NOW GENERATION SEQUENCING
SEQEUNCING TIMELINE

1953: Structure of DNA

1975: Sanger method for sequencing

1985: Human Genome Sequencing Project begins

1990s: Clinical sequencing begins

1998: NHGRI $1000 Genome Goal established

2003: Human Genome Project completed. $100 million x 10 years

2005: First Next-Generation Platform released

2011: Small form-factor NGS: IonTorrent, MiSeq
EXPONENTIAL GROWTH

Rate at which sequence data can be acquired is outpacing Moore’s Law
EXPONENTIAL GROWTH

Rate at which sequence data can be acquired is outpacing Moore’s Law.

Monday, December 5, 11
PLATFORM NUTS+BOLTS

Roche 454™

Illumina Genome Analyzer
http://www.youtube.com/watch?v=77r5p8lBwJk

Life Technologies SOLiD

Pacific Biosciences
http://www.youtube.com/watch?v=_B_cUZ8hSYU

IonTorrent
http://www.youtube.com/watch?v=yVf2295JqUg

Platform mechanistic underpinnings have real bioinformatic consequences
# QUICK COMPARISON

<table>
<thead>
<tr>
<th>Platform</th>
<th>Read length</th>
<th>Run time</th>
<th>Gb/run</th>
<th>Cost</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Roche/454 Titanium GS FLX</strong></td>
<td>400-500bp</td>
<td>8-12h</td>
<td>0.45</td>
<td>$500k</td>
<td>Longer read length; Short run time.</td>
<td>Relatively expensive compared to Illumina or SOLID. Polynucleotide tract errors.</td>
</tr>
<tr>
<td><strong>Illumina GA II, HiSeq, MiSeq</strong></td>
<td>36-100bp</td>
<td>1-4d</td>
<td>18-500</td>
<td>$550k</td>
<td>Massive throughput; Inexpensive</td>
<td>Higher error rates; Rapidly evolving platform. Less ability to multiplex.</td>
</tr>
<tr>
<td><strong>Lifetech SOLiD and ABI 5500</strong></td>
<td>50bp</td>
<td>4-5d</td>
<td>30-170</td>
<td>$595k</td>
<td>Colorspace (2bit) bioinformatics.</td>
<td>Colorspace bioinformatics; Lower adoption rate than Illumina. Long run times.</td>
</tr>
<tr>
<td><strong>PacBio</strong></td>
<td>850-1500bp</td>
<td>40m</td>
<td>0.40</td>
<td>$695k</td>
<td>Speed; Run length</td>
<td>Perpetual beta; Strobing; Low throughput; Cost</td>
</tr>
<tr>
<td><strong>Ion Torrent</strong></td>
<td>&gt;200bp</td>
<td>2h</td>
<td>0.1-1.0</td>
<td>$75k</td>
<td>Low instrument cost and convenient form factor; Short run-time. Good run length.</td>
<td>Low throughput</td>
</tr>
</tbody>
</table>

Source: Metzker, Nat. Rev. Genetics, Jan 2010 + Additional research
## Experiment Types

<table>
<thead>
<tr>
<th>Genomic DNA</th>
<th>RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genome discovery</strong></td>
<td><strong>Transcript discovery</strong></td>
</tr>
<tr>
<td>- De novo assembly</td>
<td>- De novo assembly</td>
</tr>
<tr>
<td>- Reference-guided assembly</td>
<td>- Reference-guided discovery</td>
</tr>
<tr>
<td>- Variant identification</td>
<td></td>
</tr>
<tr>
<td>- SNPs</td>
<td></td>
</tr>
<tr>
<td>- Indels and rearrangements</td>
<td></td>
</tr>
<tr>
<td><strong>Enrichment-Seq</strong></td>
<td><strong>Small molecule profiling</strong></td>
</tr>
<tr>
<td>- ChIPseq</td>
<td>- miRNA sequencing</td>
</tr>
<tr>
<td>- C3</td>
<td>- Small RNA sequencing</td>
</tr>
<tr>
<td>- Histone occupancy</td>
<td></td>
</tr>
<tr>
<td>- DNAse I Hypersensitivity</td>
<td></td>
</tr>
<tr>
<td><strong>Methylation profiling</strong></td>
<td><strong>Quantitative Measures</strong></td>
</tr>
<tr>
<td>- Bisulfite sequencing</td>
<td>- RNA-seq</td>
</tr>
<tr>
<td>- MeDIP (ChIP)</td>
<td>- Allele-specific expression</td>
</tr>
<tr>
<td>- Methylation-sensitive restriction enzymes</td>
<td></td>
</tr>
</tbody>
</table>
GENOME DISCOVERY

