

• Small interfering RNAs (siRNAs)
  – 21-24 nt long molecules
  – Produced by host silencing machinery
  – Not virus specific (silencing of host genes) but high proportion of viral origin in case of viral infection
  – Cover all the virus genome
  – Produced for DNA & RNA viruses & viroids
dsRNAs extraction and amplification

• Optimized purification of dsRNAs
  – 2x CF11 cellulose chromatography
  – DNAse & RNAse treatment

• Random amplification (WGA) & Tagging
  – cDNA synthesis (pdN₆ & dT₁₈)
  – cDNA amplification with WGA primers containing MID tags

- MID = Multiplex identifiers
- 10 bp tags, tolerate up to 2 errors and still allow unambiguous identification of sample
- available for >100 multiplexing schemes
Typical bioinformatic analysis

- **Filtering** (quality, minimal read length)
- **Demultiplexing** *(Optional : additional filtering)*
- *Optional:* host genome substraction
- *Optional:* known viruses substraction
- *Optional:* duplicated reads removal
- **de novo contig assembly** *(home developed pipeline or CLC Genomics Workbench)*
- **Contig** *(Optional: singletons) annotation* by BlastN & BlastX on Genbank *(E-value cut-off : $10^{-3}$ to $10^{-5}$)*
- Parsing of annotation results to identify potential viruses/contigs of interest
- Optional: mapping of reads/contigs on reference viral genomes to identify known viruses present
- For virus characterization, further **manual or automated contig extension** *(various strategies)*
dsRNA analysis of *Prunus* samples

- 8 *Prunus* sources, containing various known of unknown agents, or with diseases of unknown etiology
  - Pair N°1 [*Prunus salicina*, unknown origin (Asia?)]: PBNSPaV, PNRSV, novel *Betaflexiviridae*
  - Peach from China (PPV, APV3....)
  - Shirofugen stunt disease (*sour cherry sample*)
  - Krikon necrosis disease (Plum sample; *Prunus domestica*)
  - False red marbling disease (2 sources)
  - Apricot vein necrosis disease (?)
  - Peach gummosis disease (?)
### Statistics on viruses detected

<table>
<thead>
<tr>
<th>Virus</th>
<th>Sample</th>
<th>Number of contigs</th>
<th>Reads in Contigs</th>
<th>% total reads</th>
<th>Contigs length</th>
<th>Genome length</th>
<th>% of genome length</th>
<th>% of genome length</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBNSPaV</td>
<td>b</td>
<td>18</td>
<td>17833</td>
<td>60,9%</td>
<td>25318</td>
<td>14214</td>
<td>178,1%</td>
<td>2 isolates</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>7</td>
<td>8098</td>
<td>34,3%</td>
<td>15304</td>
<td>14214</td>
<td>107,7%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>e</td>
<td>20</td>
<td>7142</td>
<td>43,0%</td>
<td>28236</td>
<td>14214</td>
<td>198,6%</td>
<td>2 isolates</td>
</tr>
<tr>
<td></td>
<td>j</td>
<td>38</td>
<td>13685</td>
<td>52,5%</td>
<td>30282</td>
<td>14214</td>
<td>213,0%</td>
<td>2 isolates</td>
</tr>
<tr>
<td>PNRSV</td>
<td>b</td>
<td>8</td>
<td>384</td>
<td>1,3%</td>
<td>6966</td>
<td>7874</td>
<td>88,5%</td>
<td></td>
</tr>
<tr>
<td>APV3</td>
<td>d</td>
<td>5</td>
<td>1595</td>
<td>6,7%</td>
<td>9723</td>
<td>9409</td>
<td>103,3%</td>
<td></td>
</tr>
<tr>
<td>PPV</td>
<td>d</td>
<td>3</td>
<td>12289</td>
<td>52,0%</td>
<td>9839</td>
<td>9736</td>
<td>101,1%</td>
<td></td>
</tr>
<tr>
<td>LChV1</td>
<td>g</td>
<td>5</td>
<td>5332</td>
<td>34,1%</td>
<td>16894</td>
<td>16936</td>
<td>99,8%</td>
<td></td>
</tr>
<tr>
<td>CVA</td>
<td>i</td>
<td>8</td>
<td>28480</td>
<td>79,3%</td>
<td>9141</td>
<td>7383</td>
<td>123,8%</td>
<td></td>
</tr>
<tr>
<td>New Betaflexi</td>
<td>b</td>
<td>1</td>
<td>3153</td>
<td>10,8%</td>
<td>8126</td>
<td>8287</td>
<td>98,1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>e</td>
<td>5</td>
<td>335</td>
<td>2,0%</td>
<td>6141</td>
<td>8287</td>
<td>74,1%</td>
<td></td>
</tr>
</tbody>
</table>

