Genetics 211 - 2014
Lecture 1

Genome Sequencing
Gavin Sherlock
gsherloc@stanford.edu
January 7th 2014
Overview of My Lectures

• Genome Sequencing
  – Sanger Sequencing (Lecture 1)
    • Whole Genome Sequencing
    • Sequencing Theory
    • Genome Assembly
  – High Throughput Sequencing Technology (Lecture 1)
    • Illumina
    • PacBio
    • Oxford Nanopore

• Resequencing + Functional Genomics
  – High Throughput Sequencing
    • Applications (Lecture 2)
    • Analysis (Lecture 3)
  – Expression
    • RNA-Seq (Lecture 4)
    • Cluster Analysis (Lecture 4)
What to Sequence and Why?

- **De novo whole genome sequencing**
  - requires de novo whole genome assembly

- **Polymorphism discovery** (distinct from genotyping!)
  - Targeted approaches (exome)
  - Whole genome
  - SNPs, copy number variations, insertions, deletions, etc.

- **Expressed sequence discovery**
  - ESTs
  - cDNAs
  - miRNAs, etc

- **Functional genomics**
  - ChIP
  - Expression profiling
  - Nucleosome positioning
  - Etc.
Four Fundamentally Different Approaches to DNA Sequencing

- Chemical degradation of DNA
  - Maxam-Gilbert
  - obsolete
- Sequencing by synthesis (“SBS”)
  - uses DNA polymerase in a primer extension reaction
  - most common approach
  - Sanger developed it (“Sanger sequencing”)
  - Illumina, Pacific Biosciences, Ion Torrent, 454
- Ligation-based
  - sequencing using short probes that hybridize to the template
  - SOLiD, Complete Genomics
- Other
  - Nanopore
5’ and 3’


<table>
<thead>
<tr>
<th>Base</th>
<th>plus sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenine</td>
<td>Adenosine</td>
</tr>
<tr>
<td>Guanine</td>
<td>Guanosine</td>
</tr>
<tr>
<td>Cytosine</td>
<td>Cytidine</td>
</tr>
<tr>
<td>Thymine</td>
<td>Thymidine</td>
</tr>
</tbody>
</table>

in DNA: “deoxyadenosine”

plus triphosphate
“deoxynucleotide”
“2’-deoxyadenosine 5’-triphosphate” = dATP

If I throw in DNA polymerase and free nucleotide, which end gets extended?

Antiparallel
Watson 5’ TAGCGTCAAGCTG 3’
Crick 3’ ATCGCAGTCA 5’

In Sanger sequencing, Crick is the template and Watson’s synthesis starts at the primer’s 3’OH
The Chain Terminator

- Dideoxy nucleotides cannot be further extended, and so terminate the sequence chain.
Original Sanger Sequencing with Radioactive Signal

Set up four reactions

- dATP, dGTP, dCTP, ddATP, ddGTP, ddCTP
- dATP, dGTP, dCTP, ddATP, ddGTP, ddCTP
- dATP, dGTP, ddATP, ddGTP, ddCTP
- dATP, ddATP, ddGTP, ddCTP

A nested series of DNA fragments ending in the base specified by the terminator-ddNTP

very low concentration of ddNTPs compared to dNTPs

Exposure to x-ray film (to make an 'auto-radiogram')

Watson's

Template (Crick)
This is great, but…

Wouldn’t it be great to run everything in one lane?
- Save space and time, more efficient
Also, would be nice to read everything at the same point in the gel
- Unable to read sequence near the top, as the bands get closer and closer together.

Fluorescently label the ddNTPs so that they each appear a different color, and can be read by a laser at a fixed point
Fluorescent Sanger Sequencing: “Dye-terminators”

Each of the 4 ddNTPs is labeled with a different fluorescent dye (instead of radioactivity)
Fluorescent Sanger Sequencing

One-tube sequencing reaction
(note: cycle sequencing with modified Taq Polymerase)

Load on gel
(modern machines use capillaries, not slab gels)

Direction of electrophoresis

Scanning laser excites fluorescent dyes as DNA fragments pass by during electrophoresis

Data are sent to a computer

Emitted light is collected by optical detector
Fluorescent Sanger Sequencing Trace

Lane signal

(Real fluorescent signals from a lane/capillary are much uglier than this).

