MiSeq pathogenomics applications at CVI

13 September 2012 Nbic::BioAssist::NGS

Alex Bossers, Freddy de Bree, Frank Harders, et al.
Department of Infection Biology :: PathogenOmics-group
CVI – Lelystad (part of Wageningen-UR)

- 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

www.cvi.wur.nl
PathogenOmics

- Focus on
  - Understanding (micro)pathogenesis
  - Understanding pathogen variation
  - Identification of targets for vaccines and diagnostics
  - Using **NGS**, arrays (**HD/tubes/luminex, NA/protein/peptides**), and smart bioinformatics (**analysis/design**)
Why the MiSeq?!

- **Smaller genomes**
  - Viral (segmented >13kb)
  - Plasmids (50-100kb)
  - Bacteria (2-5Mb)
  - Parasites (~60Mb)

- **Price of instrument:** ~€ 85k

- **Price per run:** ~€ 1.5k +80/index (incl. depreciation, Key-user hours, IT and some maintenance. PE150 kit ~850 euro excl. libprep time)

- **Throughput** (currently ~2-2.6Gb/run)

- **Speed** (~27h/PE150-run) and Ease-of-use

- **Near certification for diagnostics**
MiSeq Performance Improvements

Next Six Months

- 3.4 M clusters/run
- 2 x 150 bp read length
- 5 min cycle time

100’s Mb

1-1.5 Gb

- 5-7 M clusters/run
- 2 x 150 bp read length
- 5 min cycle time

7 Gb

- ~15 M clusters/run
- 2 x 250 bp read length
- 4 Gb at 2 x 150
- 6 - 7 Gb at 2 x 250
- <5 min cycle time

Read lengths get longer - 66% increase
Output grows - 3 fold improvement in number of clusters run
### MiSeq experience (since February 2012)

