Analysis of High-Throughput Sequencing Data with R and Bioconductor

Workshop Introduction

Tyler Backman, Rebecca Sun & Thomas Girke

October 29, 2010
Traditional DNA Sequencing Technologies
   Chemical Sequencing
   Sanger Sequencing

High-Throughput Sequencing Methods
   Solexa/Illumina: Reversible Terminator Method
   Helicos: Single Molecule Sequencing
   454/Roche: Pyrosequencing Method
   SOLiD/ABI: Supported Oligo Ligation Method

Research Applications

Examples of Short Read Alignment Algorithms

Sequence Processing and Analysis

Hands-on Manual

References and Books
Outline

Traditional DNA Sequencing Technologies
  Chemical Sequencing
  Sanger Sequencing

High-Throughput Sequencing Methods
  Solexa/Illumina: Reversible Terminator Method
  Helicos: Single Molecule Sequencing
  454/Roche: Pyrosequencing Method
  SOLiD/ABI: Supported Oligo Ligation Method

Research Applications
  Examples of Short Read Alignment Algorithms
  Sequence Processing and Analysis
  Hands-on Manual
  References and Books
History of DNA Sequencing

1977 "DNA Sequencing by Chemical Degradation" is published by Allan Maxam and Walter Gilbert.

1977 "DNA Sequencing by Enzymatic Synthesis" is published by Fred Sanger.

1980 Fred Sanger and Walter Gilbert receive the Nobel Prize in Chemistry.

1982 GenBank starts as a public repository of DNA sequences.

1986 Leroy Hood’s laboratory at the California Institute of Technology announces the first semi-automated DNA sequencing machine.

1997 Genome sequence of E. coli is published.

2001 Draft sequence of the Human genome is published.

2004 Next generation sequencing technologies become available to the public.
Outline

Traditional DNA Sequencing Technologies

Chemical Sequencing
Sanger Sequencing

High-Throughput Sequencing Methods
Solexa/Illumina: Reversible Terminator Method
Helicos: Single Molecule Sequencing
454/Roche: Pyrosequencing Method
SOLiD/ABI: Supported Oligo Ligation Method

Research Applications

Examples of Short Read Alignment Algorithms
Sequence Processing and Analysis
Hands-on Manual
References and Books
Chemical Sequencing by Maxam & Gilbert

1. Uses radioactive labeled DNA fragments of 500 bp.
2. Four separate chemical treatments generate DNA breaks at the positions: G, A+G, C, C+T.
3. The fragments are size-separated by gel electrophoresis in four separate lanes.
4. Visualization of the fragments by autoradiography on an X-ray film.

![Diagram showing the process of chemical DNA degradation and gel electrophoresis.](image)

**Chemical DNA Degradation**

**Gel Electrophoresis**
Traditional DNA Sequencing Technologies

Chemical Sequencing

Sanger Sequencing

High-Throughput Sequencing Methods

Solexa/Illumina: Reversible Terminator Method

Helicos: Single Molecule Sequencing

454/Roche: Pyrosequencing Method

SOLiD/ABI: Supported Oligo Ligation Method

Research Applications

Examples of Short Read Alignment Algorithms

Sequence Processing and Analysis

Hands-on Manual

References and Books
Illustration of Sanger Sequencing

Sequencing Principle

Radioactive Fluorescence Labeling
Assign quality score to each peak

The frequently used Phred scores provide log(10)-transformed error probability values:

- score = 20 corresponds to a 1% error rate
- score = 30 corresponds to a 0.1% error rate
- score = 40 corresponds to a 0.01% error rate

The base calling (A, T, G or C) is performed based on Phred scores.

Ambiguous positions with Phred scores ≤ 20 are labeled with N.
Outline

Traditional DNA Sequencing Technologies
  Chemical Sequencing
  Sanger Sequencing

High-Throughput Sequencing Methods
  Solexa/Illumina: Reversible Terminator Method
  Helicos: Single Molecule Sequencing
  454/Roche: Pyrosequencing Method
  SOLiD/ABI: Supported Oligo Ligation Method

Research Applications

Examples of Short Read Alignment Algorithms

Sequence Processing and Analysis

Hands-on Manual

References and Books
Common Synonyms

- High-throughput sequencing: HTS or HT-Seq
- Flow cell sequencing (FCS)
- Massively parallel sequencing (MPS)
- Next/this generation sequencing (NGS/TGS)
- Deep sequencing
- Sequencing by synthesis
- Many other synonyms