Various algorithms to boost signal/noise, correct for dye-effects, mobility differences, etc., generates the ‘final’ trace (for each capillary of the run)

Trace
Sanger Base Calling

Quality score = -10 * log(probability of error)
For Q20, probability of error = 1/100
For Q99, probability of error ~10^{-10}
Phred: The base-calling program

• Algorithm based on ideas about what might go wrong in a sequencing reaction and in electrophoresis

• Tested the algorithm on a huge dataset of “gold standard” sequences (finished human and C. elegans sequences generated by highly-redundant sequencing)

• Compared the results of phred with the ABI Basecaller

• Phred was considerably more accurate (40-50% fewer errors), particularly for indels and particularly for the higher quality sequences

(Ewing et al., 1998, Genome Research 8: 175-185; Ewing and Green 1998, Genome Research 8: 186-194)
Progress of Sanger Sequencing Technology

~1,000-fold increase in throughput since 1985 accomplished by incremental improvements of the same underlying technology

2nd Generation Sequencing Technologies have ~500 - 30,000x more throughput than 3730: Illumina, SOLiD, 454 Pyrosequencing ( + PacBio, Ion Torrent, Complete Genomics…)

Radioactive polyacrylamide slab gel
Low throughput, labor intensive

AB slab gel sequencers
(370, 373, 377)
Fluorescent sequencing
1990-1999
6 runs/day
96 reads/run
500 bp/read
288,000 bp/day

AB capillary sequencers
(3700, 3730)
1998-now
24 runs/day
96 reads/run
550 – 1,000 bp/read
1-2 million bp/day
Whole Genome Sequencing

• Two main challenges:
  – Getting sufficient “coverage” of the genome
    • A function of read length, number of reads, complexity of library, and size of genome
  – Assembling the sequence reads into a complete genome
    • A function of coverage, and repeat size and repeat frequency
How much sequence do I need?

- Let $L$ = read Length; $G$ = Genome size.
- Assume $L \ll G$.
- $P_{\text{obs\_with\_one\_part\_fragment}} = \frac{L}{G}$
- $P_{\text{not\_obs\_with\_one\_part\_fragment}} = 1 - \frac{L}{G}$
- $P_{\text{not\_obs\_with\_N\_fragments}} = (1 - \frac{L}{G})^N$
- $P_{\text{covered\_by\_at\_least\_one\_fragment}} = 1 - (1 - \frac{L}{G})^N$
- Rearranging gives: $N = \ln(1 - P) / \ln(1 - \frac{L}{G})$
Example Calculation

- *E. coli* genome $G = 4.6$Mb, read length $L = 800$bp
- How many reads do I need to have a certain probability of observing any particular piece of my genome?
- Remember $N = \ln(1-P)/\ln(1-L/G)$
- \( P = 0.9 \Rightarrow \sim 13,000 \sim 2.3\times\) coverage
- \( P = 0.95 \Rightarrow \sim 17,000 \sim 3\times\) coverage
- \( P = 0.99 \Rightarrow \sim 26,500 \sim 4.6\times\) coverage
Back of the Envelope

- Remember, $P = 1 - (1-L/G)^N$
- Given $(1-L/G)^N \approx e^{-NL/G}$
- And, coverage, $R = NL/G$
- Then, $P \approx 1-e^{-R}$
- This is a widespread back of the envelope calculation for any project involving redundancy.
Probability as a Function of Coverage
Overcoming repeats

• Most problematic when:
  – Repeats are longer than read lengths
  – Repeats are present in many copies

• Recognize based on coverage

• Resolve with longer range continuity information:
  – Paired-end reads
  – Multiple insert size libraries
    • Plasmids
    • Fosmids
    • BAC ends
Whole Genome Sequencing Approaches

Hierarchical Shotgun Approach

Genomic DNA

BAC library

Organized, Mapped Large Clone Contigs

Shotgun Clones

Reads

Assembly

GCAATGAAATATGTTCTTGTAATTAGCTGACACTCTAATTAGCTTTGGCACAGCTCTGCTCCTACCTGAGTCTACCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAG

AGCTCTTGCTCTGACTGAGCTCCTACCTAATTTATGTTATGGAATTGACTTGGGTTTTCTCTTTTTCTTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAG

ATGTTCCTGTAATTAGCTGACACTCTAATTAGCTTTGGCACAGCTCTGCTCCTACCTGAGTCTACCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTTTTCTCTTTTTCTTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAG
Whole Genome Sequencing Approaches

Shotgun Approach

Genomic DNA

Shotgun Clones

Reads

Assembly
Rationale for Hierarchical Strategy

- Better for a repeat-rich genome
  - less misassembly of finished genome
    - long-range misassembly largely eliminated and short-range reduced

- Better for an outbred organism
  - each clone from an individual and no polymorphisms in the final sequence.
  - (Added bonus: get SNPs from regions of overlapping clones)
  - Can also get some haplotype information, if individual BACs shotgun sequenced.