Reference-guided assembly

related genome

algorithm

Many sequence reads, randomly generated

Contigs/chromosomes

De novo assembly
GENOME DISCOVERY

Many sequence reads, randomly generated

Reference genome

Algorithm

Consensus sequence
Deletion
Rearrangement

Genome sequence alignments
ENRICHMENT

Many sequence reads, enriched for specific areas of the Reference Genome

Complex biochemistry involving ultrasonic DNA breakage, antibodies, liquid buffers and other things that will ruin your computer....
Visualization is key to ChIP

ChIP-seq data in Broad Institute IGV

Distribution of reads is decidedly non-random

Visible peak structures in ChIP data


Empirical method to assess saturation

Sub-sample 0.1, 0.3, 0.5, 0.7, 0.9 x

Run feature-finding algorithm on each sample

Plot features vs coverage depth

Genome alignments

Features detected

Subsampling

0.1 0.3 0.5 0.7 0.9
TRANSCRIPT DISCOVERY

Reference-guided discovery

De novo discovery

Many RNA sequence reads, semi-randomly generated

Reference genome

Algorithm

Transcript contigs
QUANTITATIVE RNA-SEQ

RNA libraries from \( n \) treatments

Sample 1/\( n \)

Sample 2/\( n \)

Reference genome

Algorithm

Transcriptome alignment 1/\( n \)

Transcriptome alignment 2/\( n \)

Sequence alignments

Algorithm

Correct for

- Library size
- Transcript length
- Sequencing biases
- Platform biases
- Paralogous genes
- Polymorphism

Tabular results

**Monday, December 5, 11**
ASSEMBLERS

1. Hash read space
2. Build De Bruijn graph
3. Walk graph, starting at arbitrary nodes.
4. Pop “bubbles”
5. Perform scaffolding
# ASSEMBLERS

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Special details</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSAKE/VCAKE</td>
<td>Overlap and extend. Limited assembly size.</td>
</tr>
<tr>
<td>ALLPATHS-LG</td>
<td>Winner of the recent Assemblathon. Up-front sequencing recipe required.</td>
</tr>
<tr>
<td>Velvet</td>
<td>De Bruijn. Robust repeat resolution. Support for mixed assemblies. Large RAM requirement</td>
</tr>
<tr>
<td>Newbler</td>
<td>454™ only. Unpublished mechanics.</td>
</tr>
<tr>
<td>Trinity</td>
<td>Optimized for transcriptome assembly. Large RAM requirement*</td>
</tr>
<tr>
<td>ABYSS</td>
<td>MPI-Parallel. Complex implementation. Can still require large RAM node at one step.</td>
</tr>
<tr>
<td>ALLPATHS</td>
<td>Assumes specific sequencing and library construction strategy. Ranked well in Assemblathon.</td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>Assumes specific sequencing and library construction strategy. Ranked well in Assemblathon.</td>
</tr>
<tr>
<td>clcBIO</td>
<td>Doesn’t do scaffolding. Commercial. But fast and considered to be good.</td>
</tr>
<tr>
<td>Celera WGS</td>
<td>Older bioinformatics formats, Supports mixed assemblies. Not very fast :-/</td>
</tr>
</tbody>
</table>
1. Do you need to do full *de novo* assembly?
2. Simulation is good practice
3. How will you evaluate your assembly qualities?
META-ASSEMBLY

<table>
<thead>
<tr>
<th>Assembler</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCAP/CAP3</td>
<td><a href="http://seq.cs.iastate.edu/">http://seq.cs.iastate.edu/</a></td>
</tr>
</tbody>
</table>

Combine results from multiple methods to leverage best of features of various algorithms.
Get up to date on the Assemblathon
http://assemblathon.org/pages/results