Review article: [Holt et al 2008]
Overview: 454, SOLiD and Illumina

From review article: [Medini et al 2008]
Similarities and Differences of HT-Seq Technologies

Common components

- Flow cells as reaction chambers
- Iterative sequencing process
- Massive parallelization
- Clonally amplified or single molecule templates

Differences

- Template preparation
- Sequencing chemistry
- Flow cell configuration
HT Sequencing Methods

Reversible Terminator Methods (e.g. Illumina/Solexa)
- Use reversible versions of dye-terminator reactions.
- Principle steps: adding one nucleotide at a time, detecting fluorescence corresponding to that position, then removing the blocking group to allow the polymerization of another nucleotide.

Single Molecule Methods (e.g. Helicos)
- Sequences one of the four nucleotides per cycle.

Pyrosequencing Methods (e.g. 454)
- Also use DNA polymerization to add nucleotides.
- Principle steps: adding one type of nucleotide at a time, then detecting and quantifying the number of nucleotides added to a given location through the light emitted by the release of attached pyrophosphates.

Supported Oligonucleotide Ligation Methods (e.g. SOLiD, Complete Genomics)
- Uses ligation-based approach
- Principle steps: stepwise ligation of labeled random octamers to obtain sequence of attached dinucleotides; the ligated dinucleotides of each ligation round are spaced by several nucleotides; continuous sequence information is obtained by offsetting the sequencing primer.
Traditional DNA Sequencing Technologies
   Chemical Sequencing
   Sanger Sequencing

High-Throughput Sequencing Methods
   Solexa/Illumina: Reversible Terminator Method
   Helicos: Single Molecule Sequencing
   454/Roche: Pyrosequencing Method
   SOLiD/ABI: Supported Oligo Ligation Method

Research Applications
   Examples of Short Read Alignment Algorithms
   Sequence Processing and Analysis
   Hands-on Manual

References and Books
Example: Illumina/Solexa Technology

Illumina Sequencer

Flow Cell
Basic Steps of Illumina/Solexa Sequencing Technology

Compare with illustration on next three slides!

Flow Cell Loading

1. Generate DNA library (genomic- or cDNA-based) with insert length of \( \sim 200 \) bp.
2. Load library onto flow cell (nano device for liquid handling).
3. PCR-based bridge amplification of loaded fragments to obtain DNA clusters (serves signal amplification)

Sequencing Cycles

4. Start reversible dye-terminator reaction containing primer and labeled dNTPs among other components.
5. Image scan to detect the identity of first base of each cluster via the characteristic fluorescence signal for each labeled nucleotide.
6. De-protection step removes the blocking group and fluorescence group of the incorporated nucleotide.
7. Repeat steps 4-6 about 30-60 times.
Loading of Flow Cell

**DNA Fragments With Adaptors**

**Bind DNA to Surface of Flow Cell**

**Bridge Amplification**

**X Amplification Rounds**
10 Mio molecule clusters each diameter of 1um
1000 molecules per cluster
Details of Sequencing Reaction

Sequencing Reaction


(A): Polymerase attaches labeled and blocked nucleotides

(B): Deblocking and removal of label

(A) (B) (A) (B) (A) (B) (A) (B)

Data Collection

Legend

Primer
Adaptors
Dyes & Blocking Agent
Signals on Images

Illustration shows the sequencing cycles for a single template molecule!
Single End, Paired End and Mate Pair Sequencing

**Single End**

- Genomic DNA
- Fragment (200–500 bp)
- Ligate Adaptors
- Generate Clusters
- Sequence

**Paired End**

- Genomic DNA
- Fragment (200–500 bp)
- Ligate Adaptors
- Generate Clusters
- Sequence First End
- Regenerate Clusters and Sequence Paired End

**Mate Pair**

- Genomic DNA
- Fragments (2–5 kb)
- Circularize
- Enrich Bisulfate Ends
- Ligate Adaptors
- Generate Clusters
- Sequence First End
- Regenerate Clusters and Sequence Paired End

*AP1/AP2: flow cell adaptors; SP1/SP2: sequencing primers*
Paired End Chemistry: Step I

Grafted Flow Cell

Single End
Periodate

Paired End
Two different enzymes

Linearization:

Cluster Generation: Initial Extension

Slide 21/63
Paired End Chemistry: Step II

Cluster Generation: Amplification

1\textsuperscript{st} cycle denaturation → 1\textsuperscript{st} cycle annealing → 1\textsuperscript{st} cycle extension → 2\textsuperscript{nd} cycle denaturation

\[ n=35 \text{ total} \]