- Better if there are cloning biases
  - use minimum tiling path, so the same coverage for each region

- Easier to identify and fill gaps (from unclonable regions) sooner

BUT

- Time consuming and expensive to make minimum tiling path
De Novo Whole Genome Sequencing

Make millions of random clones: “Shotgunning”
Paired End Sequencing Reads

GCAATGAAATATTTGTTTTGTTTTTGAATTTAAGCTGACACTCCTAATTTTAGCTTTGCTCTCTACTGAGTCTACCTAATTATATGTAT
GGATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAGCCTGGGAGAGAGAAACCCCCCAAGCTAGGATTTCTG
CAGCTCATGAAGCCCTTGGAGATAAATGAGTAAGTGGGGAAAATCTTGCTGTTAAAAAGGAAATCTCATCCTCTTTTCTGAAATAT
ATTAGTTGCCATTGATAGGATACTTAAATCTGCAATTTGAACTGAGGATTATTGTTATTACCTCTACTTTAAA
AAGTTTTTTTTTTAATTGAAGGACAGCCACCATGTGAGGAGTTTATTTTAGACTTTTTTTAATTGAATTCAATGGGTTCCTTTGCTGATCC

---

TAAATTTCTTGGAGCTATCCCTAGGATATTGTGAGGATATAAAAAAATTTACATTCTTTTCATCTATCTAAAAACATTCT
GACAGGAAAATTTTCAGATGTAATGCTCATTGACCTAGAAGACATTCTTGAGGAAATTCTGATGTTGAGGAAGAAACATG
ATAAACATAGGTTTAAAAAAATAATTTAGAATAATTTAAGAACGCTTTCAAGATTTAAGACAAAGATAAGAGGAAATGAGGTTCAATTG
CTCAGAAAAATGAGAAAGATTCTTAAAGTT
GGAGTCATAAAAATCTGAGGTTGGCAGAGACCTTAAAGGTCATTAGCTGAACCCACACTCTGGTAGTTGACACTCACA
CTATTGTAGCAATGAGCTGTTTTAATTTTATGATTCTTTTTTCTGAAAGGTTTACAGAATCTTTTCTGAGATCTTTAGGG
Assembly: Contigs and Supercontigs

"Supercontig" or "Scaffold"

"Contig"

Seq gap

"Contig"

number of N's in sequence = estimated size
Why Different Insert Sizes are Useful

Longer (fosmid) mate pairs connect assembly pieces that are not connected by shorter (plasmid) paired ends
Key Concepts in Assembly

- **Contig N50**
  - 50% of the genome assembly is in contigs larger than this size
- **Supercontig (scaffold) N50**
  - same, but for scaffolds

- **k-mer**
  - string of bases of length k
  - for computational efficiency, long sequences such as sanger reads are often chopped up into their constituent k-mers; usually *overlapping* k-mers are used because converting a sequence into nonoverlapping k-mers loses information

- **High-quality mismatch**
  - A position in two well-aligning reads in which the base calls are *high quality but disagree*
  - Indicative of *allelism* or *paralogy*

---

The first three overlapping 22-mers and their positions in a Sanger read

<table>
<thead>
<tr>
<th>Read</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read</td>
<td>tagcgactacctgaactggacctttgaacgag...</td>
<td>tagcgactacctgaactggacc</td>
<td>agcgactacctgaactggacct</td>
</tr>
<tr>
<td>0</td>
<td>tagcgactacctgaactggacc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>agcgactacctgaactggacct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>gcgactacctgaactggacctt</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A high-quality mismatch: High Phred scores (like Q99) on both mismatched bases

Read 1 ..actacctgaactggacctttgaacg...
Read 2 ..actacctgaactagaacctttgaacg...
Assemblies are not Perfect

• Sequence coverage may vary
  – missing regions; strong fragmentation
• Some regions don’t clone well
  – results in low sequence coverage
  – which causes gaps in assembly
• Some regions don’t sequence well
  – extreme GC content
  – homopolymeric or otherwise low-complexity runs
• Some regions don’t assemble well
  – mobile elements
    • high identity, large copy number
  – segmental duplications
    • Repeats are the single biggest impediment to assembly
• Polymorphism
• Best way to improve assemblies is longer reads and better long range continuity
High Throughput Sequencing
The Players

