HT-Sequence Data Analysis: Challenges and Solutions

IIGB Illumina Workshop

Thomas Girke

October 19, 2009
Introduction

Data Analysis
  Overview
  Example: Short Read Alignments

Project Overview
  Project Types
  Example of a New Application

IIGB Bioinformatics Infrastructure
  Hardware & Software
  HTS Bioinformatics Services
  Online Resources

References
Outline

Introduction

Data Analysis
  Overview
  Example: Short Read Alignments

Project Overview
  Project Types
  Example of a New Application

IIGB Bioinformatics Infrastructure
  Hardware & Software
  HTS Bioinformatics Services
  Online Resources

References
### What Has Changed?

<table>
<thead>
<tr>
<th></th>
<th>Sanger</th>
<th>Illumina</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read length in bp</td>
<td>600-800 bp</td>
<td>36-2×76 (2×152) bp</td>
</tr>
<tr>
<td>Gb per run</td>
<td>&lt;0.001 Gb</td>
<td>&gt;10 Gb</td>
</tr>
<tr>
<td>Human genome (3Gb)</td>
<td>5 years; $1 billion</td>
<td>1 week; $3000</td>
</tr>
<tr>
<td>Arabidopsis genome (0.125Gb)</td>
<td>4 years; $70 million</td>
<td>$100</td>
</tr>
<tr>
<td>Yeast genome (0.012Gb)</td>
<td>3 years; $10 million</td>
<td>$10</td>
</tr>
</tbody>
</table>

⇒ Sequence volume has changed by a factor of $>10,000$. 
Many classical molecular and genetics approaches have now HTS solutions.

Microarray-based approaches are being replaced by HTS.

New applications are being invented every day.
Opportunities: Application Spectrum of HTS

- **SNP-Seq**: Genome-wide detection of SNPs and mutations
- **BS-Seq**: Methylome profiling by bisulphite sequencing
- **ChIP-Seq**: DNA-protein interactions
- **RNA-Seq**: Transcriptome profiling
- **DGE**: mRNA expression profiling
- **small RNA-Seq**: Small RNA profiling and discovery
- ...
Bioinformatics Challenges

Data Management
- What: data types to store?
- How: data formats?
- Where: databases?

Computer Hardware
- Data storage systems that scale to several hundred TB
- Hundreds CPUs for running slower algorithms
- Extensive memory: e.g. assembly problems

Algorithms
- Time efficiency and utility

Analysis Pipelines
- Very difficult to design for all analysis needs
Skills Required for Data Analysis

- Most analysis steps cannot be performed on desktop or laptop computers anymore.
- Much more computational experience required than in pre-HTS times.
- Important skills:
  - Training in sequence analysis algorithms and statistics
  - Considerable programming experience
  - Proficiency with Linux/Unix environments
- How to Create Required Knowledge Base
  - Recruit bioinformatics postdocs
  - Collaborate with CS and statistics labs
    - Priority: biological findings ↔ advancement of methods
  - Educate: new generation of graduate students
  - Core facility model: staff bioinformatics programmers
Outline

Introduction

Data Analysis

Overview
Example: Short Read Alignments

Project Overview
Project Types
Example of a New Application

IIGB Bioinformatics Infrastructure
Hardware & Software
HTS Bioinformatics Services
Online Resources

References
Outline

Introduction

Data Analysis

Overview

Example: Short Read Alignments

Project Overview

Project Types

Example of a New Application

IIGB Bioinformatics Infrastructure

Hardware & Software

HTS Bioinformatics Services

Online Resources

References
Main Analysis Steps

- Raw Data Analysis
- Quality Control
- Application Specific Analysis
- Data Visualization
<table>
<thead>
<tr>
<th>Type</th>
<th>Software Packages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Raw Data Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Image Analysis</td>
<td>Firecrest (Illumina)</td>
</tr>
<tr>
<td>Base Calling</td>
<td>Bustard (Illumina), Alta-Cyclic, Rolexa</td>
</tr>
<tr>
<td><strong>Quality Control</strong></td>
<td></td>
</tr>
<tr>
<td>Instrument QC</td>
<td>QC (Illumina), ShortRead (Bioconductor)</td>
</tr>
<tr>
<td>Sequence Trimming</td>
<td>Bio* Tools, Biostrings (Bioconductor)</td>
</tr>
<tr>
<td>Alignment</td>
<td>Eland (Illumina), Soap, Maq, Bowtie, etc.</td>
</tr>
<tr>
<td><strong>Application Specific Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>SNP-Seq</td>
<td>MAQ, Bowtie, etc.</td>
</tr>
<tr>
<td>RNA-Seq/DGE</td>
<td>ERANGE, TopHat, edgeR (Bioconductor)</td>
</tr>
<tr>
<td>ChIP-Seq, BS-Seq</td>
<td>ChIP-Seq (Bioconductor)</td>
</tr>
<tr>
<td>small RNA-Seq</td>
<td>Various pipelines, custom tools</td>
</tr>
<tr>
<td><em>De novo</em> Assembly</td>
<td>Velvet, ABySS, Edena, etc.</td>
</tr>
<tr>
<td>Multipurpose Tools</td>
<td>Bioconductor, BioPerl, BioPython</td>
</tr>
<tr>
<td><strong>Data Visualization</strong></td>
<td></td>
</tr>
<tr>
<td>Generic scientific graphics</td>
<td>R, Matlab, etc.</td>
</tr>
<tr>
<td>Genome Browsers</td>
<td>G/JBrowse, AnnoJ, UCSC, rtracklayer, etc.</td>
</tr>
</tbody>
</table>
Outline

Introduction

Data Analysis
  Overview
  Example: Short Read Alignments

Project Overview
  Project Types
  Example of a New Application

IIGB Bioinformatics Infrastructure
  Hardware & Software
  HTS Bioinformatics Services
  Online Resources

References
Why Can’t We Use BLAST Anymore?

It is far too slow

- To BLAST over 10 million reads against the human genome, can take several months.

It is the wrong algorithm for imperfect matching of short reads

- Example: BLASTN cannot find an alignment like this one with its word size limit of 7 nucleotides:

```
Query: ATCTACTACTACTAGATCGAGCGTACGTGTTGACACACTATCTAC

Subject: ATCTACCACTACTGAGATCGTGCGTACATGTTGAAACACTAGCTAC
```

BLAST output is very IO and storage intensive

- All we need are match/mismatch positions in genomes
Performance Requirements of Alignment Software

Aligning tens of millions of sequences requires:

- Ultra fast search algorithms (100-1000x faster than BLAST)
- Small memory footprint
- Economic data structures and containers

Alignment requirements

- The requirements for short-read mapping applications are very different from traditional sequence database search approaches for ortholog identification.
- Lower scoring alternative hits (more mismatches) are less interesting.
- Often only perfect matching required, but with the possibility to allow 1-2 mismatches and only sometimes very short gaps.
- Most short-read alignment algorithms will not work for longer sequences!
- However, most of them are more sensitive for short-reads than BLAST, because they lack its word size limitation.
- Often they can only align sequences of constant length.
Solution: Indexing and Hashing Approaches

Indexing and searching used by MAQ

Ref4  TACGCGAT  2 continuous mismatches
Ref3  AACCGGAT  2 spaced mismatches
Ref2  AACGCGAT  1 mismatch
Ref1  ATCGCGAT  perfect match
Read  ATCGCGAT  Ref1 Ref2 Ref3 Ref4
TMPA  11110000  Y  N  N  N
TMPA  00001111  Y  Y  N  Y
TMPB  11000111  Y  N  N  N
TMPB  00111000  Y  Y  N  Y
TMPB  11001100  Y  N  N  N
TMPC  00110011  Y  Y  Y  Y
TMPC  00110011  Y  Y  Y  Y

Fig. 1: Indexing and Search Strategy of MAQ Algorithm
Bowtie

Algorithm [Langmead et al 2009]
- Burrows-Wheeler index based on the full-text minute-space (FM) index.

Performance
- Aligns sequences of 4-1,024 bases.
- Handles sequences of variable length in a single run.
- Index requires only 1.3GB memory for the human genome (works on laptop!).
- Aligns 25 million 35bp reads per CPU-hour.
- Fastest of all short-read alignment programs: 35 faster than MAQ and 3 times faster than SOAP.
- With default settings, sensitivity similar to SOAP, but slightly lower than MAQ.

Limitations
- Requires BWT index, which takes several hours to compute.
- It reports inexact matches, but does not guarantee to find the match with the highest quality alignment.
- With its highest performance settings, it may fail to align a small number of reads with multiple mismatches.
- Increased accuracy options can overcome some of these limitation at the cost of speed performance.
(a) The Burrows-Wheeler matrix and transformation for 'acaacg'. Data compression is facilitated by forming stretches of the same characters.

(b) An unpermute step repeatedly applies the last first (LF) mapping to recover the original text (in red on the top line) from the Burrows-Wheeler transform (in black in the rightmost column).

(c) Steps taken by an exact search to identify the range of rows, and thus the set of reference suffixes, prefixed by 'aac'.
Bowtie: Searching for Inexact Matches

- If the range of an exact search becomes empty, then the algorithm selects an already-matched query position and substitutes its base by different ones (mismatch). Then the exact search resumes from the substituted position.
- If there are multiple candidate substitution positions, then the algorithm greedily selects a position with a minimal quality value.

![Diagram of exact and inexact matches](image)

**Fig. 2**: Inexact match for query 'GGTA'

IIGB Illumina Workshop  
Data Analysis  
Example: Short Read Alignments
Very Incomplete List of Short-Read Alignment Tools

- Indexing Reference with Suffix Array/Burrows-Wheeler
  - Bowtie [Langmead et al 2009]
  - SOAPv2
- Indexing Reads with Hash Tables
  - ZOOM: uses spaced seeds algorithm
  - RMAP: simpler spaced seeds algorithm [Smith et al 2008]
  - SHRiMP: employs a combination of spaced seeds and the Smith-Waterman
  - MAQ [Li et al 2008b]
  - Eland (Illumina)
- Indexing Reference with Hash Tables
  - SOAPv1 [Li et al 2008]
  - Mosaik
- Merge Sorting
  - Slider [Malhis et al 2009]
- Almost Complete Link Collection by Heng Li
HTS Data Depository

Short Read Archive (SRA) from NCBI
  • URL: http://www.ncbi.nlm.nih.gov/Traces/sra
Outline

Introduction

Data Analysis
  Overview
  Example: Short Read Alignments

Project Overview
  Project Types
  Example of a New Application

IIGB Bioinformatics Infrastructure
  Hardware & Software
  HTS Bioinformatics Services
  Online Resources

References
Outline

Introduction

Data Analysis
  Overview
  Example: Short Read Alignments

Project Overview
  Project Types
  Example of a New Application

IIGB Bioinformatics Infrastructure
  Hardware & Software
  HTS Bioinformatics Services
  Online Resources

References
HTS Projects at IIGB

- RNA-Seq
- new genome-Seq
- small RNA-Seq
- BS-Seq
- Knockout-Seq
- ChIP-Seq
- SNP-Seq
Outline

Introduction

Data Analysis
  Overview
  Example: Short Read Alignments

Project Overview
  Project Types
  Example of a New Application

IIGB Bioinformatics Infrastructure
  Hardware & Software
  HTS Bioinformatics Services
  Online Resources

References
Example: RNA-Seq of Unsequenced Genomes

- Read length improvements beyond 50bp made it possible.
- Typical strategy:
  1. Long PE (2x 76bp) sequencing, followed by assembly of reference transcriptome
  2. Annotate reference transcriptome
  3. Quantitative RNA-Seq of many samples by aligning against reference transcriptome
- Project Examples
  - Spider transcriptome; gene discovery objective; PI: Cheryl Hayashi, UC Riverside
  - Aphid transcriptome; mRNA profiling objective; PI: Isgouhi Kaloshian, UC Riverside
  - Fraser fir transcriptome; breeding marker discovery; PI: Ross Whetten, NC State University
RNA-Seq versus DGE

RNA-seq

mRNA ➔ AAAA
Fragment ➔ □□□□□□
Library ➔ □□□□□□

Sequencing ➔
1. Alternative splicing
2. Limited expression profiling
3. SNP detection
4. Many other applications

DGE

mRNA ➔ AAAA
Fragment ➔ □□□□□□
Library ➔ □□□□□□

Sequencing ➔
1. Expression profiling ➔ more appropriate for many biosamples

NlaIII/DpnII
TAG

Example: Assembly of Potato Aphid Transcriptome

- Initial data based on *de novo* assembly with Velvet
- Assembly pipeline: *de novo* → reference assisted → gene boosted
- Challenges:
  - Assembly of >100 million reads can require >100GB of RAM
  - Abundance differences of mRNA species

<table>
<thead>
<tr>
<th>Contig Size</th>
<th>Distinct Contigs</th>
</tr>
</thead>
<tbody>
<tr>
<td>200-500bp</td>
<td>13,328</td>
</tr>
<tr>
<td>500-1000bp</td>
<td>6,249</td>
</tr>
<tr>
<td>&gt;1000bp</td>
<td>4,666</td>
</tr>
</tbody>
</table>
Outline

Introduction

Data Analysis
  Overview
  Example: Short Read Alignments

Project Overview
  Project Types
  Example of a New Application

IIGB Bioinformatics Infrastructure
  Hardware & Software
  HTS Bioinformatics Services
  Online Resources

References
Outline

Introduction

Data Analysis
   Overview
   Example: Short Read Alignments

Project Overview
   Project Types
   Example of a New Application

IIGB Bioinformatics Infrastructure
   Hardware & Software
   HTS Bioinformatics Services
   Online Resources

References
High-Performance Compute Infrastructure at IIGB

Head node: 16 CPU cores
Linux cluster nodes: 256 CPU cores + 64 CPU cores
High memory node: 64GB (192GB) RAM
SAN high-availability storage: 30TB of 15k rpm disks
Long-term storage: 40TB
Server Room: 500 sqft, 40 Ton AC, UPS, backup generator

Full access: $1000 annual registration fee for an entire lab!
High-Performance Compute Infrastructure at IIGB