- **Generally for each virus**
  - One to a few (<10-15) contigs
  - from 1,3 to 79% of reads for the sample
  - >70% of genome length recovered, often close to 100%
Coverage of PNRSV genome

8 contigs spanning the genome
5’ and 3’ genome ends missing as generally observed
dsRNA assembly of novel *Betaflexiviridae*

**3153 reads**
**94x average coverage**

**8,13kb contig, missing sequence: 33nt 3' & ~125nt 5'**
Some viruses characterized

- PBNSPaV strains including one that is not detected by current PCR assays
- Non-cherry CVA strain potentially associated with the Krikon necrosis disease
- Novel *Betaflexiviridae* with affinities to Citriviruses but with an extra 3’ ORF
siRNAs vs dsRNAs seq in Prunus

• **siRNAs**
  - Illumina GAIIx sequencing
  - After filtering $\sim 3 \times 10^6$ siRNAs per sample, 21-24nt long

• **dsRNAs**
  - 454 pyrosequencing
  - After filtering $\sim 2 \times 10^4$ reads per sample ($\sim 300$nt long)

• Demultiplexing & contig building CLC Genomics Workbench. Contigs > 60-100bp

• Contig annotation: BlastN & BlastX on Genbank
  - highest e-value of BlastN & BlastX
  - E-value cut-off: $10^{-3}$
### siRNAs vs dsRNAs seq in Prunus

<table>
<thead>
<tr>
<th></th>
<th>Number of contigs</th>
<th>Longest contig</th>
<th>Best e-value</th>
<th>Total contigs length</th>
<th>Number of reads in contigs</th>
<th>% of total reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBNSPaV</td>
<td>83</td>
<td>869 nt</td>
<td>0</td>
<td>11 301</td>
<td>177 863</td>
<td>6,3%</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>2229 nt</td>
<td>0</td>
<td>13 717</td>
<td>1398</td>
<td>12,1%</td>
</tr>
<tr>
<td>APV2+APV3</td>
<td>57</td>
<td>189 nt</td>
<td>1.8e-58</td>
<td>4843</td>
<td>17 094</td>
<td>0,6%</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>2652 nt</td>
<td>0</td>
<td>16 797</td>
<td>4376</td>
<td>37.7%</td>
</tr>
<tr>
<td>ACLSV</td>
<td>4</td>
<td>86 nt</td>
<td>1.5e-8</td>
<td>332</td>
<td>547</td>
<td>0,02%</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>674 nt</td>
<td>4.8e-96</td>
<td>2822</td>
<td>43</td>
<td>0,3%</td>
</tr>
<tr>
<td>New Betaflexiviridae</td>
<td>12</td>
<td>100 nt</td>
<td>7.1</td>
<td>850</td>
<td>943</td>
<td>0,02%</td>
</tr>
<tr>
<td>PLMVd viroid</td>
<td>1</td>
<td>8126 nt</td>
<td>0</td>
<td>8 126</td>
<td>3 153</td>
<td>10,8%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>337 nt</td>
<td>7.1e-106</td>
<td>409</td>
<td>183 047</td>
<td>6,5%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>290 nt</td>
<td>3e-123</td>
<td>290 nt</td>
<td>1</td>
<td>0,008%</td>
</tr>
</tbody>
</table>
PBNSPaV assembly from « Pair N°1 » sample siRNAs

104 PBNSPaV contigs (for a total of 76230 sequences siRNAs)
11.7 kb total, 82.2% of PBNSPaV genome

75245 siRNAs aligned on the PBNSPaV genome
near complete coverage (average depth 115x) largest gap <20 nt
Conclusions

• Both technologies OK, feasible.

• With *Prunus* (woody hosts ?) siRNAs results less effective than initial report with sweet potato samples

• siRNA contig building has to be optimized, annotation may be inefficient if low number of reads and novel virus

• mapping offers a more sensitive alternative but (1) needs a negative control and (2) inefficient to detect novel agents

• Need >300K (better several millions) siRNAs for detection of some agents through *de novo* assembly

• At this stage, *dsRNA*-based approach very efficient and more cost effective than siRNAs-sequencing *(but viroids and DNA viruses not efficiently detected through dsRNAs)*
More questions, perspectives