• Commercially available now:
  – Illumina – most prevalent technology
  – SOLiD (Life Technologies)
  – Ion Torrent (Life Technologies)
  – Pacific Biosciences
  – (Complete Genomics)
  – 454, Helicos – both commercially dead

• Next next generation approaches
  – Oxford Nanopore (in public beta testing)
  – Illumina Nanopore
    • Recently licensed an alternative nanopore technology
  – NABsys
Sequencing Template Approaches

- **Clonal Amplification** of Single Molecules
  - Single molecule only briefly needed as a template
  - Thousands of identical molecules boost signal
  - Two different methods
    - Bridge amplification of molecules immobilized on surface
      - Illumina
    - Emulsion PCR
      - SOLiD and Ion Torrent, 454

- Approaches that use a single DNA molecule as a sequencing template do exist (Pacific Biosciences, Oxford Nanopore, Helicos).
  - Challenges include:
    - Keeping single molecules stable during insults of sequencing
    - Signal to noise ratio in base detection
    - BUT
    - Avoid amplification biases
Differences in Throughput

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sanger (AB 3730)</th>
<th>Illumina (HiSeq 2500)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read L (bp)</td>
<td>800</td>
<td>2 x100</td>
</tr>
<tr>
<td>Number of reads per run [days]</td>
<td>96 [&lt;1]</td>
<td>6,000,000,000 [11 days]</td>
</tr>
<tr>
<td>Throughput</td>
<td>6Mb/day</td>
<td>50Gb/day</td>
</tr>
<tr>
<td>SNP error rate</td>
<td>low</td>
<td>high (~0.5%)</td>
</tr>
<tr>
<td>Indel error rate</td>
<td>low</td>
<td>low</td>
</tr>
<tr>
<td>Costs</td>
<td>$500/Mb</td>
<td>&lt;$0.05/Mb</td>
</tr>
</tbody>
</table>
Illumina: Flow Cells with “Molecular Colonies”

- flow cell with randomly spaced molecular clusters
- spacing depends on initial seeding of the single molecules onto the flow cell
Detection, Chemistry

- Massively Parallel Detection on immobilized “molecular colonies”
- Means you have to measure (image) every cycle, instead of the Sanger model (letting reaction go to completion and then separating products by size)
- Requires specially designed chemistry, using reversible dye-terminators and a polymerase
Illumina Sequencing Technology
Robust Reversible Terminator Chemistry Foundation

DNA (0.1-1.0 ug)

Sample preparation

Cluster growth

Sequencing

1 2 3 4 5 6 7 8 9

Image acquisition

Base calling

TGCTACGAT
Illumina Sequence Visualization

250+ Million Clusters Per Flow Cell

100 Microns

20 Microns
Illumina Sequencing: Reversible Terminators

**Incorporate**

- Fluorophore
- Cleavage site

**Detection**

- **3’ OH is blocked**
- **Deblock and Cleave off Dye**

**Free 3’ end**

Ready for Next Cycle
Flow Cells and Imaging

Illumina: single 8-channel
First chemistry, then imaging

Reagents flowed in here

Out to waste

etc (four images per panel, one for each color)

Number of image panels in each channel:

3 columns X 100 rows
2400 for each flow cell
times 4 images
times 35 cycles
=336,000 images

~1 TB image data per run
Image Processing, Base Calling

• Image processing algorithms find signals in each panel, align signals from different panels, etc.
  – Machines ship with server or small cluster that does image analysis while run is happening

• Sequence data after base calling much reduced in size (tens of gigabytes) => more manageable but still large amounts that add up over time

• It is simply unsustainable to keep image data for anything but the short term. People discard the images, and just keep the sequences (fastq format).
Pacific Biosciences

- Single Molecule Real Time DNA Sequencing
- Read lengths now averaging ~5kb, max 20kb
- Strobe sequencing
- Observation of DNA modifications
- Throughput per run is low, but run time is short
- Initial release in late 2010; up to 80,000 reads, of ~1.5kb each in ~15 minutes for $100.
- Error rate is high, though hybrid approaches can significantly improve assemblies generated by short reads alone.
Oxford Nanopore

• MinION and GridION products
• Not yet on market, but in early release
• DNA “sequenced” as it is dragged through a nanopore
• 4% error rate? Supposed to be 0.1-1% by release
• Reads as long as 100kb?
• No data yet released – keep an eye on the AGBT meeting next month
Recommended Reading

Early Sequencing Technology:


New Sequencing Technologies:

Recommended Reading

Landmark Genome Sequencing Papers:


Assembly Algorithms:


Recommended Reading

Reviews:


Sequencing Theory: