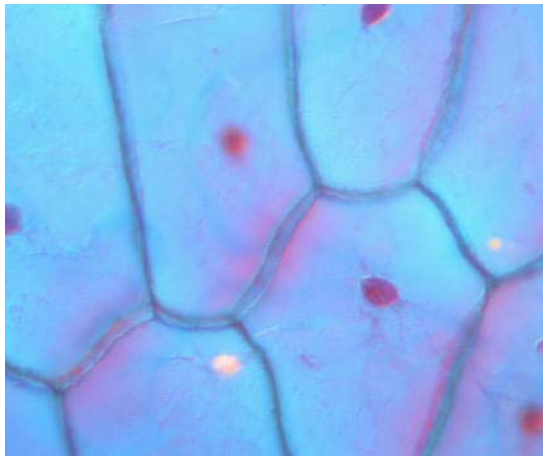
How To Build A Horse: Midyear Report

Megan Smedinghoff

DNA

- Cells in all organisms have a DNA molecule in their nucleus
- DNA is an abbreviation for DeoxyriboNucleic Acid
- DNA is translated into proteins, which determine the structure, function, and behavior of the cell



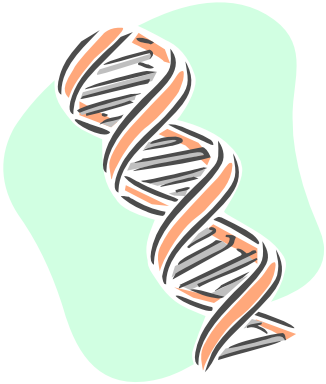
What does DNA look like?

- Linear, double-stranded, helical molecule.
- Consists of four types of nucleotides (bases):

  adenine (A)
  thymine (T)
  guanine (G)
  cytosine (C)

What is a genome?

- A genome is the linear sequence of bases of the DNA molecule: …..GATGACATGTAT…..
- Bacteria: ~2-5 million bases
- Insects (fly): ~200 million bases
- Mammals: ~3 billion bases



Background

- In February 2007, Broad Institute released a draft assembly of the horse genome
- The sequencing cost $15 million
- The assembly contains approximately 2.7 billion bases and was done using 6.8-fold coverage

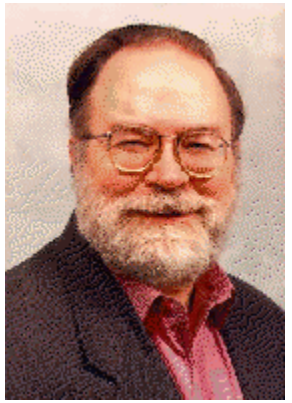
The horse that was sequenced

Why Sequence the Horse?

- Allows scientists to study diseases that primarily affect horses such as Glanders
- Over 80 known genetic conditions exist in both horses and humans; comparative genomics can lead to better treatments for both species
- Work done on the horse will lead to improvements in the assembly pipeline and allow easier assembly of large mammals in the future

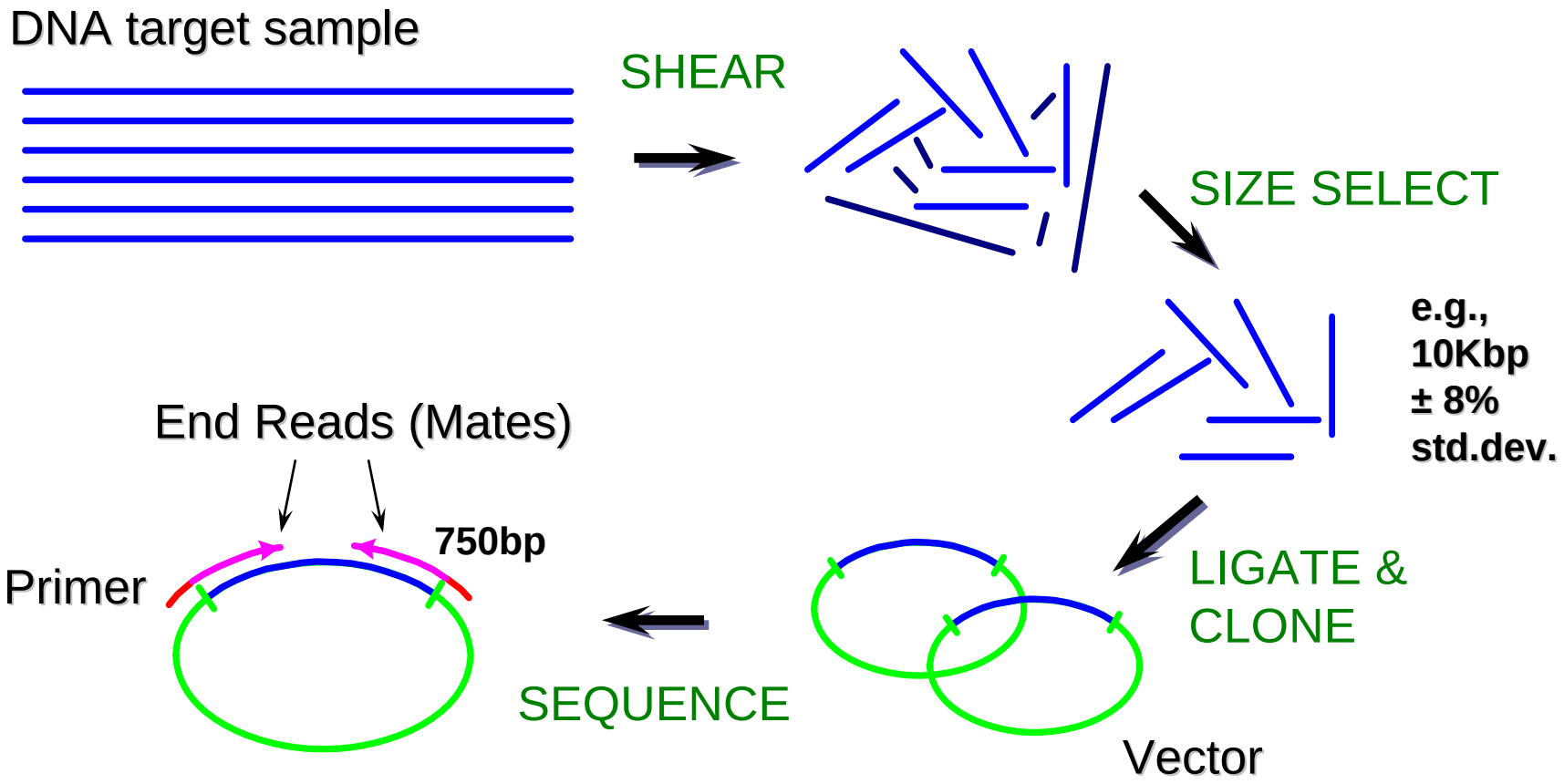
Project Goals

- Reassemble the horse genome using the Celera Assembler
- Compare my assembly with the Broad assembly and produce a reconciled horse genome
- Deposit the improved assembly in GenBank
- Improve existing process for assembly of large mammals wherever possible



