Analysis of high-throughput sequencing data using Galaxy (ChIP-seq and RNA-seq)

Denis Puthier, Claire Rioualen & Jacques van Helden
Aix-Marseille Univ, INSERM, TAGC lab, Marseille, France
Goals of the workshop

- **Target audience**
  - Biologists involved in NGS projects.
  - No prior experience of NGS bioinformatics.

- **Approach**
  - Practice-driven.
  - Elements of theory interspersed in the tutorials.

- **Scope**
  - Study cases from ChIP-seq and RNA-seq.
  - However many concepts and tools are also used by many other applications.

- **Software environment**
  - Mainly Galaxy
  - Visualisation with IGV
  - Web sites for specific resources.
  - R under RStudio convivial environment? To be discussed...
Schedule

- **Days 1 - 2: ChIP-seq analysis**
  - NGS Technologies
  - ChIP-Seq analysis - Intro
  - Short read file formats
  - Quality control of the reads
  - Trimming
  - Read mapping
  - Data visualization (IGV)
  - Coverage normalisation
  - Peak calling
  - Peak annotation
  - Motif analysis

- **Days 3-4: RNA-seq**
  - RNA-Seq method intro
  - Preprocessing (Quality control, Trimming)
  - Splice-aware alignment
  - Transcript discovery
  - Data visualization
  - Quantification
  - Differential analysis
  - Functional annotation
  - Motif analysis (continued)

- **Day 5: tutorship and/or R ?**
  - Customized analytic flow charts + playing with your own data.
  - Optional: first steps with R.
Presentation of the teachers

- **Denis Puthier**
  - Bioinformatics analysis of high-throughput data.
  - Teaching domains: bioinformatics, genomics, programming, statistics.

- **Claire Rioualen**
  - Bioinformatics analysis of high-throughput data.
  - Development of workflows for NGS data (ChIP-seq, RNA-seq).

- **Jacques van Helden**
  - Teaching domains: bioinformatics, statistics, genomics.
Presentation of the participants

Participants introduce themselves in 4 sentences.

1. Name and affiliation
2. Background in biology/bioinformatics
3. Research project involving NGS / interest for NGS.
4. Prior experience with NGS bioinformatics?
## Resources used during the training

<table>
<thead>
<tr>
<th>Resource</th>
<th>Description + URL</th>
<th>To install locally</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>R statistical package <a href="https://www.r-project.org/">https://www.r-project.org/</a></td>
<td>X</td>
</tr>
<tr>
<td>RStudio</td>
<td>An environment to manage R programming and projects <a href="https://www.rstudio.com/">https://www.rstudio.com/</a></td>
<td>X</td>
</tr>
<tr>
<td>ArrayExpress</td>
<td>Gene expression database <a href="https://www.ebi.ac.uk/arrayexpress/">https://www.ebi.ac.uk/arrayexpress/</a></td>
<td></td>
</tr>
</tbody>
</table>
High-throughput sequencing
Breakthrough in DNA Sequencing

- 1977-1990, 500bp, manual analysis
- 1990-2000, 500bp, computed assisted analysis (1D capillary sequencers)
- 2005-2014, 20-1000bp (2D sequencers “Next Generation Sequencing.”)
Cost per megabase (1 million base)
Cost per human genome

**NB:** most of the methods rely on fragmented DNA/RNA material.
Important things to consider

- Sequencer throughput
  - Some applications require good coverage
    - High dynamic range, sensitivity
    - e.g. transcriptome analysis, ChIP-Seq
  - May offer multiplexing
- Read length produced
  - May be important to resolve low complexity regions
  - i.e. a word of size 20 is more ambiguous than a word of size 500

Important things to consider (continued)

- Fidelity
  - Some sequencer may be error prone
  - Fidelity may be important for variant calling (...)

