NGS, omics and applied bioinformatics at CVI

The first steps in NGS…

Alex Bossers, Freddy de Bree, and Frank Harders et al.

Department of Infection Biology
Pathogenomics and Applied Bioinformatics group

Nbic::Bioassist 17 December 2010
CVI - Lelystad

- Animal disease control
- Prevention of transmission to humans
- Diagnoses and crisis organization
- Development of animal models and methods/pathobiology
- Development of models (epidemiology)
- Development of diagnostic tests
- Development of intervention tools (vaccines, therapeutics)
- Pathobiology, animal models and clinical studies
Host-pathogen interactions

- **OMICS:**
  - Proteome / transcriptome / metabolome
  - Immunology and pathogenesis

Vaccine and diagnostics development
Bioinformatics at CVI

- **Applied bioinformatics**
  - CVI: Frank, Freddy, Alex
  - Collaborations include
    - Life stock Research (ASG) and Plant Research International (PRI)
    - Academic Medical Centre Amsterdam
    - UManchester, UMaryland/NBACC, Roslin, Sanger, and others…

- **Current/past focus**
  - Reversed vaccinology approaches
  - NGS
  - CGH/tiling array approaches (molecular epidemiology)
  - *In situ* proteome array approaches
  - Transcriptomics and analysis
  - HT genotyping and automation

- **Future focus (2011-2012)**
  - Virus discovery and pathogen identification (NGS and DEFRA biochip :: EpiZone chip)
  - NGS transcriptomics vs arrays
  - Pathogen detection/typing solutions (arrays (solid/solution/tubes))
  - Metagenomics (NGS)
NGS - CVI

- CVI: Sanger capillary sequencing and array platforms
- Roche 454 at PRI-WUR (Wageningen)
- Sequencing unit AMC (Amsterdam)
  - Roche 454 / SOLID
- Roche 454 at Onderstepoort Veterinary Institute (SA)
- Illumina at BaseClear (Leiden)
- Illumina and HiSeq at DNAVision (B)
Pathogens - focus

- *Coxiella burnetii* (Q fever)
- *Streptococcus Suis*
- others...

- Blue Tongue (BTV)
- Avian influenza
- AngHV1 (eel)
- Prions/TSEs
- others…
Knowledge-based vaccine design

- Reversed vaccinology approaches
  - Sequence (knowledge) based vaccine design
  - Vaccine-candidate reduction
    - In silico
      - Localization
      - Motifs for post-translational modifications and immunomodulation
    - In vitro by proteome approaches
      - 2D, SELDI/mass-spec, Western blots,…
      - In situ protein/proteome arrays
  - Knowledge building host::pathogen interactions
    - Transcriptome of host and pathogen

Bossers et al. 2010
Reversed vaccinology

- Successful examples where others failed
  - *Neisseria Menigitidis* (subtype B)
  - *S pneumoniae* (group B, effective in infants)
  - *Porphyromonas gingivalis*
  - *Chlamidia pneumoniae*
  - *Influenza* (universal vaccine)
  - *S suis* (?)
  - ....
Reversed vaccinology

NGS

Genome
2-4Mb

Proteome
2-4k ORFs

Genome variation/plasticity

In silico approaches

Localization
Predict exposed antigens
Motifs for modification (lipo)

In situ protein arrays

2D SELDI

Candidate reduction for vaccines/diagnostics

Bossers et al. 2010
NGS at CVI

- **First series** *Streptococcus suis*
  - Several serotypes (different pathogenicity)
  - 454-FLX
    - 61-135 contigs *de novo* assembled (1/3 fc each)

- **Second series** *S. suis* (ongoing)
  - 454 titanium using paired-end 8kb and 20kb
    - 5-20 scaffolds *de novo* assembled
  - Illumina MP 5kb
    - >2600 scaffolds *de novo*

- **Hot series**
  - Q fever 10 isolates
    - 454 titanium shotgun and PE runs
    - Illumina MP
  - BTV and AHSV
    - 454 titanium (OVI)
CVI Christmas tree

Raw 454 data

de novo assembly (contigs/scaffolds)

\textit{gsAssembler::Newbler}

\textbf{IMAGE}
Iterative Mapping and Assembly for Gap Elimination
1. align the paired end reads onto draft sequence

2. local assembly of the aligned reads; new contigs are produced

3. gaps are now shortened. Repeat the whole procedure in a few iterations

4. The gap is now closed

Overview of the IMAGE process. Step one, Illumina reads are aligned against the initial assembly. Step two, Illumina reads that align to contig ends, along with their non-aligning mate adjacent to gaps, are assembled into new contigs, which are subsequently mapped back to the initial assembly. Step three, Illumina reads are aligned against the updated assembly and the whole process is repeated iteratively until the gap is closed.