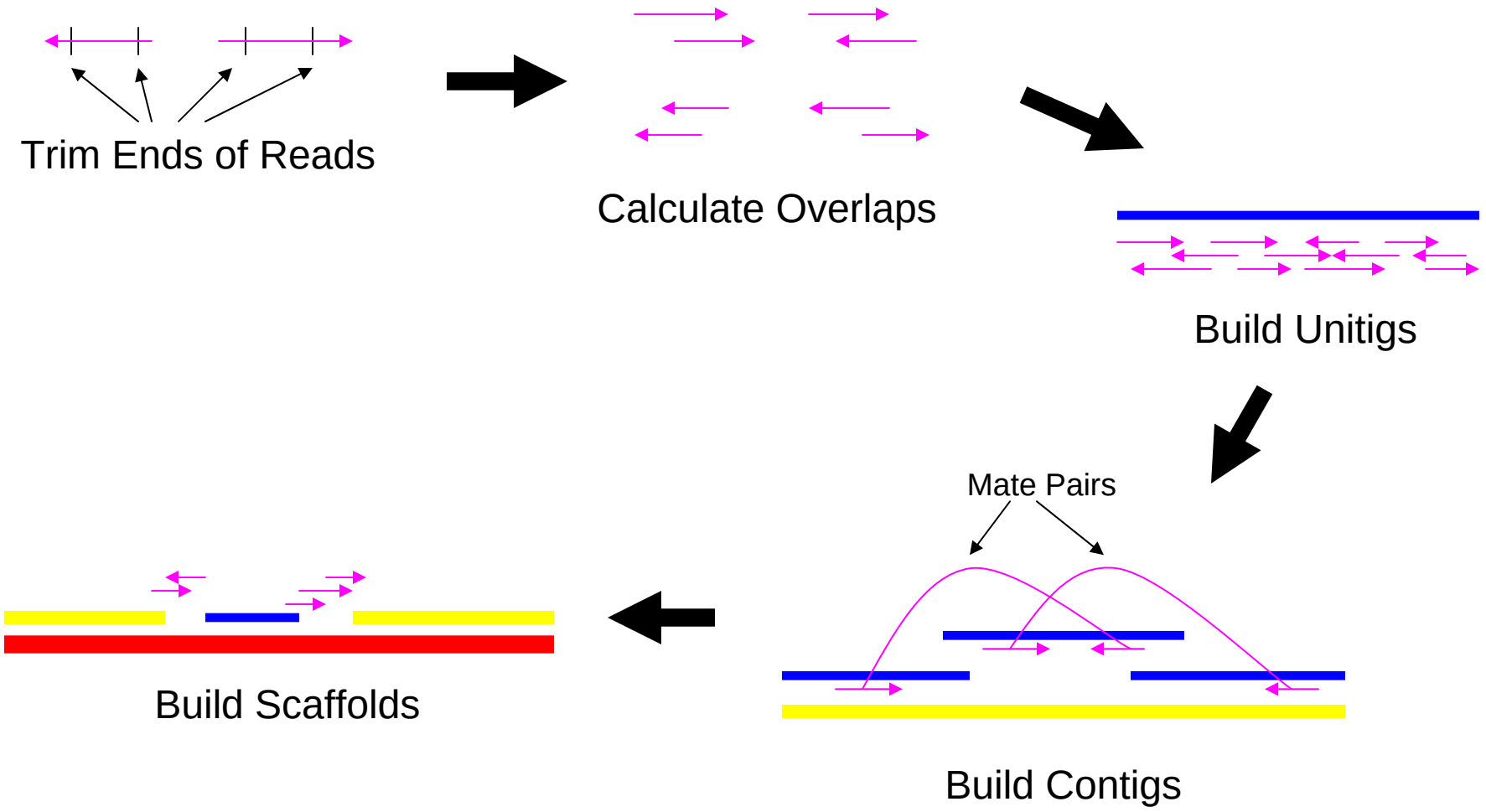
Advisor: Jim Yorke

Introduction to Genome Sequencing



Slide courtesy of Art Delcher

How Genomes Are Assembled



Running Celera Assembler at UMD

- Step 1: Download the data
- Step 2: Examine/Modify the data
- Step 3: Trim the data
- Step 4: Run UMD Overlapper
- Step 5: Run Celera Assembler

Step 1: Downloading the Data

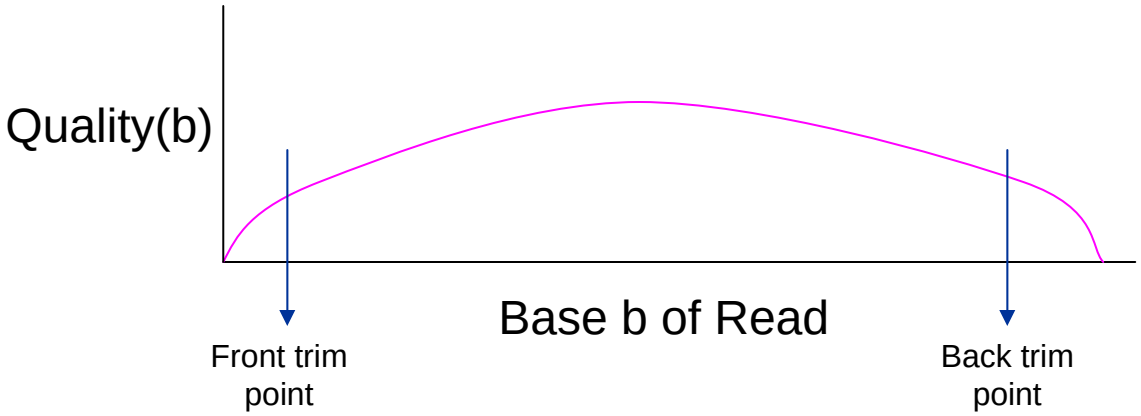
- The data is downloaded from the National Center for Biotechnology Information (NCBI) website
- Fasta files contain the sequence for each read
- Qual files contain a score for each base representing the probability that the base was recorded correctly

