xiaomanshawnli@gmail.com
Office: HEC210
Phone: 823-4811
Office hour: TuTh9:30-10:30
   http://www.cs.ucf.edu/~xiaoman/fall/

For attendance today, answer the following questions:
1. Do you know how to code? If yes, which language?
2. Undergrads: Do you plan to go to graduate school?
   Master: do you plan to pursue PhD at UCF?
3. Do you want to work on a bioinformatics project with me? How much time do you have to spend on the bioinformatics project this semester?
Next Generation Sequencing Technologies

Some slides are modified from Robi Mitra’s lecture notes
What will you do to understand a disease?
What will you do to understand a disease?

- Phenotype
- Genotype
- Hypothesis
- Test Hypothesis
- By Genetic Manipulation
Forward Genetics

Mutation in APC Gene

Two groups:
1. Develop Colorectal cancer At Young Age
2. Do not

Genotype

Phenotype

Hypothesis

APC is a Tumor Suppressor Gene

Test Hypothesis By Genetic Manipulation

Delete APC in Mouse
Control: Isogenic APC+
The Cycle of Forward Genetics

- Observation
- Phenotype
- Genotype
- Thinking
- Hypothesis
- Test Hypothesis
- By Genetic Manipulation
  - Gene Deletion/Replacement
  - Recombinant Technology

In 2005
$9 million/genome
Not feasible

?Sequencing?
The Problem with Forward Genetics

- Phenotype
- Genotype
- Hypothesis
- Test Hypothesis By Genetic Manipulation
- Gene Deletion/Replacement Recombinant Technology
- Sequencing

- Observation
- Thinking

2008
$60,000 /genome
Cost is rapidly dropping

2008
$60,000 /genome
Cost is rapidly dropping
Bp/US dollar: increases exponentially with time

Adapted from Jay Shendure et al 2004
Two questions:

• How was this dramatic acceleration achieved?

• What will it mean?
How was this achieved?

- Integration (Think about sequencing pipeline)
- Parallelization
- Miniaturization

Same concepts that revolutionarized integrated circuits

Plus one additional insight

Read length is Not as important for resequencing
Read Length is Not As Important For Resequencing

% of Paired K-mers with Uniquely Assignable Location

Length of K-mer Reads (bp)

E.COLI
HUMAN

Jay Shendure
Second generation sequencers

• 454 Life Sciences (Roche Diagnostics)
  – 25-50 MB of sequences in a single run
  – Up to 500 bases in length

• Solexa (Illumina)
  – 1 GB of sequences in a single run
  – 35 bases in length

• SOLiD (Applied Biosystems)
  – 6 GB of sequences in a single run
  – 35 bases in length
## Comparing Sequencers

<table>
<thead>
<tr>
<th></th>
<th>Roche (454)</th>
<th>Illumina</th>
<th>SOLiD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chemistry</strong></td>
<td>Pyrosequencing</td>
<td>Polymerase-based</td>
<td>Ligation-based</td>
</tr>
<tr>
<td><strong>Amplification</strong></td>
<td>Emulsion PCR</td>
<td>Bridge Amp</td>
<td>Emulsion PCR</td>
</tr>
<tr>
<td><strong>Paired ends/sep</strong></td>
<td>Yes/3kb</td>
<td>Yes/200 bp</td>
<td>Yes/3 kb</td>
</tr>
<tr>
<td><strong>Mb/run</strong></td>
<td>100 Mb</td>
<td>1300 Mb</td>
<td>3000 Mb</td>
</tr>
<tr>
<td><strong>Time/run</strong></td>
<td>7 h</td>
<td>4 days</td>
<td>5 days</td>
</tr>
<tr>
<td><strong>Read length</strong></td>
<td>250 bp</td>
<td>32-40 bp</td>
<td>35 bp</td>
</tr>
<tr>
<td><strong>Cost per run</strong></td>
<td>$8439</td>
<td>$8950</td>
<td>$17447</td>
</tr>
<tr>
<td>(total)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cost per Mb</strong></td>
<td>$84.39</td>
<td>$5.97</td>
<td>$5.81</td>
</tr>
</tbody>
</table>