- NGS technologies have the potential to drastically change the field of viral indexing
- How much can we multiplex? To link with novel sequencing platforms: more sequences, lower costs
- Contaminations, significativity threshold …
- Mixed infections, impact on the detection of viral species producing low amounts of dsRNAs?
- Improve the throughput further: automatisation (extraction, post-sequencing bioinformatics…)
- Overcoming the limits of virus identification by homology searches? (host genome substraction…)
- Finding a virus does not demonstrate that it causes the disease!....
Application to cherry samples
Shirofugen stunt disease (SSD)

**Shirofugen stunt disease**

- syndrome in *P. serrulata* (flowering cherry) cv. Shirofugen indicators grafted with some sweet or sour cherry sources
- strong rosetting, with dwarfed and deformed leaves, reduced vigor, gradual union necrosis and sometimes die-off of indicator after a few growth cycles
- not or few symptoms in typical cherry indicators, sometimes netting of older leaves of F12-1, Sam or Bing
- symptoms reminiscent of those reported for the « Kwanzan stunting » syndrome
- causal agent unknown, probably viral

Candresse *et al.*, 2013. *Phytopathology*, **103**:293-298,
dsRNA analysis of V2356 Shiro Stunt source

- Multiplexed 454 pyrosequencing of cDNAs produced by random WGA amplifications of purified cDNAs
- 21766 reads
- 15646 reads after quality filtering and removal of reads <60nt
- 279 contigs >100 bp assembled using CLC Genomics Workbench
  - Average 399nt, max 5 kb
  - Average 38 reads, max 2482
dsRNA analysis of V2356 SSD source (1)

• BlastN and BlastX annotation of contigs against GenBank database

• 5 contigs with high homology to LChV1, no other viral agent significantly identified
  - Ct91: 3,9kb; 1620 reads, 112x average coverage, e-value=0
  - Ct121: 0,3kb, 24 reads, 13x coverage, e-value=1,04xe^{-8}
  - Ct208: 5kb, 896 reads, 45x coverage, e-value=0
  - Ct212: 2,7kb, 310 reads, 27x coverage, e-value=0
  - Ct266: 5kb, 2482 reads, 131x coverage, e-value=0

• Total length of contigs 16,89kb (99,7% of LChV1)

• Total of 5332 reads in contigs (34% of reads)
• Manual assembly: 3 supercontigs, leaving only two gaps (17 and 55 bases, missing also 2nt at 5’ end)

• Gaps filled by targeted PCR, partial confirmation resequencing (no errors, 85x average coverage).

• Typical LChV genetic organization, colinear with reference sequence with very few indels, mostly in non-coding regions
• highly divergent but still within the 25% aa divergence in Pol, HSP70 or CP accepted for species members in the *Closteroviridae* family

• first full genome sequence for second phylogenetic cluster of LChV1 isolates
LChV1 & SSD conclusions

• Very efficient characterization of LChV1 isolate by 454 pyrosequencing of dsRNAs

• First complete genome for second LChV1 cluster

• No other viral agent identified in SSD source

• A LChV1 isolate identified in a second source of Shirofugen stunt disease

• LChV1 identified by Matic et al., 2012 in a Kwanzan stunting disease source

• No complete demonstration by strong suspicion that LChV1 isolates are associated with these two probably related syndromes
Going after novel cherry viruses...

- Cherry samples from various sources tested over time with the polyvalent PDO-RT-PCR (Betaflexiviridae)

- PCR products with sequences suggestive of novel flexiviruses identified in:
  - Old Italian cherry variety (collaboration with M. Barone, D. Alioto and A. Raggozzino, Napoli)
  - Czech cherry variety (collaboration with M. Navratil, Olomouc)

- Perform multiplexed 454 pyrosequencing of purified dsRNAs (14 samples, 1/8th of a 454 run, imbalance in multiplexing pool !)
<table>
<thead>
<tr>
<th></th>
<th>Italian cherry</th>
<th>Czech cherry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total reads</td>
<td>1415</td>
<td>1456</td>
</tr>
<tr>
<td>Filtered reads (&gt;60nt)</td>
<td>1151</td>
<td>1125</td>
</tr>
<tr>
<td>Contigs (&gt;100nt)</td>
<td>104</td>
<td>91</td>
</tr>
<tr>
<td>Reads in contigs</td>
<td>822</td>
<td>938</td>
</tr>
<tr>
<td>% reads in contigs</td>
<td>71.4%</td>
<td>83.3%</td>
</tr>
</tbody>
</table>

- **Assembly of contigs and analysis performed using CLC Genomics Workbench**
- **Contigs & singletons annotated using BlastN and BlastX (e-value 10^{-5} cut-off)**
- **Contigs & singletons mapped against reference Prunus virus genomes**
• Multiple viral contigs identified for 4 viruses