<table>
<thead>
<tr>
<th>MiSeek_run_no:</th>
<th>Run_name:</th>
<th>Species</th>
<th>Sample</th>
<th>Samples</th>
<th>Genome_size</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Training test run met PhiX</td>
<td>CB</td>
<td>Genome</td>
<td>1</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>2</td>
<td>120209_M00374_0002_AMS0008595-00300</td>
<td>CB</td>
<td>Genome</td>
<td>1</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>3</td>
<td>problems during run, valve?</td>
<td>CB</td>
<td>Genome</td>
<td>1</td>
<td>13.000</td>
<td>ScriptSeq v2</td>
</tr>
<tr>
<td>4</td>
<td>testrun with Phix and SBV</td>
<td>CB</td>
<td>Genome</td>
<td>1</td>
<td>12.000</td>
<td>ScriptSeq v2 (RNAseq)</td>
</tr>
<tr>
<td>5</td>
<td>120217_M00374_0005_AMS0009336-00300</td>
<td>SVB</td>
<td>Virus</td>
<td>1</td>
<td>12.000</td>
<td>ScriptSeq v2 (RNAseq)</td>
</tr>
<tr>
<td>6</td>
<td>120222_M00374_0006_AMS0009501-00300</td>
<td>Plasmid</td>
<td>Plasmid</td>
<td>12</td>
<td>100.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>7</td>
<td>120409_M00374_0007_AMS0005263-00300</td>
<td>Ssu</td>
<td>Phage</td>
<td>3</td>
<td>35.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>8</td>
<td>120410_M00374_0008_AMS0005324-00300</td>
<td>CB</td>
<td>plasmid</td>
<td>5</td>
<td>40.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>9</td>
<td>120420_M00374_0010_AMS0005287-00300</td>
<td>Ssu</td>
<td>Genome</td>
<td>4</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>10</td>
<td>120420_M00374_0012_AMS0005287-00300</td>
<td>Ssu</td>
<td>Genome</td>
<td>4</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>11</td>
<td>testrun with Phix and samples PW16, 21995</td>
<td>CB</td>
<td>Genome</td>
<td>3</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>12</td>
<td>120420_M00374_0013_AMS0005287-00300</td>
<td>Ssu</td>
<td>Genome</td>
<td>4</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>13</td>
<td>013_MS0005410</td>
<td>CB</td>
<td>Genome</td>
<td>3</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>14</td>
<td>Pump of the Miseq is replaced</td>
<td>CB</td>
<td>Genome</td>
<td>3</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>15</td>
<td>Error runs</td>
<td>CB</td>
<td>Genome</td>
<td>3</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>16</td>
<td>Error runs</td>
<td>CB</td>
<td>Genome</td>
<td>3</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>17</td>
<td>Error runs</td>
<td>CB</td>
<td>Genome</td>
<td>3</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>18</td>
<td>Error runs</td>
<td>CB</td>
<td>Genome</td>
<td>3</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>19</td>
<td>Error runs</td>
<td>CB</td>
<td>Genome</td>
<td>3</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>20</td>
<td>020_MS0005259</td>
<td>CB</td>
<td>Genome</td>
<td>3</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>21</td>
<td>Error runs</td>
<td>CB</td>
<td>Genome</td>
<td>3</td>
<td>2.000.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>22</td>
<td>022_MS0005275</td>
<td>SBV</td>
<td>Genome</td>
<td>1</td>
<td>12.000</td>
<td>ScriptSeq v2</td>
</tr>
<tr>
<td>23</td>
<td>023_8014160-002_RNAseq</td>
<td>CB</td>
<td>RNA</td>
<td>1</td>
<td>12.000</td>
<td>ScriptSeq v2 (RNAseq)</td>
</tr>
<tr>
<td>24</td>
<td>120622_M00374_0024_AMS00000000-A19T6</td>
<td>AI</td>
<td>Genome</td>
<td>12</td>
<td>12.000</td>
<td>ScriptSeq v2</td>
</tr>
<tr>
<td>25</td>
<td>120702_M00374_0025_AMS00000000-A1A4B</td>
<td>Ecoli</td>
<td>Genome</td>
<td>3</td>
<td>5.500.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>26</td>
<td>120703_M00374_0026_AMS00000000-A1A45</td>
<td>Ecoli</td>
<td>Genome</td>
<td>3</td>
<td>5.500.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>27</td>
<td>120704_M00374_0027_AMS00000000-A1A0W</td>
<td>Ecoli</td>
<td>Genome</td>
<td>3</td>
<td>5.500.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>28</td>
<td>120705_M00374_0028_AMS00000000-A1A8H</td>
<td>Ecoli</td>
<td>Genome</td>
<td>3</td>
<td>5.500.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>29</td>
<td>120709_M00374_0029_AMS00000000-A1AE5</td>
<td>Ecoli</td>
<td>Genome</td>
<td>3</td>
<td>5.500.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>30</td>
<td>120710_M00374_0030_AMS00000000-A1A9P</td>
<td>Ecoli</td>
<td>Genome</td>
<td>3</td>
<td>5.500.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>31</td>
<td>120711_M00374_0031_AMS00000000-A1AET</td>
<td>Ecoli</td>
<td>Genome</td>
<td>3</td>
<td>5.500.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>32</td>
<td>120713_M00374_0032_AMS00000000-A1A46</td>
<td>Ecoli</td>
<td>Genome</td>
<td>3</td>
<td>5.500.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>33</td>
<td>120714_M00374_0033_AMS00000000-A1AFM</td>
<td>Ecoli</td>
<td>Genome</td>
<td>2</td>
<td>5.500.000</td>
<td>Nextera_DNA_Sample_prep_kit</td>
</tr>
<tr>
<td>34</td>
<td>120716_M00374_0034_AMS00000000-A1AA4</td>
<td>MicroBioTa</td>
<td>16s</td>
<td>2</td>
<td>na</td>
<td>Custom primer set V3 and V4</td>
</tr>
</tbody>
</table>
## MiSeq experience (since February 2012)

<table>
<thead>
<tr>
<th>run</th>
<th>clusters</th>
<th>PE Reads (M)</th>
<th>&gt;Q30 (%)</th>
<th>PastFilter (%)</th>
<th>Final PE reads (M)</th>
<th>Final Gb/run</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>1.185</td>
<td>9.9</td>
<td>88.4</td>
<td>89.56</td>
<td>8.82</td>
<td>2.66</td>
</tr>
<tr>
<td>25</td>
<td>815</td>
<td>6.6</td>
<td>88.7</td>
<td>91.33</td>
<td>6.04</td>
<td>1.82</td>
</tr>
<tr>
<td>27</td>
<td>1.111</td>
<td>8.7</td>
<td>85.6</td>
<td>88.45</td>
<td>7.71</td>
<td>2.33</td>
</tr>
<tr>
<td>29</td>
<td>1.393</td>
<td>9.5</td>
<td>80.7</td>
<td>78.34</td>
<td>7.48</td>
<td>2.26</td>
</tr>
<tr>
<td>31</td>
<td>1.421</td>
<td>10.1</td>
<td>80.9</td>
<td>80.51</td>
<td>8.11</td>
<td>2.45</td>
</tr>
<tr>
<td>32</td>
<td>1.396</td>
<td>8.2</td>
<td>74.2</td>
<td>76.13</td>
<td>6.25</td>
<td>1.89</td>
</tr>
<tr>
<td>33</td>
<td>1.225</td>
<td>8.5</td>
<td>73.7</td>
<td>77.98</td>
<td>6.63</td>
<td>2.00</td>
</tr>
<tr>
<td>34</td>
<td>963</td>
<td>7.7</td>
<td>80.9</td>
<td>92.7</td>
<td>7.09</td>
<td>2.14</td>
</tr>
</tbody>
</table>
NGS – overall work flow