From Stefan Bekiranov, Univ of Virginia, 2008
### NGS Technology Comparison

<table>
<thead>
<tr>
<th></th>
<th>ABI SOLiD</th>
<th>Illumina GA</th>
<th>Roche FLX</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cost</strong></td>
<td>SOLiD 4: $495k</td>
<td>Ile: $470k</td>
<td>Titanium: $500k</td>
</tr>
<tr>
<td></td>
<td>SOLiD PI: $240k</td>
<td>Ilx: $250k</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HiSeq: $690k</td>
<td></td>
</tr>
<tr>
<td><strong>Quantity of</strong></td>
<td>SOLiD 4: 100Gb</td>
<td>Ile: 20 - 38 Gb</td>
<td>450 Mb</td>
</tr>
<tr>
<td><strong>Data per run</strong></td>
<td>SOLiD PI: 50Gb</td>
<td>Ilx: 50 – 95 Gb</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HiSeq: 200Gb +</td>
<td></td>
</tr>
<tr>
<td><strong>Run Time</strong></td>
<td>7 Days</td>
<td>4 Days</td>
<td>9 Hours</td>
</tr>
<tr>
<td><strong>Pros</strong></td>
<td>Low error rate due to dibase probes</td>
<td>Most widely used NGS platform. Requires least DNA</td>
<td>Short run time. Long reads better for de novo sequencing</td>
</tr>
<tr>
<td><strong>Cons</strong></td>
<td>Long run times. Has been demonstrated certain reads don’t match reference</td>
<td>Least multiplexing capability of the 3. Poor coverage of AT rich regions</td>
<td>Expensive reagent cost. Difficulty reading homopolymer regions</td>
</tr>
</tbody>
</table>

London Local Genomics Center April 1, 2010 NGS meeting
Technology Overview: Solexa/Illumina Sequencing

Sequencing-By-Synthesis Demo

Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments

http://www.illumina.com/
Immobilize DNA to Surface

**Sequencing-By-Synthesis Demo**

- Adapter
- DNA fragment
- Dense lawn of primers
- Adapter

**Attach DNA to surface**

Bind single stranded fragments randomly to the inside surface of the flow cell channels.

Source: www.illumina.com
Technology Overview: Solexa Sequencing

Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Sequence Colonies

Sequencing-By-Synthesis Demo

First chemistry cycle: determine first base

To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
Sequence Colonies

Image of second chemistry cycle is captured by the instrument. After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
Call Sequence

Sequencing-By-Synthesis Demo

Cycle 1: T G C A
Cycle 2: A C T G
Cycle 3: A T T C
Cycle 4: G A C G
Cycle 5: T A G G

Sequence read over multiple chemistry cycles
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

GCTGA....
Sequencer Output

Image acquisition

Base calling

Quality Scores

Sequence Files

TGCTAAGCAT...
Sequence

- ~10 million sequences per lane
- ~500 MB files
Illumina FASTQ format

FASTQ format stores sequences and Phred qualities in a single file. It is concise and compact. FASTQ is first widely used in the Sanger Institute and therefore we usually take the Sanger specification and the standard FASTQ format, or simply FASTQ format. Although Solexa/Illumina read file looks pretty much like FASTQ, they are different in that the qualities are scaled differently. In the quality string, if you can see a character with its ASCII code higher than 90, probably your file is in the Solexa/Illumina format.

Each sequence entry consists of 4 lines:
- sequence name after @
- sequence
- quality score name after + (optional)
- quality scores in phred format ([http://maq.sourceforge.net/qual.shtml](http://maq.sourceforge.net/qual.shtml))
FASTQ examples

@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCCTCAGCGTTTCTCC
+
;3;7;88

@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;7;3;83

@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGTTGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;9;7;7;393333
Quality Score

• Given a character $q$, the corresponding Phred quality can be calculated with:
  $Q = \text{ord}(q) - 33$; where $\text{ord}()$ gives the ASCII code of a character.

