Assembly and annotation using Galaxy

Konrad Paszkiewicz
Sequencing Service, University of Exeter, UK.
25th May 2011
Overview

- Why de-novo assembly?
- What is de-novo assembly?
- Types of assemblers
- Annotation
- A toy example in Galaxy
- Future developments
Sequencing - 2007

**PRODUCTION**
- Rooms of equipment
- Subcloning > picking > prepping
- 35 FTEs
- 3-4 weeks

**SEQUENCING**
- 74x Capillary Sequencers
- 10 FTEs
- 15-40 runs per day
- 1-2Mb per instrument per day
- 120Mb total capacity per day
2nd generation sequencing - Today

**PRODUCTION**

1x Cluster Station
1 FTE
1 day

**SEQUENCING**

1x Genome Analyzer
Same FTE as above
1 run per 3-10 days
- 90Gb per instrument per run
Why de-novo assembly?
Why is de-novo sequence assembly useful?

- No reference genome available
- What is the most suitable reference genome? (e.g. species definition problem in bacteria)

– What’s new in a genome?
  • Remapping will not tell you what is new in a genome (e.g. plasmids, novel genes, novel chromosomes)

– What’s really missing from a genome?
  • Remapping may fail to detect homologous regions
What is de-novo assembly?
De-novo sequence assembly

1. Sequence DNA fragments from each end
De-novo Sequence Assembly

1. Sequence DNA fragments from each end
2. Reads aligned to generate contigs
De-novo Sequence Assembly

1. Sequence DNA fragment from each end
2. Reads aligned to generate contigs
De-novo Sequence Assembly

1. Sequence clones from each end
2. Reads aligned to generate contigs
3. Supercontigs derived from paired reads on different contigs
De-novo Sequence Assembly

1. Sequence reads from each end
2. Reads aligned to generate contigs
3. Supercontigs derived from paired reads on different contigs
4. Ordering of contigs is determined
5. Different insert lengths and read lengths can resolve ambiguities
De-novo assembly: It’s not just for genomes.

1. Traditional single homogenous genome assembly
2. Single organism transcriptomes without a reference - Estimates of expression
3. Genomic/transcriptomic assembly of symbionts and metagenomes
Metagenomics
Denovo Sequence Assembly

• Caveats
  ▪ No assembly is perfect
  ▪ Assemblies from 2\textsuperscript{nd} generation tend to be worse in a number of ways than Sanger based-assemblies
    + Easier to generate data
    + Easier to generate lots of assemblies
    - Shorter reads/higher error rates
    - Man/brainpower is more thinly spread
    - Harder to evaluate assemblies
    - Harder to annotate and compare between samples
Types of assemblers
Types of assemblers

- 4 categories, many variations
- Each tends to have its own niche
- Memory and hardware requirements can differ substantially
- Galaxy has support either in-built or via Galaxy Tool-shed for Velvet, MIRA, AbySS, Phrap Newbler
- Typically a parameter scan is need to get the ‘best’ assembly


De novo assembly of short sequence reads
Paszkiewicz, K. Studholme, D.
Briefings in Bioinformatics
August 2010 11(5): 457-472
Annotation
Annotation

Identification of

genes
exons
promoters
signal peptides
regulatory regions
alleles
non-coding RNAs
repeats...

2 broad categories of annotation methodology:
Sequence homology-based (e.g. Blast)
Profile/HMM-based (e.g. PFAM, TMHMM, SignalP)
Annotation

To do this effectively it is often necessary to gather additional data:

e.g.

ChIP-Seq
RNA-seq
Annotation

Exon structure
Transcription start sites

Annotated gene structure
A toy example in Galaxy
Denovo sequencing project

A new beta-proteobacterium which secretes elemental metal

60% GC content

Approximately 8 Mb genome

Aim: Which gene(s) are responsible for translocation?

Method: 1 lane Illumina sequencing

Mass spectrometry
Process

1. Uploading files from Illumina sequencing

2. Filtering reads

3. De-novo assembly

4. Annotation

5. Locating secretion protein using mass-spectrometry information
1. Uploading files
2nd generation sequencing output formats

- Illumina
  - FASTQ (various flavours)
- SoLID/ABI-Life
  - Colourspace FASTA
- Roche 454
  - SFF
- Ion Torrent
  - SFF or FASTQ
Uploading FASTQ files

<table>
<thead>
<tr>
<th>File Format:</th>
<th>Auto-detect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which format? See help below.</td>
<td></td>
</tr>
</tbody>
</table>

**File:**

```
[raid6-storage/projects/54/s_4/Browse...]
```

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

**URL/Text:**

```
http%5C/users%40mypassword.bioby.nuff.ox.ac.uk/572/s_4_1_sequence.txt.nral
http%5C/users%40mypassword.bioby.nuff.ox.ac.uk/572/s_4_2_sequence.txt.nral
```

Here you may specify a list of URLs (one per line) or paste the contents of a file.