- Biological question
- Choose method/technique
  - Sample prep
  - Sample QC
  - Library prep + QC
  - Sequencing
- Data QC
- Data analysis
Sample (free of host material) → DNA (genomic, plasmid, phage, amplicons**) → Quality Control (Qubit, gel) → TruSeq (DNA library prep, indexing)

DNA sample input ~5ug → Shearing (Covaris or Nebulizer***) → Library prep (and indexing) (~8h) → QQC (Qubit and Bioanalyzer) → Sample ready for the MiSeq

Nextera (DNA library prep, indexing)

DNA sample input ~50ng → Library prep (and indexing) (~2h) → QQC (Qubit and Bioanalyzer) → Sample ready for the MiSeq

**not available at this time, discuss with key-users for possible solutions

***recommended
Sample (free of host material)

RNA (totRNA, mRNA, smalRNA, bacterial, viral)

Quality Control (Qubit, gel)

RNA (totRNA, mRNA, microbial, viral)
  add. sample preps (add ~5h, RiboZero kits)
  ScriptSeq v2 (totRNA library prep, ~8h, indexing)
  500pg – 50ng
  Library prep (~12h)
  QQC (Qubit and Bioanalyzer)

RNA# (totRNA, microbial, mRNA, viral)
  add. sample preps (add ~5h, RiboZero kits)
  TruSeq v2 (totRNA library prep, ~8h, indexing)
  0.1–4ug(tot), 10–400ng(m)
  Library prep (~12h)
  QQC (Qubit and Bioanalyzer)

RNA (small RNA, microbial, viral)
  add. sample preps (add ~5h, RiboZero kits)
  ScriptMiner (small RNA library prep, ~12h, directional, indexing)
  5ug totRNA, 100 pg size fract RNA
  Library prep (~12h)
  QQC (Qubit and Bioanalyzer)

Sample ready for the MiSeq
MiSeq experience

- Only PE150 runs
- Main applications;
  - Nextera genome sequencing (TAGmentation)
  - ScriptseqV2
  - Amplicon 16S metagenomics
- Routinely 1.8-2Gb/run (~6.5M PE reads)
- >75% but mostly >81% quality over Q30
ScriptSeqV2 (example Schmallenberg virus)

- ScriptSeqV2: random hexamer-tagging (epibio.com)

- Genome RNA 3 segments
  - S: ~830b
  - M: ~4.4k
  - L: ~6.8k

- Mapping against “reference” and de novo assemblies
ScriptSeqV2 (example Schmallenberg virus)

Coverage: 4k/b

Coverage: 21k/b

Coverage: 58k/b
Nextera genome sequencing

- Currently the MAIN application
- TAGmentation sequencing (epibio.com)
- PE150

- Analysis
NGS analysis work flow

**Raw NGS data**
- 454 shotgun, PE 3kb
- Illumina mp50, PE150 (250)

**Mapping**
- (Bowtie)

**Filter data**
- (de novo) assembly

**Assembly**
- Newbler (454)
- MIRA (mixed)
- ABYSS (mixed)
- MUMmer / BLA(S)T (reference)
- SSPACE (MP/PE data)
- IMAGE (MP/PE data)

**Scaffolding contigs**

**Gap closure**

**Annotation (BLAST)**

**Biological Interpretation**

**FastQC**

**BioInformatics**
De novo assembly (which one to choose)

Main application: ABYSS

EHEC ~5.2Mb

Ssuis ~2.2Mb
De novo assembly (which one to choose)

Main application: ABYSS

Cummulative contigsize over contignr (length descending)
Sample 3262

Cb ~2Mb
De novo assembly (which one to choose)

Main application: ABYSS

[Graphs and charts showing data points and relationships]
MiSeq: metagenomics (16S)

Complex sample-Deep sequencing

INTESTINAL MICROFLORA

$10^{14}$ micro-organisms, >500 different species

<table>
<thead>
<tr>
<th>Environment</th>
<th>Micro-organisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stomach</td>
<td>$10^2$ to $10^3$</td>
</tr>
<tr>
<td>Duodenum</td>
<td>$&lt;10^{4.5}$</td>
</tr>
<tr>
<td>Jejunum</td>
<td>$10^3$ to $10^7$</td>
</tr>
<tr>
<td>Ileum</td>
<td>$10^9$ to $10^{12}$</td>
</tr>
<tr>
<td>Colon with appendix</td>
<td></td>
</tr>
</tbody>
</table>
MiSeq: metagenomics (16S)

- Variable regions ribosomal 16S -> species classification
Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample

Caporasoa et al. 2010 PNAS 10.1073

Figure 1: Amplification Strategy and Perfect Paired-End Read

A. V4 was amplified from each sample using primers 515F and 806R paired with P5 and P7 sequences, respectively. Paired 150 bp sequencing gives a full-length 254 bp fragment of V4 with a 46 bp overlap. B. Raw intensities (matrix and phasing corrected) for an example perfect 254 bp paired-end read from the V4 library.