Solexa/Illumina Read Format
The syntax of Solexa/Illumina read format is almost identical to the FASTQ format, but the qualities are scaled differently. Given a character $sq$, the following Perl code gives the Phred quality $Q$:
  $Q = 10 \times \log(1 + 10 \times (\text{ord}(sq) - 64) / 10.0) / \log(10)$;
Bioinformatics Challenges

• Rapid mapping of these short sequence reads to the reference genome

• Visualize mapping results
  – Thousand of enriched regions

• Peak analysis
  – Peak detection
  – Finding exact binding sites

• Compare results of different experiments
  – Normalization
  – Statistical tests
Mapping of Short Oligonucleotides to the Reference Genome

• Mapping Methods
  – Need to allow mismatches and gaps
    • SNP locations
    • Sequencing errors
  – Indexing and hashing
    • genome
    • oligonucleotide reads

• Use of quality scores
• Use of SNP knowledge
• Performance
  – Partitioning the genome or sequence reads
Mapping Methods: Indexing the Genome

• Fast sequence similarity search algorithms (like BLAST)
  – Not specifically designed for mapping millions of query sequences
  – Take very long time
    • e.g. 2 days to map half million sequences to 70MB reference genome (using BLAST)
  – Indexing the genome is memory expensive
Sequence analysis

SOAP: short oligonucleotide alignment program

Ruiqiang Li\textsuperscript{1,2}, Yingrui Li\textsuperscript{1}, Karsten Kristiansen\textsuperscript{2} and Jun Wang\textsuperscript{1,2,*}

\textsuperscript{1}Beijing Genomics Institute at Shenzhen, Shenzhen 518083, China and \textsuperscript{2}Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, DK-5230, Denmark

Received on November 10, 2007; revised on December 20, 2007; accepted on January 14, 2008

Advance Access publication January 28, 2008

Associate Editor: Keith Crandall

\textbf{ABSTRACT}

\textbf{Summary}: We have developed a program SOAP for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences. The program is designed to handle the huge amounts of short reads generated by parallel sequencing using the new generation Illumina-Solexa sequencing technology. SOAP is compatible with numerous applications, including single-read or pair-end resequencing, small RNA discovery and mRNA tag sequence mapping. SOAP is a command-driven program, which supports multi-threaded parallel computing, and has a batch module for multiple query sets.

\textbf{Availability}: http://soap.genomics.org.cn

\textbf{Contact}: soap@genomics.org.cn

SOAP will allow either a certain number of mismatches or a continuous gap for aligning a read onto the reference. The best hit of each read which has minimal mismatches or smaller gap will be reported. For multiple best hits, the user can instruct the program to report randomly report one, or disregard all of them. Since the read length is 25-50 bp, hits with too many mismatches are unreliable which are hard to distinguish with random. By default, the program will allow at most two mismatches. Between two haplotype genome sequences, occurrence of nucleotide polymorphism is much higher than the insertions or deletions, so ungapped hits have precise gapped hits. For gapped alignment only one continue with a size ranging from 1 to 3 bp is accepted.
Mapping Methods:
Indexing the Oligonucleotide Reads

- **ELAND** (Cox, unpublished)
  - “Efficient Large-Scale Alignment of Nucleotide Databases” (Solexa Ltd.)

