Visualizing NGS data with GenomeView

June 2\textsuperscript{nd}, 2010, Espoo, Finland

Thomas Abeel – thomas@abeel.be
VIB - Ghent University, Gent, Belgium
Broad Institute of MIT and Harvard, Cambridge, MA, USA
Overview

• Introduction to NGS
  – What is NGS? Platforms?
  – Applications of NGS:
    • Assembly
    • Mapping: RNA-seq, chip-seq, resequencing, etc.

• Visualizing NGS and other data
  – Description of GenomeView
NGS INTRODUCTION
Next-gen sequencing

- Next-generations sequencing machine
  - ABI Solid, Illumina (Solexa), and Roche 454
  - Helicos Biosciences, Pacific Biosciences
- Pour DNA in machine and you get reads
- Read = fragment of sequenced DNA

<table>
<thead>
<tr>
<th>Gen.</th>
<th>1st</th>
<th>next</th>
<th>next</th>
<th>next</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform</td>
<td>Sanger</td>
<td>Roche 454</td>
<td>Illumina</td>
<td>SOLiD</td>
</tr>
<tr>
<td>Data/run</td>
<td>100kb</td>
<td>500 Mb</td>
<td>8 Gb</td>
<td>16Gb</td>
</tr>
<tr>
<td>Read len.</td>
<td>1000 nt</td>
<td>500 nt</td>
<td>100 nt</td>
<td>50 nt</td>
</tr>
<tr>
<td>$/genome</td>
<td>1,000k</td>
<td>25k</td>
<td>25k</td>
<td>25k</td>
</tr>
</tbody>
</table>
1\textsuperscript{st}, next, 3\textsuperscript{rd}?

- Discussion what’s ‘next’-generation?

<table>
<thead>
<tr>
<th>Generation</th>
<th>Platforms</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sanger</td>
<td>Production</td>
</tr>
<tr>
<td>2 (next)</td>
<td>454</td>
<td>Production</td>
</tr>
<tr>
<td>2.5 (next)</td>
<td>Illumina, Solid</td>
<td>Production</td>
</tr>
<tr>
<td>3</td>
<td>Helicos, Pacbio</td>
<td>First commercial models available</td>
</tr>
<tr>
<td>3.x (not available)</td>
<td>Nanopore, ....</td>
<td>Prototype</td>
</tr>
</tbody>
</table>

- For ease of discussion, NGS will be used for anything not Sanger seq.

June 2, 2010
NGS workshop, Espoo, Finland, Thomas Abeel
What to do with short reads

• Assembly
  – At least 40x coverage
  – Result is large number of contigs

• Map to a reference genome
  – Re-sequencing → SNP discovery
  – ChIP-seq
  – RNA-seq
Re-sequencing

- Sequence the genome of an organism for which you already have a genomic sequence
  - Genomic diversity: SNPs, indels, but also structural variation
RNA-seq

- Whole genome transcriptome sequencing
- Extract mRNA by picking up the poly-A tail
- Sequence cDNA
RNA-seq applications

• Gene expression
  – If a read maps to a gene, that gene is expressed

• Expression measure
  – If 10 reads map to the same place in a gene, it’s 10 fold expressed

• Condition/cell type specific expression

• Microarray replacement

• Annotation
ChIP-seq

- **ChIP** produces a library of target DNA to which a protein of interest binds in-vivo
- **-seq** sequences the library
- Less bias than ChIP-on-chip because not limited to probes
- **Application:**
  - Find transcription factor binding sites
Short read mapping

• Each of the proposed applications boil down to aligning the reads to a reference sequence

• Characteristics:
  – Many short queries (reads)
  – Single large reference (reference)
  – Mismatches allowed
Alignment challenges

• Efficiency
  – Need to align several billion reads (300 Gb)
  – To a genome of several billion nucleotides (3 Gb)
  – Preferably over the weekend

• Ambiguity
  – Sequencing errors
  – Genetic variation
  – Alignment with mismatches
Assigning reads