- With current technologies:
  - The longer the reads \(i.e\) several kbs the weaker the fidelity and coverage
Sequencing is continuously evolving

- Technologies are subject to rapid changes!
- From this 2011 table, only a few survived in 2016.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Current company</th>
<th>Former company</th>
<th>Sequencing method</th>
<th>Amplification method</th>
<th>Claim to fame</th>
<th>Primary applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>454</td>
<td>Roche</td>
<td>454</td>
<td>Synthesis (pyrosequencing)</td>
<td>emPCR</td>
<td>First Next-Gen Sequencer, Long reads</td>
<td>1*, 2, 3*, 4, 7, 8*</td>
</tr>
<tr>
<td>Illumina</td>
<td>Illumina</td>
<td>Solexa</td>
<td>Synthesis</td>
<td>BridgePCR</td>
<td>First short-read sequencer; current leader in advantages†</td>
<td>1*, 2, 3*, 4, 5, 6, 7, 8</td>
</tr>
<tr>
<td>SOLiD</td>
<td>Life Technologies</td>
<td>Applied Biosystems</td>
<td>Ligation</td>
<td>emPCR</td>
<td>Second short-read sequencer; low error rates</td>
<td>3*, 5, 6, 8</td>
</tr>
<tr>
<td>HeliScope</td>
<td>Helicos Life</td>
<td>N/A</td>
<td>Synthesis</td>
<td>None</td>
<td>First single-molecule sequencing</td>
<td>5, 8</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>Technologies</td>
<td>Ion Torrent</td>
<td>Synthesis (H* detection)</td>
<td>emPCR</td>
<td>First Post-light sequencer; first system &lt;$100 000</td>
<td>1, 2, 3, 4, 8</td>
</tr>
<tr>
<td>PacBio</td>
<td>Pacific Biosciences</td>
<td>N/A</td>
<td>Synthesis</td>
<td>None</td>
<td>First real-time single-molecule sequencing</td>
<td>1, 2, 3, 7, 8</td>
</tr>
<tr>
<td>Starlight‡</td>
<td>Life Technologies</td>
<td>N/A</td>
<td>Synthesis</td>
<td>None</td>
<td>Single-molecule sequencing with quantum dots</td>
<td>1, 2, 7, 8</td>
</tr>
</tbody>
</table>

Bold indicates applications that are most often used, economical or growing.
1 = *de novo* BACs, plastids, microbial genomes.
2 = transcriptome characterization.
3 = targeted re-sequencing.
4 = *de novo* plant and animal genomes.
5 = re-sequencing and transcript counting.
6 = mutation detection.
7 = metagenomics.
8 = other (ChIP-Seq, μRNA-Seq, Methyl-Seq, etc.; see Brautigam & Gowik 2010, Shendure & Ji 2008).
Illumina sequencers
## Illumina sequencers

<table>
<thead>
<tr>
<th>Popular Applications &amp; Methods</th>
<th>MiniSeq System</th>
<th>MiSeq Series</th>
<th>NextSeq Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large Whole-Genome Sequencing (human, plant, animal)</td>
<td>Key Application</td>
<td>Key Application</td>
<td>Key Application</td>
</tr>
<tr>
<td>Small Whole-Genome Sequencing (microbe, virus)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exome Sequencing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted Gene Sequencing (amplicon, gene panel)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole-Transcriptome Sequencing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Expression Profiling with mRNA-Seq</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted Gene Expression Profiling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>miRNA &amp; Small RNA Analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA-Protein Interaction Analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methylation Sequencing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16S Metagenomic Sequencing</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run Time</th>
<th>4–24 hours</th>
<th>4–55 hours</th>
<th>12–30 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Output</td>
<td>7.5 Gb</td>
<td>15 Gb</td>
<td>120 Gb</td>
</tr>
<tr>
<td>Maximum Reads Per Run</td>
<td>25 million</td>
<td>25 million*</td>
<td>400 million</td>
</tr>
<tr>
<td>Maximum Read Length</td>
<td>2 × 150 bp</td>
<td>2 × 300 bp</td>
<td>2 × 150 bp</td>
</tr>
</tbody>
</table>
# NextSeq 500 Illumina sequencer