Step 2: Examine and Modify Data

- Determined that I had 31,240,954 reads in 126 libraries
- Checked to make sure the reads were in same order in the fasta and qual files
- Made sure reads were the same length in the fasta and qual files
- Eliminated duplicate reads (535,739)
- Set quality scores for base “N” to be zero

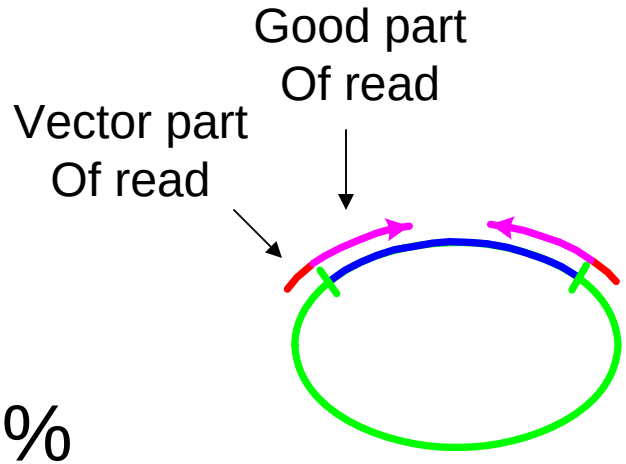
Step 3: Vector Trimming

We trim the read so that we have the longest possible sequence where each base has less than a 5% probability of being incorrect



Step 3: Vector Trimming (cont.)

- Eliminated reads with types “EST”, “FINISHING”, and “PCR”
- Trimmed the reads at both 5% error rate and 10% error rate (ended up using 5% trims)
- After trimming I had 29,318,901 reads and 21,095,993,892 total bases



Step 4: Run UMD Overlapper

Features of UMD Overlapper:

- Corrects sequencing errors and changes corresponding quality values
- Distinguishes between different copies of repeat regions
- Produces two sets of overlaps, “all” and “reliable”
- Does not require huge computational resources

Step 4: Run UMD Overlapper (cont.)

- I split the reads into five groups and then calculated overlaps between each pair of groups (and the group with itself)
- I ran five passes of the overlapper (each pass took about 12 hours to run)
- There was a second round of vector trimming during pass 3
- After running the overlapper, I had 29,078,173 reads and 232,162,427 overlaps

Step 5: Run Celera

- I am currently running Celera Assembler!
- Running Celera produces over a terabyte of data
- Running Celera on a large mammal requires about 5 days
- Celera will output statistics regarding scaffold size, contig size, number of reads used, and coverage

Parallelizing the Overlapper

Reasons to Upgrade the overlapper:

- The overlapper does not currently run well on large genomes
- The overlapper could be introduced into the Celera pipeline and improve assembly quality
- There are several parts of the overlapper that could be easily parallelized which would vastly improve the running time

Project Status

Fall 2007:

- Produce an assembly (in progress)

January 2008:

- Compare my assembly to the Broad assembly
- Work on parallelizing the overlapper

Spring 2008:

- Finish parallelizing the overlapper
- Produce consensus assembly

Acknowledgements

- Jim Yorke
- Aleksey Zimin
- Mike Roberts
- Roberts, M, Hunt, BR, Yorke JA, Bolanos R, and Delcher A. *A preprocessor for shotgun assembly of large genomes*. Journal of Computational Biology (2004). **11**(4), 734-52.
- Interpreting Celera Assembler Output (http://www.cbcb.umd.edu/research/castats.shtml)