**Convert spaces to tabs:**

- [ ] Yes

Use this option if you are entering intervals by hand.

**Genome:**

- [ ] Click to Search or Select

**Execute**

Or (maybe) directly from SRA/ENA
Uploading FASTQ files
2. Filtering reads
All platforms have errors and artefacts

1. Removal of low quality bases
2. Removal of adaptor sequences
3. Platform specific artefacts (e.g. homopolymers)
Illumina artefacts

Sequence-specific error profile of Illumina sequencers

Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C. Linak, Aki Hirai, Hiroki Takahashi, Md. Altaf-Ul-Amin, Naotake Ogasawara and Shigehiko Kanaya

Received February 3, 2011; Revised April 25, 2011; Accepted April 26, 2011

ABSTRACT

We identified the sequence-specific starting positions of consecutive miscalss in the mapping of reads obtained from the Illumina Genome Analyser (GA). Detailed analysis of the miscall pattern indicated that the underlying mechanism involves sequence-specific interference of the base elongation process during sequencing. The two major sequence patterns that trigger this sequence-specific error (SSE) are: (i) inverted repeats and (ii) GGG sequences. We speculate that these sequences favor dephasing by inhibiting single-base platforms [Illumina/ Solexa Genome Analyser (4), Life Technologies/ ABI SOLiD System (5) and Roche 454 Genome Sequencer FLX (6)], the Illumina Genome Analyser (GA) is, at the moment, the most popular choice for the analysis of genomic information (7). The Illumina/ Solexa sequencers are characterized by: (i) solid-phase amplification and (ii) a cyclic reversible termination (CRT) process, also termed sequencing-by-synthesis (SBS) technology (8). The sequencer can generate hundreds of millions of relatively short (30–100 bp) read sequences per run.

The application of data obtained from this NGS technology can be roughly categorized into the following three

Nakamura, K. et al Sequence-specific error profile of Illumina sequencers

Illumina artefacts

1. GC rich regions are under represented
   a. PCR
   b. Sequencing
2. Substitutions more common than insertions
3. GGC/GCC motif is associated with low quality and mismatches
4. Filtering low quality reads exacerbates low coverage of GC regions

Assembly and/or filtering software should account for this technology specific bias but doesn’t yet
Quality controlling workflow
Quality controlling workflow

Running workflow "Fastq quality control"

Step 1: Input dataset
Input FASTQ read 2
1: s_4_1_sequence.txt SMALL

Step 2: Input dataset
Input FASTQ read 1
2: s_4_2_sequence.txt SMALL

Step 3: Fastq Groomer

Step 4: Fastq Groomer

Step 5: Fastq Quality Trimmer

Step 6: Fastq Quality Trimmer

Step 7: Fastq Joiner

Step 8: Fastq Splitter

Step 9: Prepare paired-end reads for use with Velvet (FASTQ)
Quality controlling workflow
Quality visual summaries
3. De-novo Assembly
Assembly workflow
Velvet optimiser for genomic de-novo assembly

- De-bruijn graph assembler
- Runs a selection of k-mer lengths and parameters
- Selects optimum assembly based on contig length and N50 size (adjustable)
- Originally written by Simon Gladman, CSIRO
- Available at the Galaxy Tool Shed
However...

- We need a method of benchmarking the assembly using biological knowledge
- GC value
- Genome size ~ Total number of bp in contigs?
- Fraction of genes fully assembled
  - Measured against closely related genome
- Manual finishing, gap closure only if really necessary
- Most assemblies only need to be ‘good-enough’... whatever that means...
Assembly results
Assembly statistics

```
<table>
<thead>
<tr>
<th>Statistics for contig lengths:</th>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min contig length:</td>
<td></td>
<td>61</td>
</tr>
<tr>
<td>Max contig length:</td>
<td></td>
<td>13,323</td>
</tr>
<tr>
<td>Mean contig length:</td>
<td></td>
<td>1,016.25</td>
</tr>
<tr>
<td>Standard deviation of contig length:</td>
<td></td>
<td>520.71</td>
</tr>
<tr>
<td>Median contig length:</td>
<td></td>
<td>740</td>
</tr>
<tr>
<td>Statistics for numbers of contigs:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of contigs:</td>
<td></td>
<td>17,188</td>
</tr>
<tr>
<td>Number of contigs &gt;1kb:</td>
<td></td>
<td>2,061</td>
</tr>
<tr>
<td>Number of contigs in N50:</td>
<td></td>
<td>3,447</td>
</tr>
<tr>
<td>Statistics for bases in the contigs:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of bases in all contigs:</td>
<td></td>
<td>5,071,502</td>
</tr>
<tr>
<td>Number of bases in contigs &gt;1kb:</td>
<td></td>
<td>3,238,022</td>
</tr>
<tr>
<td>GC Content of contigs:</td>
<td></td>
<td>41.19%</td>
</tr>
<tr>
<td>Simple dinucleotide repeats:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of contigs with over 70% dinucleotide repeats:</td>
<td></td>
<td>0.00% (0 contigs)</td>
</tr>
<tr>
<td>Simple mononucleotide repeats:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of contigs with over 50% mononucleotide repeats:</td>
<td></td>
<td>0.01% (1 contigs)</td>
</tr>
</tbody>
</table>
```