- Head node: 16 CPU cores
High-Performance Compute Infrastructure at IIGB

- Head node: 16 CPU cores
- Linux cluster nodes: 256 CPU cores + 64 CPU cores

- High memory node: 64GB (192GB) RAM
- SAN high-availability storage: 30TB of 15k rpm disks
- Long-term storage: 40TB
- Server Room: 500 sqft, 40 Ton AC, UPS, backup generator

Full access: $1000 annual registration fee for an entire lab!
High-Performance Compute Infrastructure at IIGB

- Head node: 16 CPU cores
- Linux cluster nodes: 256 CPU cores + 64 CPU cores
- High memory node: 64GB (192GB) RAM
High-Performance Compute Infrastructure at IIGB

- Head node: 16 CPU cores
- Linux cluster nodes: 256 CPU cores + 64 CPU cores
- High memory node: 64GB (192GB) RAM
- SAN high-availability storage: 30TB of 15k rpm disks
High-Performance Compute Infrastructure at IIGB

- Head node: 16 CPU cores
- Linux cluster nodes: 256 CPU cores + 64 CPU cores
- High memory node: 64GB (192GB) RAM
- SAN high-availability storage: 30TB of 15k rpm disks
- Long-term storage: 40TB
High-Performance Compute Infrastructure at IIGB

- Head node: 16 CPU cores
- Linux cluster nodes: 256 CPU cores + 64 CPU cores
- High memory node: 64GB (192GB) RAM
- SAN high-availability storage: 30TB of 15k rpm disks
- Long-term storage: 40TB
- Server Room: 500 sqft, 40 Ton AC, UPS, backup generator
High-Performance Compute Infrastructure at IIGB

- Head node: 16 CPU cores
- Linux cluster nodes: 256 CPU cores + 64 CPU cores
- High memory node: 64GB (192GB) RAM
- SAN high-availability storage: 30TB of 15k rpm disks
- Long-term storage: 40TB
- Server Room: 500 sqft, 40 Ton AC, UPS, backup generator

Full access: $1000 annual registration fee for an entire lab!
Software

- Over 500 software tools
- Mostly Linux-based open-access and open-source tools
- Advantages: remote access, multitasking, multiuser, scriptable, modifiable, up-to-date, free, etc
- Commercial software only where required
Outline

Introduction

Data Analysis
   Overview
   Example: Short Read Alignments

Project Overview
   Project Types
   Example of a New Application

IIGB Bioinformatics Infrastructure
   Hardware & Software
   HTS Bioinformatics Services
   Online Resources

References
Bioinformatics Service Team

- Rebecca Sun - Bioinformatics Data Analyst
- Tyler Backman - Bioinformatics Data Analyst
- Thomas Girke - The one to blame...
Service Overview

Primary data analysis
- QC of sequencing run
- Quality filtering and adaptor trimming
- Basic mapping to reference genome
- Access to externally and internally developed analysis pipelines and packages

Available upon request
- Online genome browser support
- Web BLAST server support

Custom data analysis
- Comprehensive custom analysis: annotations, mapping, predictions etc.
- Specialty algorithms for base calling, alignment, etc.
- Detailed data analysis reports
- Statistical analysis
- Any custom request if reasonable
Bioinformatics Service Rates

General charges

- Sequence/quality control: no cost
- Sequencing run storage on tapes: included in sequencing run
- Online genome browser support: $1000 annually

Rates for custom data analysis

- 1 week projects: $24/hr
- 1-2 week project: $18/hr
- Projects >2 weeks: $12/hr
Outline

Introduction

Data Analysis
  Overview
  Example: Short Read Alignments

Project Overview
  Project Types
  Example of a New Application

IIGB Bioinformatics Infrastructure
  Hardware & Software
  HTS Bioinformatics Services
  Online Resources

References
IIGB HTS Web Site

- Overview: technology, protocols, links, etc.
- Sample submission
- Data download
Anno-J Genome Browser

- Excellent performance for large datasets
GBrowse Genome Browser

- Flexible visualization options
Outline

Introduction

Data Analysis
  Overview
  Example: Short Read Alignments

Project Overview
  Project Types
  Example of a New Application

IIGB Bioinformatics Infrastructure
  Hardware & Software
  HTS Bioinformatics Services
  Online Resources

References
URL http://www.hubmed.org/display.cgi?uids=18519653

Langmead, B, Trapnell, C, Pop, M, Salzberg, S L (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome Genome Biol, 10:.
URL http://www.hubmed.org/display.cgi?uids=19261174

URL http://www.hubmed.org/display.cgi?uids=18227114

URL http://www.hubmed.org/display.cgi?uids=18714091

URL http://www.hubmed.org/display.cgi?uids=18974170

URL http://www.hubmed.org/display.cgi?uids=18307793