• 36 nt → with a couple of mismatches reads may align to multiple locations in the genome

• Unique: where it maps

• Non-unique:
  – all locations
  – random
  – nowhere
### Short read aligners

- Bfast
- BioScope
- Bowtie
- BWA
- CLC bio
- CloudBurst
- Eland/Eland2
- GenomeMapper
- GnuMap
- Karma
- MAQ
- MOM
- Mosaik
- MrFAST/MrsFAST
- NovoAlign
- PASS
- PerM
- RazerS
- RMAP
- SSAHA2
- Segemehl
- SeqMap
- SHRiMP
- Slider/SliderII
- SOAP/SOAP2
- Srprism
- Stampy
- vmatch
- ZOOM
- ...
Typical NGS analysis pipeline

• Extracted DNA (or somebody else did it)
• Put it in machine, got a lot of data in return
• We have aligned it to a reference sequence
• We now have an even bigger file with even more information.
Analyses

• SNP calling
• Peak detection
• Gene prediction
• Transcript identification/quantification
• Visualization
Reasons for visualization

• Taking a look at the data
  – Sanity check on the data
  – Hypothesis generation

• Provide insights in large-scale data sets
  – Augment ability to reason about complex data
  – Make it easier to develop algorithms
  – The appropriate image makes the solution obvious
VISUALIZING NGS DATA

GenomeView
GenomeView

• Interactive genome browser/editor
• Annotations and mapped experimental data
• Short read alignments (*-seq)
• Multiple alignments
GUI Overview
Basic features
Short read mappings

Green = forward mapped read
Blue = reverse mapped read
Yellow = mismatch
Black = read insertion
Red = read deletion
Purple = connector for paired-end reads and spliced reads
Quality, indels and mismatches
SNP track
Re-sequencing demo
SNPs

<table>
<thead>
<tr>
<th>12137</th>
<th>12145</th>
<th>12153</th>
<th>12161</th>
<th>12169</th>
<th>12177</th>
<th>12185</th>
<th>12193</th>
<th>12201</th>
<th>12209</th>
<th>1221</th>
</tr>
</thead>
</table>

http://www.youtube.com/watch?v=KPgARXGbDaM
RNA-seq demo
RNA-seq

http://www.youtube.com/watch?v=PgBU1gWCkWU
Chip-seq
Other features

Plug-ins, editor, integration, multiple alignments
Plug-ins

• Computational analyses can be plugged in
  – Gene prediction
    • Core promoter prediction (EP3, Prosom)
    • Splice site prediction (SpliceMachine)
    • Translation start site prediction (StartScan)
    • Coding potential prediction
  – Physical properties of DNA
  – Blast
  – ...
Browser integration

Select polymorphism

Go to polymorphism detail page

SNP in short read alignment

View polymorphism in GV

Select strain with alternative allele

June 2, 2010
Editor integration

Protein Domains

- **External ID**: 50474
- **From Database**: superfamly
- **Description**: alpha/beta-Hydrolase

- **External ID**: Q3OSA3:40.60.1820
- **From Database**: Gene3D
- **Description**: no description

Protein Homologs

- **n/a**

Gene Structure

- **View in ConemsvView**
- **View in Aronmini**

Structure:

```
07704 07817 07994 09944 100173 30105 50191 7018 7111
CDACGTTTCAATCAGAAGGATCGGATCGATGGTTAATCTTGGCT
```
Saving back to the server...
Multiple alignments
Next-gen multiple alignments
Things to come

• Indexed sequence data structure
• Indexed feature data structure
• Faster and more interactive
Summary

• NGS data sets are large and daunting
• Visualization makes them less daunting
• Visualization is essential
• GenomeView can visualize a broad set of NGS data sets (and other types of data)

• GenomeView is freely available @ http://genomeview.org
Acknowledgements

- Yves Van de Peer
- Yvan Saeys
- James Galagan
- Thomas Van Parys

http://genomeview.org