**Statistics**
- 20 lines, 5 comments
- format: tabular, database: ?

**History**
- 155: histogram of contig sizes
- 154: Sorted contigs
- 153: Assembly statistics
  - 29 lines, 5 comments
  - format: tabular, database: ?

**Assembly**
- 209.7 kb
- Data is in pdf format
Assembly statistics
Taxonomy of contigs
4. Annotation
Annotation workflow
Still to be included

- De-novo gene prediction
- EST and other evidence needs to be included
- tRNAs
- RepeatMasker
- Non-coding features
- Other annotation software pipelines
Can we incorporate these?
Can we incorporate these?
Do we want to incorporate these?

If so: locally? Web services?

Is the service sustainable if it becomes really popular?
Denovo sequencing project

A new bacterium which secretes elemental metal

60% GC content

Approximately 8 Mb genome

**Aim:** Which genes(s) are responsible for translocation?
5. Where is the secretory protein?
Mass spectrometry evidence

MTITASQSRTEVVVRSAN
Locate peptide within contigs ORFs using BlastP

MTITASQSRTEVVVRSA....

Contig 204 ORF 17
Check with annotation tools

- SignalP predicts a signal peptide using both NN and HMM
- TMHMM also predicts that the peptide is external
- PFAM reports a DUF (Domain of Unknown Function)
- BlastP NR reports Hypothetical proteins

ORF located and characterised as coding for a novel metal export factor
Summary

- Filtered and formatted raw data
- Assembled a draft 8 Mb genome – no finishing
- Evaluated metrics and taxonomy of contigs
- Called ORFs bacterial codon usage table
- Basic annotation with BlastP against NCBI NR
- PFAM, SignalP, TMHMM
- Identified peptide within contigs
- No hits in PFAM, NCBI NR. Signal peptide present
- Time frame < 1 day
Other assemblers

• Minimus2 (Galaxy wrapper by Edward Kirton)
  – Merge contigs from different assemblies
• MIRA (Galaxy wrapper by Peter Cock, SCRI)
  – Recent upgrades for PacBio and Ion Torrent
• AbySS (Galaxy wrapper by Edward Kirton)
• Newbler (Galaxy wrapper by Edward Kirton)
  – Roche/454 proprietary assembler and remapper
• Phrap (Galaxy wrapper by Edward Kirton)
  – Sanger read assembly
• String Graph Assembler (Jared Simpson, Sanger)
  – Useful for large (> human) genomes with short reads
Available at Galaxy Toolshed

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly</td>
<td>Tools for working with assemblies</td>
<td>9</td>
</tr>
<tr>
<td>Convert Formats</td>
<td>Tools for converting data formats</td>
<td>5</td>
</tr>
<tr>
<td>Data Source</td>
<td>Tools for retrieving data from external data sources</td>
<td>2</td>
</tr>
<tr>
<td>Fasta Manipulation</td>
<td>Tools for manipulating fasta data</td>
<td>8</td>
</tr>
<tr>
<td>Graphics</td>
<td>Tools producing images</td>
<td>2</td>
</tr>
<tr>
<td>Next Gen Mappers</td>
<td>Tools for the analysis and handling of Next Gen sequencing data</td>
<td>11</td>
</tr>
<tr>
<td>Ontology Manipulation</td>
<td>Tools for manipulating ontologies</td>
<td>1</td>
</tr>
<tr>
<td>SAM</td>
<td>Tools for manipulating alignments in the SAM format</td>
<td>2</td>
</tr>
<tr>
<td>Sequence Analysis</td>
<td>Tools for performing Protein and DNA/RNA analysis</td>
<td>14</td>
</tr>
<tr>
<td>SNP Analysis</td>
<td>Tools for single nucleotide polymorphism data such as WGA</td>
<td>1</td>
</tr>
<tr>
<td>Statistics</td>
<td>Tools for generating statistics</td>
<td>2</td>
</tr>
<tr>
<td>Text Manipulation</td>
<td>Tools for manipulating data</td>
<td>5</td>
</tr>
<tr>
<td>Visualization</td>
<td>Tools for visualizing data</td>
<td>2</td>
</tr>
</tbody>
</table>
Other applications
Oases optimiser for de-novo RNA-seq

- Sister program of Velvet
- Runs a selection of kmer lengths
- Combines all results
- Uses these as a scaffold to assemble transcripts at shortest kmer length
Galaxy denovo RNA-seq Pipeline
Future developments
Community to-do/wish list

- Adding tools dedicated to evaluating assembly quality (e.g. Using EST sequences or related sequences)
- Tools to aid in finishing assemblies
- AFG or other assembly-format visualisation
- Collating and formatting annotation (e.g. GFF files)
- Metagenomics/transcriptomics (e.g. MetaVelvet)
- Gene prediction software
- Blast2Go
- Comparison of GO or PFAM terms between samples
- Enabling workflows of workflows
- AMOS tools (Amos validate etc), web-services
Future developments

A single Illumina GAIIx run can produce data for ~ 100 bacterial genomes in less than a week.

Cost: ~10,000 Euro

**Question:** How do we deal with 100s of comparisons between datasets in Galaxy?

Do we want to?
Do we have a choice?
Rapid Pneumococcal Evolution in Response to Clinical Interventions

Nicholas J. Croucher,^1^ Simon R. Harris,^1^ Christophe Fraser,^2^ Michael A. Quail,^1^ John Burton,^1^ Mark van der Linden,^3^ Lesley McGee,^4^ Anne von Gottberg,^5^ Jae Hoon Song,^6^ Kwan Soo Ko,^7^ Bruno Pichon,^8^ Stephen Baker,^9^ Christopher M. Parry,^9^ Lotte M. Lamberts,^10^ Dea Shahinian,^11^ Dylan R. Pillai,^11^ Timothy J. Mitchell,^12^ Gordon Dougan,^1^ Alexander Tomasz,^13^ Keith P. Klugman,^4,5,14^ Julian Parkhill,^1^ William P. Hanage,^2,15^ Stephen D. Bentley^1^*

Epidemiological studies of the naturally transformable bacterial pathogen *Streptococcus pneumoniae* have previously been confounded by high rates of recombination. Sequencing 240 isolates of the PMEN1 (Spain^23F-1) multidrug-resistant lineage enabled base substitutions to be distinguished from polymorphisms arising through horizontal sequence transfer. More than 700 recombinations were detected, with genes encoding major antigens frequently affected. Among these were 10 capsule-switching events, one of which accompanied a population shift as vaccine-escape serotype 19A isolates emerged in the USA after the introduction of the conjugate polysaccharide vaccine. The evolution of resistance to fluoroquinolones, rifampicin, and macrolides was observed to occur on multiple occasions. This study details how genomic plasticity within lineages of recombinogenic bacteria can permit adaptation to clinical interventions over remarkably short time scales.
DNA sequencing generations

<table>
<thead>
<tr>
<th>Then + Now</th>
<th>Now</th>
<th>Now + anticipated</th>
<th>Anticipated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Gen Sanger</td>
<td>2nd Gen -parallelised</td>
<td>3rd Gen -single mol or electronic</td>
<td>Next -single mol AND electronic</td>
</tr>
</tbody>
</table>

- Low throughput
- High cost
- Accurate
- Broad user base
- Optical
- Amplification needed
- Highly parallel
- Improved cost and throughput
- New applications
- Or electronic, clonal

- Sanger
- GAI (Solexa/Illumina)
- SOLiD (ABI/LIFE)
- 454 FLX (454/Roche)
- Helicos
- Pacific Biosciences
- Ion Torrent (LIFE Starlight)
- Nanopores

Estimated cost of a human genome using these technologies

- $70M
- $200k --- $50k ---- $20k --- 15k---
- ?$5k - $1k?
Questions?

Konrad Paszkiewicz

k.h.paszkiewicz@exeter.ac.uk
“We need to start thinking about how to train people, both health-care professionals and scientists, to be facile in bioinformatics. We need to foster development of professionals who have expertise analyzing large data sets of the size that biologists haven't had to think about. We need to entice smart people into genomics.”

Eric Green,

Director National Human Genome Research Institute
Acknowledgements

University of Exeter

• Murray Grant
• Karen Moore
• Alex Moorhouse

Peter Cock – James Hutton Institute
Galaxy team and tool shed contributors