<table>
<thead>
<tr>
<th>Key Methods</th>
<th>Everyday genome, exome, transcriptome sequencing, and more.</th>
<th>Production-scale genome, exome, transcriptome sequencing, and more.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Run Mode</th>
<th>Mid-Output</th>
<th>High-Output</th>
<th>Rapid Run</th>
<th>High-Output</th>
<th>N/A</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow Cells per Run</td>
<td>1</td>
<td>1</td>
<td>1 or 2</td>
<td>1 or 2</td>
<td>1</td>
<td>1 or 2</td>
</tr>
<tr>
<td>Output Range</td>
<td>20-30 Gb</td>
<td>30-120 Gb</td>
<td>10-300 Gb</td>
<td>50-1000 Gb</td>
<td>125-750 Gb</td>
<td>125-1500 Gb</td>
</tr>
<tr>
<td>Run Time</td>
<td>15-26 hours</td>
<td>12-30 hours</td>
<td>7-60 hours</td>
<td>&lt;1-6 days</td>
<td>&lt;1-3.5 days</td>
<td>&lt;1-3.5 days</td>
</tr>
<tr>
<td>Reads per Flow Cell</td>
<td>130 million</td>
<td>400 million</td>
<td>300 million</td>
<td>2 billion</td>
<td>2.5 billion</td>
<td>2.5 billion</td>
</tr>
<tr>
<td>Maximum Read Length</td>
<td>2 x 150 bp</td>
<td>2 x 150 bp</td>
<td>2 x 250 bp</td>
<td>2 x 125 bp</td>
<td>2 x 150 bp</td>
<td>2 x 150 bp</td>
</tr>
<tr>
<td>System Overview</td>
<td>Speed and simplicity for everyday genomics.</td>
<td>Power and efficiency for large-scale genomics.</td>
<td>Maximum throughput and lowest cost for production-scale genomics.</td>
<td>Maximum throughput and lowest cost for production-scale genomics.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[https://www.illumina.com/systems/sequencing.html](https://www.illumina.com/systems/sequencing.html)
HiSeq X 10: a factory scale sequencer

- **Illumina**
- **A set of 10 sequencers.**
  - Each producing 1.8 Terabases / 3 days
- **18,000 genome / year**
  - Factory-scale sequencing technology
Some computing issues

http://glennklockwood.blogspot.nl/

- 18,000 / year ~ 340 / week
- 30-50To storage / week
  - Cost of long term storage?
- 518 core hours / genome
- 175,000 core hours per week

But 1000$ genome coming true....
Is the 1000 $ genome for real?

Whole Genome Sequencing

Description

Average 30X Coverage

Gene By Gene's whole genome sequencing service allows for a high degree of accuracy in identifying variants across the entire scope of the human genome.

Gene By Gene is excited to announce the implementation of the Arpeggi engine pipeline. Arpeggi, Inc., now part of Gene By Gene, developed internal bioinformatics software for alignment and variant calling of next generation sequencing (NGS) data that is now exclusive to Gene By Gene. If you want us to complete the alignment, mapping, and variant calling on your exome sequencing, we will show you a full report comparing different pipeline options and why the Arpeggi engine delivers you the best possible VCF file so you have the highest quality data for your project.

Results are delivered to the customer via electronic FTP transfer and are only stored by Gene By Gene for 30-60 days.

Quantity

1

Analysis

- None
- Alignment and Variant Calling ($400)

Price

$9,995.00

( $9,995.00 per sample )

Add to Cart
U.S. proposes effort to analyze DNA from 1 million people

WASHINGTON | BY TONI CLARKE AND SHARON BEGLEY

Human Longevity Inc. (HLI) Launched to Promote Healthy Aging Using Advances in Genomics and Stem Cell Therapies

HLI is Building World’s Largest Genotype/Phenotype Database by Sequencing up to 40,000 Human Genomes/Year Combined with Microbiome, Metabolome and Clinical Data to Develop Life Enhancing Therapies

HLI has Purchased Two Illumina HiSeq X Ten Sequencing Systems

SAN DIEGO, CA (March 4, 2014) — Human Longevity Inc. (HLI), a cell therapy-based diagnostic and therapeutic company focused on the healthy, life performance enhancement market, today announced it had purchased two of the most advanced next generation sequencers in the world: the Illumina HiSeq X Ten and paired HPLC.

Illumina’s Jay Flatley at #PMWC14: Get Sequence of 1 million cancer patients in next 5 years

January 27, 2014 by nextgenseek - 1 Comment

Illumina’s Jay Flatley said at #PMWC14 that Illumina wants to help sequence 1 million cancer patients in a database in the next five years. And one of his goals is to make cancer a “chronic” disease within 10 years. Jay Flatley said Illumina is working with large population datasets with researchers and clinicians at #PMWC14.