<table>
<thead>
<tr>
<th>Virus</th>
<th>Reads</th>
<th>% of total reads</th>
<th>Contigs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple chlorotic leaf spot virus</td>
<td>36</td>
<td>3.1%</td>
<td>9</td>
</tr>
<tr>
<td>Little cherry virus 1</td>
<td>240</td>
<td>20.8%</td>
<td>14</td>
</tr>
<tr>
<td>Prune dwarf virus RNA1</td>
<td>19</td>
<td>1.6%</td>
<td>3</td>
</tr>
<tr>
<td>Prune dwarf virus RNA2</td>
<td>4</td>
<td>0.3%</td>
<td>1</td>
</tr>
<tr>
<td>Prune dwarf virus RNA3</td>
<td>37</td>
<td>3.2%</td>
<td>4</td>
</tr>
<tr>
<td>New Flexiviridae</td>
<td>353</td>
<td>30.7%</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>689</td>
<td>59.8%</td>
<td>43</td>
</tr>
</tbody>
</table>
**New Flexiviridae in Italian cherry (1)**

**Difficulties**

- only 5 contigs (~3,8kb) annotated as viral, other relevant contigs annotated as unknown (7 of 12 !) and not available for initial assembly
- largest viral contig is only 1,5 kb
- do all contigs belong to a single virus? [contigs had best BlastX values against *Scaevaola virus A, Banana mild mosaic virus, Mint virus 2, Potato virus T, Peach mosaic virus*]
- Possibility of improperly assembled contigs, of contaminations…
- Possibility of multiple strains of a single agent
- Fill the gaps and sequence the genome ends by PCR
New *Flexiviridae* in Italian cherry (2)

- Genomic organization similar to that of ACLSV (*Trichovirus*) or of *Potato virus T* (potential *Tepovirus*)

- RdRp which clusters together (bootstrap support) with PVT and, to a lesser extent to Vitiviruses

- But only 35.3% aminoacid sequence identity
New *Flexiviridae* in Italian cherry (3)

- No clear affinities in MP or CP trees
- Only 21.8% (MP) and 30.9% (CP) aa identity with PVT
- Clearly a novel *Betaflexiviridae*, possibly in the *Tepovirus* genus
- Another strain of the same virus identified by dsRNA NGS in a plum from Azerbaijan
Czech cherry (1)

- Analyzed because a PDO-RT-PCR survey suggested presence of a *Flexiviridae*, possibly close to the new virus in the Italian cherry sample.

Fragment of phylogenetic tree obtained using the PDO fragment (RNA Pol)
• dsRNAs from a pool of leaves from 3 trees was sequenced (1125 reads!)
• Multiple contigs identified for LChV1 (5 contigs, 53 reads), PDV (2 contigs, 65 reads), PNRSV (4 contigs, 36 reads)
• A single contig (4 reads, 462nt) with interesting *Flexiviridae* BlastX (*Actinidia virus A*, 3.48E-19) >>> *Flexivirus* cannot be easily “seen” at this sequencing depth >>> deeper sequencing in progress
• 9 contigs (93 reads) with significant BlastX values with RNA1 and RNA2 of *Secoviridae* members:
  – *Cucurbit mild mosaic virus (Fabavirus)*
  – *Gentian mosaic virus (Fabavirus)*
  – *Broad bean wilt virus 1 (Fabavirus)*
  – *Mikania micrantha mosaic virus (Fabavirus)*
• but protein identity <40% with known viruses >>> novel agent (Faba?)
• further assembly much complicated because low sequencing depth plus presence of at least 2 strains… >>> deeper sequencing from individual plants in progress
Conclusion: cherry viruses

- Full genome sequence of a divergent LChV1 isolate, likely associated with the Shirofugen stunt disease
  - Single virus infection, assembly easy with 15,000 reads

- Full genome of novel Betaflexiviridae completed by PCR from partial assembly of dsRNA reads (3.8 kb, 5 contigs)
  - Assembly only partial because low depth (1151 reads) and mixed infection (ACLSV, LChV1 and PDV in addition to the novel virus)

- Indications (very partial sequences) for the presence of both a second novel Flexiviridae and a novel Secoviridae in the Czech cherry sample.
  - Assembly very poor (low depth, multiple strains, mixed infection)
Thanks to

• Bordeaux Team
  A. Marais
  C. Faure
  T. Candresse

• Valencia Team
  A. Olmos
  M. Cambra

• Napoli Team
  M. Barone
  D. Alioto
  A. Ragozzino

• Olomouc Team
  M. Navratil

• Bergerac Ctifl Team
  (P. Gentit)
  Y. Brans