<table>
<thead>
<tr>
<th>Learn all about...</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N50 vs NG50</td>
<td>NG50 = Scaffold/contig length at which you have covered 50% of total genome length</td>
</tr>
<tr>
<td>Fragment Analysis</td>
<td>Count how many randomly chosen fragments from species A genome can be found in assembly</td>
</tr>
<tr>
<td>Repeat analysis</td>
<td>Choose fragments that either overlap or don’t overlap a known repeat</td>
</tr>
<tr>
<td>Gene finding</td>
<td>How many genes can be identified in the new assembly?</td>
</tr>
<tr>
<td>Bacterial contamination</td>
<td>How much non-target contamination is in the genome assembly?</td>
</tr>
<tr>
<td>MAUVE analysis</td>
<td>Miscalled, Uncalled, Missing, Extra bases; Misassemblies; Double cut &amp; join distance</td>
</tr>
<tr>
<td>BWA analysis</td>
<td>What fraction of contigs align to a reference? Lengths of coverage islands. Validity, multiplicity, parsimony,</td>
</tr>
</tbody>
</table>

Which assembler was THE BEST?
VIEWERS: THE
BIG 5

IGV

SeqMonk

Gbrowse

Ensembl

UCSC
<table>
<thead>
<tr>
<th>Browser</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Broad IGV</strong></td>
<td>Easy to configure and use. Can add new genomes. Integrates with network-accessible data</td>
<td>Limited by client side resources and network speed. No capacity to perform analyses.</td>
</tr>
<tr>
<td><strong>SeqMonk</strong></td>
<td>Easy to configure and use. Can add new genomes. Includes basic analytical workflows.</td>
<td>Limited by client side resources and network speed</td>
</tr>
<tr>
<td><strong>Ensembl</strong></td>
<td>Web client is easy to configure and use. Powerful query API. Integrates with network-accessible data. Highly responsive.</td>
<td>Server setup and administration is complex. Large, sophisticated database schema.</td>
</tr>
<tr>
<td><strong>Gbrowse</strong></td>
<td>Web client is easy to configure and use. Powerful query API. Integrates with network-accessible data. Can be highly responsive.</td>
<td>Server setup and administration is complex due to Perl module dependencies.</td>
</tr>
<tr>
<td><strong>UCSC</strong></td>
<td>Highly responsive. Nice configuration options. Powerful web-based query API. Massive # of tracks.</td>
<td>You will never figure out how to install or extend it on your own.</td>
</tr>
</tbody>
</table>
DO I NEED HPC?

• I don’t have time to learn the command line or any of this UNIX stuff. I just need to get from point A to point B!

• Galaxy [http://usegalaxy.org/]: Complexity of analysis and data management is hidden, but can still rear its head

• clcBIO Workbench: $10,000 per seat.

• Consulting arrangements: Communications-intensive. Longer turnaround times. Dependency relationship.
GALAXY

Tool selection and data entry

Data products

Shared data

Parameterize and view results of analyses

Contextual help
DO I NEED HPC?

• I can use Linux and all the associated tools but don’t need a lot of horsepower
  • Deploy a personal system
    • You have 100% control of the system -> You’re 100% responsible for the system
    • Your upgrade path is limited to your immediate financial resources
  • Use “The Cloud”
    • CPU time is cheap. But data transfer and storage fees add up rapidly
    • Limited performance and resources
  • Use a traditional HPC cluster
    • You learn in shared environment
      • Master getting data in and out of the shared system for other applications
    • Someone can help you with maintenance, support, data security
    • System is constantly being upgraded. sometimes transparently to the user

You don’t have to consume 10000 cores and a PB of disk to qualify for TACC systems. Use as little or as much of the systems as you need.
**DO I NEED HPC?**

"I need a lot of resources"

<table>
<thead>
<tr>
<th>System</th>
<th>Best features</th>
<th>Use cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranger</td>
<td>Loads of cores. Optimized for MPI.</td>
<td>Very high core count MPI jobs.</td>
</tr>
<tr>
<td>Longhorn</td>
<td>&gt;2000 GPGPU; Fast disk; Fat RAM (144GB) nodes</td>
<td>Interactive sessions; GPU computing; SciViz</td>
</tr>
<tr>
<td>Lonestar</td>
<td>Very fast cores; 1 TB RAM nodes; GPGPU nodes; Faster bandwidth than Ranger</td>
<td>Pretty much any throughput biocomputing. Genome assembly.</td>
</tr>
</tbody>
</table>

System choice is important!
PRACTICALS

• Log into Lonestar
• Survey Lonestar’s filesystems
• Accounts and allocations
• Quick introduction to modules

• Tutorial 1: A simple serial job (Picard MergeSamfiles)
• Tutorial 2: An MPI-based parallel job (mpiBLAST)
• Tutorial 3: A parameteric serial job (multi-CPU BWA)