- **SeqMap** (Jiang, 2008)
  - “Mapping massive amount of oligonucleotides to the genome”

- **RMAP** (Smith, 2008)
  - “Using quality scores and longer reads improves accuracy of Solexa read mapping”

- **MAQ** (Li, 2008)
  - “Mapping short DNA sequencing reads and calling variants using mapping quality scores”
Main features

Eland: read length 20-32nt, 2 substitutions
SOAP: 2 substitutions, 1 indel of 1-3nt without any substitutions
RMAP: ungapped mapping, take the tag quality into account
SeqMap: allow <=5 mixture of substitutions and insertions/deletions
Mapping Algorithm (2 mismatches)

GATGCATTGCTATGCCTCCCAGTCCGCAACTTCACG

↓

GATGCATTG  CTATGCCTC  CCAGTCCGC  AACTTCACG  seeds

↓

GATGCATTG  CTATGCCTC  CCAGTCCGC  AACTTCACG

↓

Indexed table of exactly matching seeds

↓

Approximate search around the exactly matching seeds
Mapping Algorithm (2 mismatches)

- Partition reads into 4 seeds \{A,B,C,D\}
  - At least 2 seed must map with no mismatches
- Scan genome to identify locations where the seeds match exactly
  - 6 possible combinations of the seeds to search
    - \{AB, CD, AC, BD, AD, BC\}
    - 6 scans to find all candidates
- Do approximate matching around the exactly-matching seeds.
  - Determine all targets for the reads
  - Ins/del can be incorporated

- The reads are indexed and hashed before scanning genome

- Bit operations are used to accelerate mapping
  - Each nt encoded into 2-bits
Sequence analysis

**SeqMap**: mapping massive amount of oligonucleotides to the genome

Hui Jiang\(^1\), Wing Hung Wong\(^2,\)\(^*\)

\(^1\) Institute for Computational and Mathematical Engineering, Stanford University, Stanford, California 94305, USA.
\(^2\) Department of Statistics, Stanford University, Stanford, California 94305, USA.

Associate Editor: Dr. Limsoon Wong

---

**ABSTRACT**

**Summary**: SeqMap is a tool for mapping large amount of short sequences to the genome. It is designed for finding all the places in a reference genome where each sequence may come from. This task is essential to the analysis of data from ultra high throughput sequencing machines. With a carefully designed index-filtering algorithm and an efficient implementation, SeqMap can map tens of millions of short sequences to a genome of several billions of nucleotides. Multiple substitutions and insertions/deletions of the nucleotide bases in the sequences can be tolerated and therefore detected. SeqMap supports FASTA input format and various output formats, and provides command line options for tuning almost every aspect of the mapping process. A typical mapping can be done in a few hours on a desktop PC. Parallel use of SeqMap on a cluster is also very straightforward.

20–32 nt to the genome, allowing up to 2 substitutions in the mapping. SOAP (Li, 2008) can also handle up to 2 substitutions, or a gap of 1–3 nt without any other substitution. It can also handle longer reads or pair-end reads. RMAP (Smith, 2008) is another program for ungapped mapping, which takes read qualities into account.

Compared to these existing programs, SeqMap offers more flexibility in the mapping. It allows up to 5 mixed substitutions and inserted/deleted nucleotides in the mapping, which is considered sufficient for most mapping applications. FASTA input format and various output formats (e.g., the ELAND format) are supported by SeqMap for the convenience of users. It also provides many command line options for tuning almost every aspect of the mapping process. For instance, SeqMap allows sequences to contain Ns, and to have unequal lengths. Such flexibility is beneficial for the analysis since both sequencing errors and SNPs may cause subti-
Using quality scores and longer reads improves accuracy of Solexa read mapping
Andrew D Smith†, Zhenyu Xuan† and Michael Q Zhang*

Address: Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
Email: Andrew D Smith - asmith@cshl.edu; Zhenyu Xuan - xuan@cshl.edu; Michael Q Zhang - mzhang@cshl.edu
*Corresponding author †Equal contributors

Published: 28 February 2008
Received: 5 October 2007 Accepted: 28 February 2008

Abstract
Background: Second-generation sequencing has the potential to revolutionize genomics and impact all areas of biomedical science. New technologies will make re-sequencing widely available for such applications as identifying genome variations or interrogating the oligonucleotide content of a large sample (e.g. ChIP-sequencing). The increase in speed, sensitivity and availability of sequencing technology brings demand for advances in computational technology to perform associated analysis tasks. The Solexa/Illumina 1G sequencer can produce tens of millions of reads, ranging in length from ~25–50 nt, in a single experiment. Accurately mapping the reads back to a reference genome is a critical task in almost all applications. Two sources of information that are often ignored when mapping reads from the Solexa technology are the 3' ends of longer reads, which contain a much higher frequency of sequencing errors, and the base-call quality scores.