2\textsuperscript{nd} cycle extension → 2\textsuperscript{nd} cycle annealing
Paired End Chemistry: Step III

Cluster Generation: Linearization

Cluster Amplification → P5 Linearization → Block with ddNTPS → Denaturation and Sequencing Primer Hybridization
Paired End Chemistry: Step IV

Sequencing

- Denaturation and Hybridization
- Sequencing First Read
- Denaturation and De-Protection
- Resynthesis of P5 Strand
- Sequencing Second Read
- Denaturation and Hybridization
- Block with ddNTPs
- P7 Linearization
Processing of Illumina Sequencing Data

- Convert cluster images to intensity values.
- Base calling based on intensity for each fluorescence dye.
- Generate quality scores similar to Phred scores.
- The length of each sequence corresponds to the number of cycles, e.g. 36 cycles → 36 bp.
- Remove sequences with low quality reads.

In 2008 a single sequencing run with 36 cycles could generate ~ 1.5 billion bp of sequence information and ~ 130,000 images or 1TB of image data.
Outline

Traditional DNA Sequencing Technologies
   Chemical Sequencing
   Sanger Sequencing

High-Throughput Sequencing Methods
   Solexa/Illumina: Reversible Terminator Method
   Helicos: Single Molecule Sequencing
   454/Roche: Pyrosequencing Method
   SOLiD/ABI: Supported Oligo Ligation Method

Research Applications

Examples of Short Read Alignment Algorithms

Sequence Processing and Analysis

Hands-on Manual

References and Books
Helicos: Single Molecule Sequencing

- Has similarities to Solexa/Illumina technology, but sequences single molecule templates.
- Attaches one of the four nucleotides at a time using proprietary nucleotide-polymerase formulations. This prevents the incorporation of more than one nucleotide in each cycle in homopolymer regions.
Outline

Traditional DNA Sequencing Technologies
  Chemical Sequencing
  Sanger Sequencing

High-Throughput Sequencing Methods
  Solexa/Illumina: Reversible Terminator Method
  Helicos: Single Molecule Sequencing
  454/Roche: Pyrosequencing Method
  SOLiD/ABI: Supported Oligo Ligation Method

Research Applications

Examples of Short Read Alignment Algorithms

Sequence Processing and Analysis

Hands-on Manual

References and Books
Pyrosequencing Methods (*e.g.* 454)

- Also uses DNA polymerization to add nucleotides.
- Principle steps: adding one type of nucleotide at a time, then detecting and quantifying the number of nucleotides added to a given location through the light emitted by the release of attached pyrophosphates.
- For more details see: [454 Web Site](http://www.454.com)
454 Pyrosequencing

Sample preparation

Pyrosequencing

454 technology

gDNA -> Adaptors -> Beads -> PCR

Enzymes on beads and primer

Polymerase

T/AT/G/C

ATP

Light

PPi
SOLiD: Sequencing by Supported Oligonucleotide Ligation and Detection

1. Prime and Ligate
   - PRIMER ROUND 1
   - Universal seq primer (a)
   - Primer Adder
   - Template Sequence
   - Cleavage Agent
   - Fluorescence

2. Image
   - Excite
   - TA

3. Cap Unextended Strands

4. Cleave off Fluor

5. Repeat steps 1-4 to Extend Sequence
   - Ligation cycle

6. Primer Reset
   - Universal seq primer (n-1)
   - Primer Reset

7. Repeat steps 1-5 with new primer

8. Repeat Reset with n-2, n-3, n-4 primers

Workshop Introduction  High-Throughput Sequencing Methods  SOLiD/ABI: Supported Oligo Ligation Method
### Comparison of HT-Sequencing Methods

**Technology in 2008!!**

<table>
<thead>
<tr>
<th>Method</th>
<th>Illumina</th>
<th>454</th>
<th>SOLiD</th>
<th>Helicos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read Length</td>
<td>36-2x100</td>
<td>300-400</td>
<td>36</td>
<td>25-45</td>
</tr>
<tr>
<td>Error Rate</td>
<td>~1%</td>
<td>&gt;1%</td>
<td>~0.1%</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>Data/Run (Gb)</td>
<td>1-3</td>
<td>0.1</td>
<td>2-3</td>
<td>8</td>
</tr>
<tr>
<td>Cost (per Gb)</td>
<td>$6,000</td>
<td>$84,000</td>
<td>$6,000</td>
<td>$2,500</td>
</tr>
</tbody>
</table>

Data largely from [Holt et al 2008]
## Comparison with Traditional Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Read Length</th>
<th>Sequences per Run</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dye-Terminator</td>
<td>500-1500 bp</td>
<td>384</td>
<td><em>de novo</em> and low-throughput applications</td>
</tr>
<tr>
<td>(Sanger)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>454/Roche</td>
<td>120-400 bp</td>
<td>~200,000</td>
<td><em>de novo</em> and medium-throughput applications</td>
</tr>
<tr>
<td>Illumina/Solexa</td>
<td>36-60 bp</td>
<td>~20,000,000</td>
<td>high-throughput applications</td>
</tr>
</tbody>
</table>

- *de novo* applications do not require guide genome sequence, *e.g.* new genomes, etc.
- high/medium-throughput applications require guide genome sequence, *e.g.* SNPs, small RNAs, etc.

All numbers are estimates and apply to the situation in Feb. 2009!
Outline

Traditional DNA Sequencing Technologies
  Chemical Sequencing
  Sanger Sequencing

High-Throughput Sequencing Methods
  Solexa/Illumina: Reversible Terminator Method
  Helicos: Single Molecule Sequencing
  454/Roche: Pyrosequencing Method
  SOLiD/ABI: Supported Oligo Ligation Method

Research Applications

Examples of Short Read Alignment Algorithms
Sequence Processing and Analysis
Hands-on Manual
References and Books
HT-Seq technologies provide endless opportunities for genomics, comparative genome biology, medical diagnostics, etc. The following examples provide only a brief overview.

**Applications**

- Genome-wide detection of SNPs and mutations (SNP-seq)
- Methylome profiling by bisulphite sequencing (BS-seq)
- DNA-protein interactions (ChIP-seq)
- Transcriptome sequencing (RNA-seq)
- mRNA expression profiling (DGE)
- Small RNA profiling and discovery
Application: DNA-Protein Interactions with ChIP-Seq

Reference for ChIP-Seq data analysis: [Jothi et al 2008]
Application: Methylome Profiling with BS-Seq

Genomic DNA

5' C C C G G G 3'

Treatment with bisulfite

5' U U mC U 3'

PCR (1st cycle)

5' U U mC U A A G A 3'

PCR (amplification)

5' T T C A A G A 3'

Sequencing
Application: Digital Gene Expression (DGE) Profiling

Illustration for Illumina’s DGE

- First and Second Strand cDNA Synthesis
- Restriction Enzyme Digestion: Nla III or Dpn II
- Ligation of 5' Adaptor
- MmeI digestion
- Ligation of 3' Primer/Adaptor
- PCR Amplification via Adaptors
- Sequencing Primer Binding Site

HT-Seq of Tags
RNA-Seq versus DGE

**RNA-seq**

- mRNA
- Fragment Library

**DGE**

- mRNA
- Fragment Library

**Sequencing**

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

1. Expression profiling
   ➔ more appropriate for many biosamples
Targeted Sequencing Using DNA Capture Microarrays

- Powerful approach to make HT-Seq more economic and versatile.

Example: usage of programmable microarrays (here NimbleGen) to enrich for DNA regions of interest [Albert et al 2007].
Outline

Traditional DNA Sequencing Technologies
   Chemical Sequencing
   Sanger Sequencing

High-Throughput Sequencing Methods
   Solexa/Illumina: Reversible Terminator Method
   Helicos: Single Molecule Sequencing
   454/Roche: Pyrosequencing Method
   SOLiD/ABI: Supported Oligo Ligation Method

Research Applications

Examples of Short Read Alignment Algorithms

Sequence Processing and Analysis

Hands-on Manual

References and Books
Performance Requirements

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.
- 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.
- Only best hits with almost perfect alignments are required.
- 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.
List of Short-Read Alignment Tools (not complete)

- Indexing Reference with Suffix Array/Burrows-Wheeler
  - Bowtie [Langmead et al. 2009]
  - SOAPv2

- Indexing Reads with Hash Tables
  - ZOOM: uses spaced seeds algorithm [Lin et al. 2008]
  - 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 (commercial)

- Indexing Reference with Hash Tables
  - SOAPv1 [Li et al. 2008]

- Merge Sorting
  - Slider [Malhis et al. 2009]
MAQ: Mapping Quality

Algorithm [Li et al 2008b]

- Uses a hashing technique that guarantees to find alignments with up to two mismatches in the first 28 bp of the reads.
- It indexes the read sequences in six hash tables and scans the reference genome sequence for seed hits that are subsequently extended and scored.
- The commercial Eland alignment program uses a very similar approach.