Thanks to awesome live tweets by Kevin Davies, @DiraBioTech, and links to the original tweets.
An overview of Illumina technology - sequencing by synthesis
Illumina sequencing: general principle

http://www.illumina.com/company/video-hub/HMyCqWhwB8E.html
Illumina sequencing: general principle
Starting with a fragment

- **Terminology:**
  - “Fragment” a piece of DNA
  - “Read” the sequence(s) associated to this fragment
First-strand synthesis

Annealing

Reverse-strand synthesis

Denaturation
Fragment released
Bridge-PCR

Annealing

Reverse-strand synthesis

Denaturation
Bridge-PCR cycles

**Annealing**

**2-copies**

**N-copies**
- Cluster
- Polymerase colonies (Polonies)
A population of DNA colonies
Getting single-stranded colonies

Reverse strand cleavage
First-end sequencing

Primer annealing

Synthesis/extension
Record color at each step

This is a parallelized process!
Barcode analysis

Release of the read

Read the barcode
(for subsequent de-multiplexing)

Release of the read
Paired-end sequencing

Bridged-annealing

Reverse-strand synthesis

Cleavage of the forward strand
Paired-end sequencing

Bridged-annealing

Parallel sequencing

Denaturation
Paired-end sequencing: sequence both ends of a fragment
- Facilitate alignment
- Facilitate gene fusion detection
- Better to reconstruct transcript model from RNA-seq
Other technologies
The MinION portable sequencer

- Alpha-hemolysine
  - A nanopore from bacteria that causes lysis of red blood cells
- Molecules that enter the nanopore cause characteristic disruption of the current.
- Potentially offers read lengths of tens of kilobases (kb) limited only by the length of DNA molecules presented to it."
- ~1Gb to 2 Gb of sequence per minION.
- Detect DNA modifications.

https://nanoporetech.com/science-technology/how-it-works
Example application of MinION

Real-time, portable genome sequencing for Ebola surveillance.

Nature doi:10.1038/nature16996

Joshua Quick, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Stephan Günther, Miles W. Carroll et al

Abstract

The Ebola virus disease epidemic in West Africa is the largest on record, responsible for over 28,599 cases and more than 11,299 deaths. Genome sequencing in viral outbreaks is desirable to characterize the infectious agent and determine its evolutionary rate. Genome sequencing also allows the identification of signatures of host adaptation, identification and monitoring of diagnostic targets, and characterization of responses to vaccines and treatments. The Ebola virus (EBOV) genome substitution rate in the Makona strain has been estimated at between $0.87 \times 10^{-3}$ and $1.42 \times 10^{-3}$ mutations per site per year. This is equivalent to 16–27 mutations in each genome, meaning that sequences diverge rapidly enough to identify distinct sub-lineages during a prolonged epidemic. Genome sequencing provides a high-resolution view of pathogen evolution and is increasingly sought after for outbreak surveillance. Sequence data may be used to guide control measures, but only if the results are generated quickly enough to inform interventions. Genomic surveillance during the epidemic has been sporadic owing to a lack of local sequencing capacity coupled with practical difficulties transporting samples to remote sequencing facilities. To address this problem, here we devise a genomic surveillance system that utilizes a novel nanopore DNA sequencing instrument. In April 2015 this system was transported in standard airline luggage to Guinea and used for real-time genomic surveillance of the ongoing epidemic. We present sequence data and analysis of 142 EBOV samples collected during the period March to October 2015. We were able to generate results less than 24 h after receiving an Ebola-positive sample, with the sequencing process taking as little as 15–60 min. We show that real-time genomic surveillance is possible in resource-limited settings and can be established rapidly to monitor outbreaks.
And now the Smidgion...

SmidgION: nanopore sensing for use with mobile devices

Using the same core technology as the handheld MinION device, we are now starting to develop an even smaller device.

In early development
Single-molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio).

- Zero-mode waveguides (ZMV)
  - Each ZMW well is several nanometres in diameter
  - The size of each well does not allow for light propagation
  - The fluorophores bound to bases can only be visualized through the glass substrate in the bottom-most portion of the well, a volume in the zeptolitre range
  - Polymerase is fixed to the bottom of the well.
  - dNTP incorporation on each single-molecule template per well is continuously analyzed by a laser and
  - The polymerase cleaves the dNTP-bound fluorophore during incorporation, allowing it to diffuse away
  - High error-rate, high cost per base
Applications of high-throughput sequencing
High-throughput sequencing: so much applications...

http://tinyurl.com/znrb9jc
Merci