Results: To investigate whether these sources of information can be used to improve accuracy when mapping reads, we developed the RMAP tool, which can map reads having a wide range of lengths and allows base-call quality scores to determine which positions in each read are more important when mapping. We applied RMAP to analyze data re-sequenced from two human BAC regions for varying read lengths, and varying criteria for use of quality scores. RMAP is freely available for downloading at http://ultra.cshl.edu/rmap/.

Conclusion: Our results indicate that significant gains in Solexa read mapping performance can be achieved by considering the information in 3' ends of longer reads, and appropriately using the basecall quality scores. The RMAP tool we have developed will enable researchers to effectively exploit this information in targeted re-sequencing projects.
RMAP (Smith et al, 2008)

• Improve mapping accuracy
  – Possible sequencing errors at 3’-ends of longer reads
  – Base-call quality scores

• Use of base-call quality scores
  – Quality cutoff
    • High quality positions always induce a match
    • Low quality positions are checked for mismatches
  – Quality control step eliminates reads with too many low quality positions

• Allow any number of mismatches
Map to reference genome

- Mapped to a unique location: 7.2 M → 3 M
  - Quality filter

- Mapped to multiple locations: 1.8 M

- No mapping: 2.5 M

- Low quality: 0.5 M

Sample sequences:

- Mapped to a unique location:
  - TGAAAAATTAATGAAAAATATGTATCCCTCGCTCACAC
  - TTAGAATTAAAAATATTAGAGGTGCTACGGCTCC
  - TTAAATATATTGTATATAGTTGATTTGATATAGGCT
  - GTTCGATTCTCTCTCTAGGAGCTTTGATGTCTCC
  - TCTCTCTGATGATATAGTTGATTTGATATAGGCT

- Mapped to multiple locations:
  - TGAAAAATTAATGAAAAATATGTATCCCTCGCTCACAC
  - TTAGAATTAAAAATATTAGAGGTGCTACGGCTCC
  - TTAAATATATTGTATATAGTTGATTTGATATAGGCT
  - GTTCGATTCTCTCTCTAGGAGCTTTGATGTCTCC
  - TCTCTCTGATGATATAGTTGATTTGATATAGGCT

- No mapping:
  - TGAAAAATTAATGAAAAATATGTATCCCTCGCTCACAC
  - TTAGAATTAAAAATATTAGAGGTGCTACGGCTCC
  - TTAAATATATTGTATATAGTTGATTTGATATAGGCT
  - GTTCGATTCTCTCTCTAGGAGCTTTGATGTCTCC
  - TCTCTCTGATGATATAGTTGATTTGATATAGGCT

- Low quality:
  - TGAAAAATTAATGAAAAATATGTATCCCTCGCTCACAC
  - TTAGAATTAAAAATATTAGAGGTGCTACGGCTCC
  - TTAAATATATTGTATATAGTTGATTTGATATAGGCT
  - GTTCGATTCTCTCTCTAGGAGCTTTGATGTCTCC
  - TCTCTCTGATGATATAGTTGATTTGATATAGGCT

- Detailed sequences:
  - TGAAAAATTAATGAAAAATATGTATCCCTCGCTCACAC
  - TTAGAATTAAAAATATTAGAGGTGCTACGGCTCC
  - TTAAATATATTGTATATAGTTGATTTGATATAGGCT
  - GTTCGATTCTCTCTCTAGGAGCTTTGATGTCTCC
  - TCTCTCTGATGATATAGTTGATTTGATATAGGCT