!!! Sample(s) must be spiked with ~60% PhiX or other reference !!!
Generation of multi-million 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads

Bartram et al. 2011 ApplEnvironMicrobiol 77:3846-52

Adapted for v3 and cluster diversity (4N)

~105 bp overlap

~200 bp

!!! No spike needed, introduction 4 N’s in sequence !!!
Both V3 and V4 region tested

Analysis 6 samples/run:

- NGS PE150 data (all PF data)
- PandaSeq (Q assisted overlap assembly)
  - V3: ~105-125bp
  - V4: ~16-46bp
- QIIME 1.5 pipeline for classification (qiime.org)

Testing overlap assembly
PANDAseq (assess minimal needed overlap)
PANDAseq (assess minimal needed overlap)
PANDAseq (assess Q threshold)

V3o20 threshold test PANDAseq

- t threshold value

Count

Duizenden

LOWQ
NOALGN
OK
Assess assembly effects on QIIME classification

QIIME classification at FAMILY level

- V3.20.2
- V3.20.6
- V3.20.9
- V4.10.6
- V4.20.6

Legend:
- OTHER-accum
- Pasteurellaceae
- Clostridiaceae
- Streptococcaceae
- Veillonellaceae
- Lactobacillaceae
Both V3 and V4 region operational

Analysis 6 samples/run:

QIIME 1.5 pipeline for classification *(based on GreenGenes DB)*

- **FAST** compared to Illumina pipeline

Testing assembly settings on different sample types

Future MiSeq Upgrade:

- Combine V3 and V4 into one 450bp fragment
MiSeq at CVI

- Quite happy with performance thus far

- Future Exome sequencing (parasites)

- Adapt indexing mechanism (avoid ~€80/sample_index)

- Planned: First FastQC and mapping directly linked to dedicated galaxy-instance.

- Machine open for shared-use (www.CAT-AGROfood.wur.nl)
Omics is TEAMPLAY!

Follow us in the social media

Linkedin

twitter

BioStar

SEQanswers the next generation sequencing community

BiomedExperts Your scientific professional network

Dik Mevius
Frank Harders
Jan Priem
Hilde Smith

Freddy de Bree

Alexander Bossers alex.bossers@wur.nl

CENTRAL VETERINARY INSTITUTE WAGENINGEN UR

nbic
Spare slides
Just in case....
Pathogen variability
Whole-genome alignment / phylogenetics

54 ESBL plasmids (~100kb)
Complete and draft
Gene alignment / phylogenetics

131 AI-variants (HA-based)
Whole-genome alignment / phylogenetics

17 Streptococci complete and draft genomes (2Mb each)
Welcome to the public GALAXY@WUR server!

This galaxy instance is a shared initiative by the Central Veterinary Institute and Plant Research International, both of Wageningen UR.

Galaxy@WUR data policy and disclaimer

Be sure to read the DISCLAIMER AND DATA POLICY of galaxy@WUR!
By continuing and using this server you agree to its policy and potential limitations.

Recent changes and updates are logged in the galaxy wiki

DISK capacity problems (27 Jan 2012)

We currently experience serious disk capacity problems.
Please cleanup unneeded files in your histories NOW! (DISCLAIMER AND DATA POLICY)

Galaxy server maintenance

MAJOR maintenance is scheduled EARLY February 2012.
At that time we will also update to the latest galaxy release.

Current galaxy@WUR development and testing team includes:
Freddy de Bree (CVI), Alex Bossers (CVI), Frank Harders (CVI),
Jan van Haarst (PRI), Henri vd Geest (PRI),
and Pas te Uitent Heikert (PRI).

Please join up to make this server work for all of us!
Galaxy tools/modules

- NGS toolbox
  - QC / Clipping / filtering
  - Mapping (bowtie, bwa, MUMmer, lastZ, ...)
  - (Assembly)
- Variation
  - MUMmer, multiMUMmer, MUMi, Bowtie, bwa, samtools
- Generic parsers
- Annotation
  - Artemis conversion/porting tools
  - Gene/ORF prediction (geneMarkHMM, Prodigal)
  - Protein classification (PSORTb, SignalP, LipoP, SPAAN, TMHMM, ...)
  - BLAT, BLAST, BLAST+, ...
- Etc.

http://galaxy.wur.nl