Tsai et al. Genome Biology 2010 11:R41 doi:10.1186/gb-2010-11-4-r41
Artificial chromosome (synteny by mapping)

- Align/map individual contigs to reference genome by BLAT
- Determine optimal contig order
- Join contigs by linkers (6 frame start/stop)
- Track contig and joiner positions for Artemis viewing
Artificial chromosome (Sanger::Artemis)

- Join contigs by linkers (6 frame start/stop)
- Track contig and joiner positions for Artemis viewing
**Genome comparison** (before and after synteny join)

*NucMer algorithm (MUMmer)*

- 8067 chronological contigs vs P1/7
- 8067 artificial chromosome vs P1/7
Quick “QC” of assembled NGS data

Coverage and quality score per base
Quick "QC" of assembled NGS data
Coverage and quality score per base

Bossers et al. 2010
Coverage plots
### Genome pairwise comparison (MUMmer matrix n x n)

<table>
<thead>
<tr>
<th></th>
<th>P1/7</th>
<th>8067</th>
<th>7917</th>
<th>BM407</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1/7</td>
<td><img src="matrix1.png" alt="Matrix" /></td>
<td><img src="matrix2.png" alt="Matrix" /></td>
<td><img src="matrix3.png" alt="Matrix" /></td>
<td><img src="matrix4.png" alt="Matrix" /></td>
</tr>
<tr>
<td>8067</td>
<td>-</td>
<td><img src="matrix5.png" alt="Matrix" /></td>
<td><img src="matrix6.png" alt="Matrix" /></td>
<td><img src="matrix7.png" alt="Matrix" /></td>
</tr>
<tr>
<td>7917</td>
<td>-</td>
<td>-</td>
<td><img src="matrix8.png" alt="Matrix" /></td>
<td><img src="matrix9.png" alt="Matrix" /></td>
</tr>
</tbody>
</table>

Bossers et al. 2010
Comparison serotype 1 - serotype 2

Genomes very homologous: additional sequences in serotype 1

Serotype 1

Serotype 2

Bossers et al. 2010
Serotype 7: antimicrobial activity

Micrococcus luteus
4) *S. suis* serotype 7
5) *S. suis* serotype 9
6) *S. suis* serotype 9
The **Galaxy @ WUR** development server (galaxy02) is up and running... Please be aware that this 6c 24G machine is for development/TESTING. Uptime is not guaranteed. Not all tool dependencies have been fulfilled yet!

For regular use, please use the stable production 32c 64G server (galaxy01).

More info on the galaxy@wur wiki: [//galaxy or //galaxy.wur.net.nl](//galaxy.wur.net.nl)

Contact the development team for tool requests, errors and mods.

Current galaxy@WUR development and testing team include of: Alex, Freddy, Ina, Arun, Janti, and Frank. Please [join-in](//galaxy.wur.net.nl) to make this server work for all of us!

The **Galaxy project** is developed by the galaxy development team at PennState University. It's supported in part by NSF, Penn State, NG4RI, and the Huck Institutes of the Life Sciences.
Pangenome tiling arrays

Molecular epidemiology (*arrays are not dead yet!*)

- **Panarray** (probe-tiling design to cover all pan-genomic features)
  - Adam Philippy *et al.* Bioinformatics Sep 2009

- **Mapping / filtering / sorting**
  - BLAT/MUMmer/…..

- **eArray** to actual Agilent array
Challenges and upcoming work

- Work
  - IMAGE assembly improvement/gap-closure
  - Complete several genomes for publication
  - Improve reversed vaccinology tools
    - Annotations/motifs
    - Easy filtering differences
  - Reshuffling annotations for changed draft genomes
  - Link galaxy directly to Artemis and ACT

- Challenges
  - Let agent expert do the work but stay in control
  - Virusdiscovery
  - NGS transcriptomics
Questions…

Alex.Bossers@wur.nl