- Summary:
  - Total mapped: 12 M
  - Quality filtered: 7.2 M → 3 M
  - Mapped to a unique location: 7.2 M
  - Mapped to multiple locations: 1.8 M
  - No mapping: 2.5 M
  - Low quality: 0.5 M
Bioinformatics Challenges

• Rapid mapping of these short sequence reads to the reference genome

• Visualize mapping results
  – Thousand of enriched regions

• Peak analysis
  – Peak detection
  – Finding exact binding sites

• Compare results of different experiments
  – Normalization
  – Statistical tests
Align/Assemble to a reference
* **Bowtie** - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hour on a workstation with 2 gigabytes of memory. [Link to discussion thread here](#). Written by Ben Langmead and Cole Trapnell.
* **ELAND** - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony Solexa 1G machine.
* **Exonerate** - Various forms of alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Birney.
* **EMBL C for POSIX.**
* **GMAP** - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentech.
* **MAQ** - Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina-Solexa 1G Genetic Analyzer, and has preliminary handle ABI SOLID data. Written by Henry Le from the Sanger Centre.
* **MUMmer** - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficient library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, Andrew Smith, Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. Perl required.
* **Novocraft** - Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Available for evaluation use and for use on open not-for-profit projects. Requires Linux or Mac OS X.
* **RMAP** - Assembles 20 - 64 bp Solexa reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinformatics) OS required.
* **SeqMap** - Works like ELand, can do 3 or more bp mismatches and also INDELS. Written by Hui Jiang from the Wong lab at Stanford. Builds available for MacOS.
* **SHRIMP** - Assembles to a reference sequence. Developed with Applied Biosystem's colurspace genomic representation in mind. Authors are Michael Bruwe and Stephen Rumble at the University of Toronto.
* **Slider** - An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a reference or a set of reference sequences. Authors are from BCGSC. Paper is [here](#).
* **SSAHA** - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases using hashing. Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.
* **SXOligoSearch** - SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illuma reads against a range of Refseq genome builds for a number of organisms. Web Portal. OS independent.

de novo Align/Assemble
* **MIRA** - MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo assemblies using reads gathered through 454 sequencing technology or GS FLX). Compatible with 454, Solexa and Sanger data. Linux OS required.
* **SHARC** - De novo assembly of short reads. Authors are Dohm JC, Lottaz C, Borodina T and Himmelbauer H. from the Max-Planck-Institute for Molecula.
* **SSAKE** - Version 2.0 of SSAGE (23 Oct 2007) can now handle error-rich sequences. Authors are René Warren, Granger Sutton, Steven Jones and Robert. Canada's Michael Smith Genome Sciences Centre. Perl/Linux.
* **VCACE** - De novo assembly of short reads with robust error correction. An improvement on earlier versions of SSAKE.
* **Velvet** - Velvet is a de novo genomic assembler specially designed for short read sequencing technologies, such as Solexa or 454. Need about 20-25X coverage paired reads. Developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI).
Visualization

• BED files are built to summarize mapping results

• BED files can be easily visualized in Genome Browser

http://genome.ucsc.edu
Visualization: Genome Browser

Visualization: Custom

300 kb region from mouse ES cells

Pacific Biosciences: A Third Generation Sequencing Technology

Eid et al 2008
Applications

- Genomes
- Re-sequencing Human Exons (Microarray capture/amplification)
- small (including mi-RNA) and long RNA profiling (including splicing)
- ChIP-Seq:
  - Transcription Factors
  - Histone Modifications
  - Effector Proteins
- DNA Methylation
- Polysomal RNA
- Origins of Replication/Replicating DNA
- Whole Genome Association (rare, high impact SNPs)
- Copy Number/Structural Variation in DNA
- ChIA-PET: Transcription Factor Looping Interactions
- ???
references