Performance

- Slower than Bowtie and SOAP. Intermediate rank with regard to memory footprint.

Features

- Versatile pipeline for SNP detection.
- Can report all hits for queries with multiple ones.
- Allows at most two mismatches.
- Performs on single reads only ungapped alignments. Gaps only possible for paired end reads by applying Smith-Waterman algorithm on small candidate set.

Limitations

- Read length limit 128bp
- No gapped alignment for single-end reads
- All sequences in one run need to have the same length.
MAQ Algorithm - Step 1: Indexing

- It builds six hash tables to index the reads and scans the reference sequence against the hash tables to find the hits. This ensures that sequences with up to two mismatches will be found.
- The six hash tables correspond to six noncontiguous seed templates. For example, the six templates for 8bp reads are 11110000, 00001111, 11000011, 00111100, 11001100 and 00110011.
- Only the nucleotides at 1’s will be indexed, but not those at 0’s using a 24-bit integer for hashing.
- When all the reads are processed, the 24-bit integers are sorted, such that reads with the same hashing integer are grouped together. Each integer and its corresponding region are then recorded in a hash table with the integer as the key.
- The indices are built for only the first 28bp of the reads, which are typically the most accurate part of the read.
MAQ: Algorithm - Step 2: Searching

- Each 28-bp subsequence of the reference will be hashed with the first two templates (TMPA in Fig. 1) used for indexing and will be looked up in the corresponding two hash tables.
- It maps the reads to a position that minimizes the sum of quality values of mismatched bases.
- When the scan of the reference is complete, the next two templates (TMPB in Fig. 1) are applied and the reference will be scanned once again (TMPC in Fig. 1) until no more templates are left.
- Using six templates guarantees to find seed hits with no more than two mismatches.
- If a hit is found to a read, MAQ will calculate the sum of qualities of mismatched bases $q$ over the whole length of the read, extending out from the 28bp seed without gaps.
MAQ: Algorithm - Illustration

**Indexing and Searching**

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sequence</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref4</td>
<td>TACGCGAT</td>
<td>2 continuous mismatches</td>
</tr>
<tr>
<td>Ref3</td>
<td>AACGGAT</td>
<td>2 spaced mismatches</td>
</tr>
<tr>
<td>Ref2</td>
<td>AACGCGAT</td>
<td>1 mismatch</td>
</tr>
<tr>
<td>Ref1</td>
<td>ATCGCGAT</td>
<td>perfect match</td>
</tr>
<tr>
<td>Read</td>
<td>ATCGCGAT</td>
<td>Ref1 Ref2 Ref3 Ref4</td>
</tr>
</tbody>
</table>

| TMPA    | 11110000 | Y N N N |
| TMPA    | 00001111 | Y Y N Y |
| TMPB    | 11000011 | Y N N N |
| TMPB    | 00111100 | Y Y N Y |
| TMPC    | 11001100 | Y N N N |
| TMPC    | 00110011 | Y Y Y Y |

*Fig. 1: Indexing and Search Strategy of MAQ Algorithm*
Error probabilities are computed for the final genotype calls. It uses a Bayesian statistical model. This model incorporates:

- mapping qualities
- error probabilities from the raw sequence quality scores
- sampling of the two haplotypes
- an empirical model for correlated errors at a site
SOAP

Algorithm [Li et al 2008]
- Uses a combination of seed and hash look-up table algorithms

Performance
- 300-1200 faster and higher sensitivity than BLASTN

Features
- Can report all hits for queries with multiple matches.
- Allows at most two mismatches.
- Performs ungapped and gapped alignments. At most one gap of 1-3 bases in length.
- Mismatches have precedence over gaps.
- Does not allow gaps in flanking regions of gapped alignments.
- Aligns paired end reads simultaneously (only one read can have gap).

Limitations
- High memory requirements for large genomes: 14GB RAM for human genome
SOAP: Algorithm

- Uses seed and hash look-up tables to accelerate search and alignment processes.
- Both the reads and the reference sequences are converted by a hashing function to numeric data using 2-bits-per-base for encoding. A search is performed by exclusive-OR comparisons with the reference sequence. Then the value is used as suffix to check the look-up table to know how many bases are different.
- To allow two mismatches, it uses a similar seed template approach as Eland and MAQ by splitting each read into four regions.
- Uses the enumeration algorithm for inserting gaps. This method tries to insert a continuous gap or deletion at each possible position of a read.
- The algorithm obtains the same alignments as dynamic programming but much faster.
- Unlike Eland and Maq, SOAP loads the reference sequences into memory. For each read, it creates seeds and searches the corresponding index table for candidate hits, computes an alignment and reports the results.
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.
The Burrows-Wheeler matrix and transformation for 'acaacg'. Data compression is facilitated by forming stretches of the same characters.

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).

Steps taken by an exact search to identify the range of rows, and thus the set of reference suffixes, prefixed by 'aac'.

(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.

Fig. 2: Inexact match for query 'GGTA'
Useful Links

- Link Collections for Short-Read Alignment Tools
  - Alignment Tools List from MAQ Developer
  - Alignment Performance Page from BWA Developer
Outline

Traditional DNA Sequencing Technologies
  Chemical Sequencing
  Sanger Sequencing

High-Throughput Sequencing Methods
  Solexa/Illumina: Reversible Terminator Method
  Helicos: Single Molecule Sequencing
  454/Roche: Pyrosequencing Method
  SOLiD/ABI: Supported Oligo Ligation Method

Research Applications
  Examples of Short Read Alignment Algorithms

Sequence Processing and Analysis
  Hands-on Manual

References and Books
## Software for Illumina Sequence Data Analysis

<table>
<thead>
<tr>
<th>Type</th>
<th>Software Packages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Analysis</td>
<td>RTA (Real Time Analysis)</td>
</tr>
<tr>
<td>Base Calling</td>
<td>OLB (Off-Line Base caller), Alta-Cyclic, Rolexa (R)</td>
</tr>
<tr>
<td>Alignment</td>
<td>CASAVA Eland, BWA, Soap, Mosaik, Bowtie, 40 more!!</td>
</tr>
<tr>
<td>Quality Control</td>
<td>Illumina pipeline, SolexaQA, TileQC, BioC</td>
</tr>
<tr>
<td>RNA-Seq, ChIP-Seq, SNP-Seq</td>
<td>Over 100 tools including BioC</td>
</tr>
<tr>
<td>Assembly</td>
<td>Velvet, SOAPdenovo, ABySS, Euler, 20 more!!</td>
</tr>
<tr>
<td>Multipurpose Tools</td>
<td>SamTools, BioC, Perl, Python, ...</td>
</tr>
</tbody>
</table>
### Data Formats for Illumina Reads

<table>
<thead>
<tr>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>fastaq</td>
<td>sequences and ASCII quality scores</td>
</tr>
<tr>
<td>sequence text file</td>
<td>sequences and numeric quality scores</td>
</tr>
<tr>
<td>quality score text file</td>
<td>Detailed quality scores (4 per base)</td>
</tr>
</tbody>
</table>
Biosequence Analysis in R and Bioconductor

R Base

- Some basic string handling utilities. Wide spectrum of numeric data analysis tools.

BioConductor

- **Biostrings**: general sequence analysis environment.
- **ShortRead**: pipeline for short read data.
- **IRanges**: infrastructure for positional data.
- **BSgenome**: BioC genome annotation data.
- **biomaRt**: interface to BioMart annotations.
- **rtracklayer**: interface to online and other genome browsers.
- **chipseq & ChIPpeakAnno**: Chip-Seq analysis.

Non-R Alignment Tools

- SOAP
- MAQ
- Bowtie
Outline

Traditional DNA Sequencing Technologies
   Chemical Sequencing
   Sanger Sequencing

High-Throughput Sequencing Methods
   Solexa/Illumina: Reversible Terminator Method
   Helicos: Single Molecule Sequencing
   454/Roche: Pyrosequencing Method
   SOLiD/ABI: Supported Oligo Ligation Method

Research Applications

Examples of Short Read Alignment Algorithms

Sequence Processing and Analysis

Hands-on Manual

References and Books
"HT Sequence Analysis with R and Bioconductor"
Outline

Traditional DNA Sequencing Technologies
  Chemical Sequencing
  Sanger Sequencing

High-Throughput Sequencing Methods
  Solexa/Illumina: Reversible Terminator Method
  Helicos: Single Molecule Sequencing
  454/Roche: Pyrosequencing Method
  SOLiD/ABI: Supported Oligo Ligation Method

Research Applications

Examples of Short Read Alignment Algorithms

Sequence Processing and Analysis

Hands-on Manual

References and Books


