Assembly of Next Generation Sequence Data

Catherine Eason (Wofford College)

Amit Upadhyay (University of Tennessee)

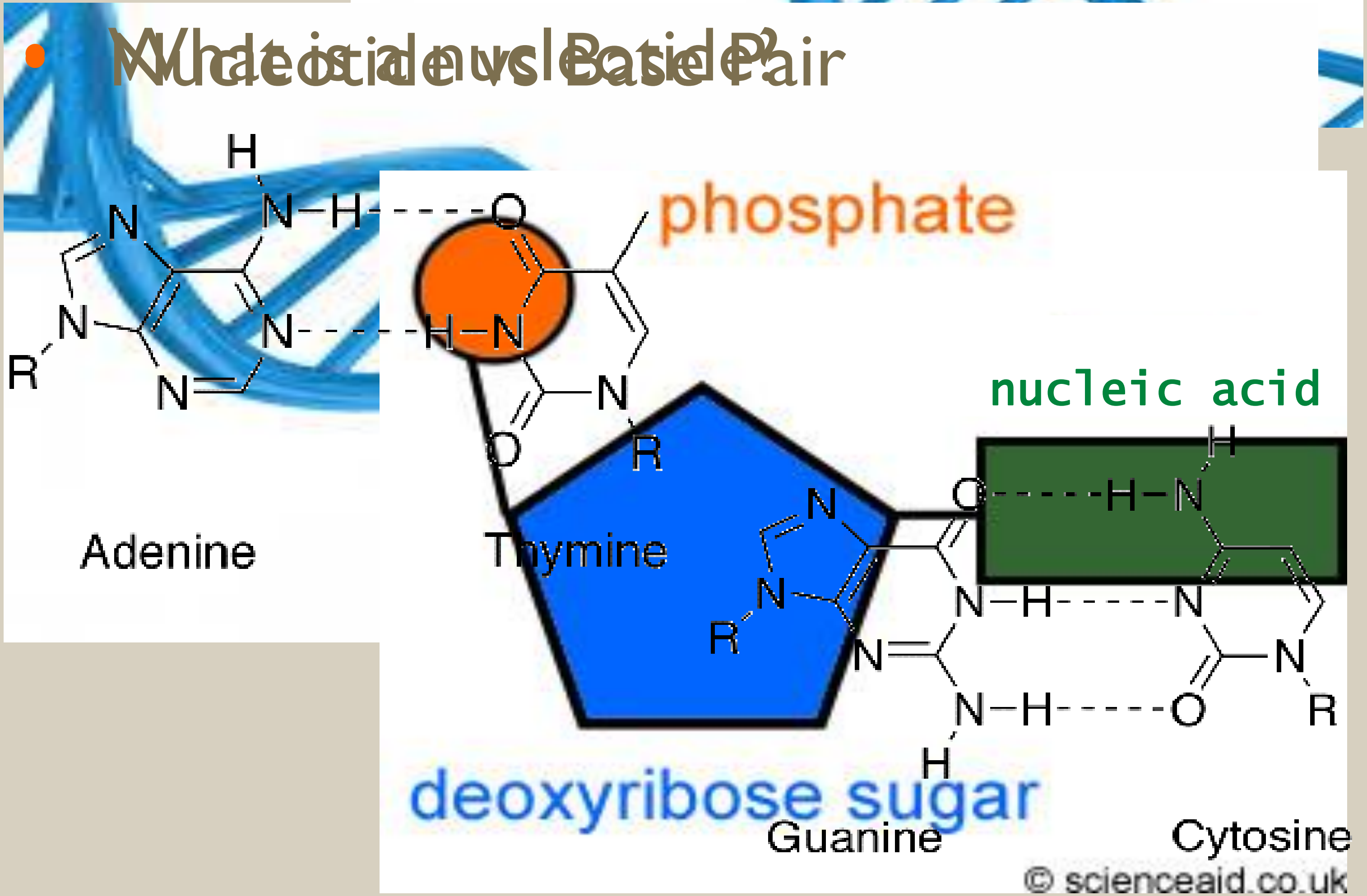
Bhanu Rekepalli (JICS)

Outline

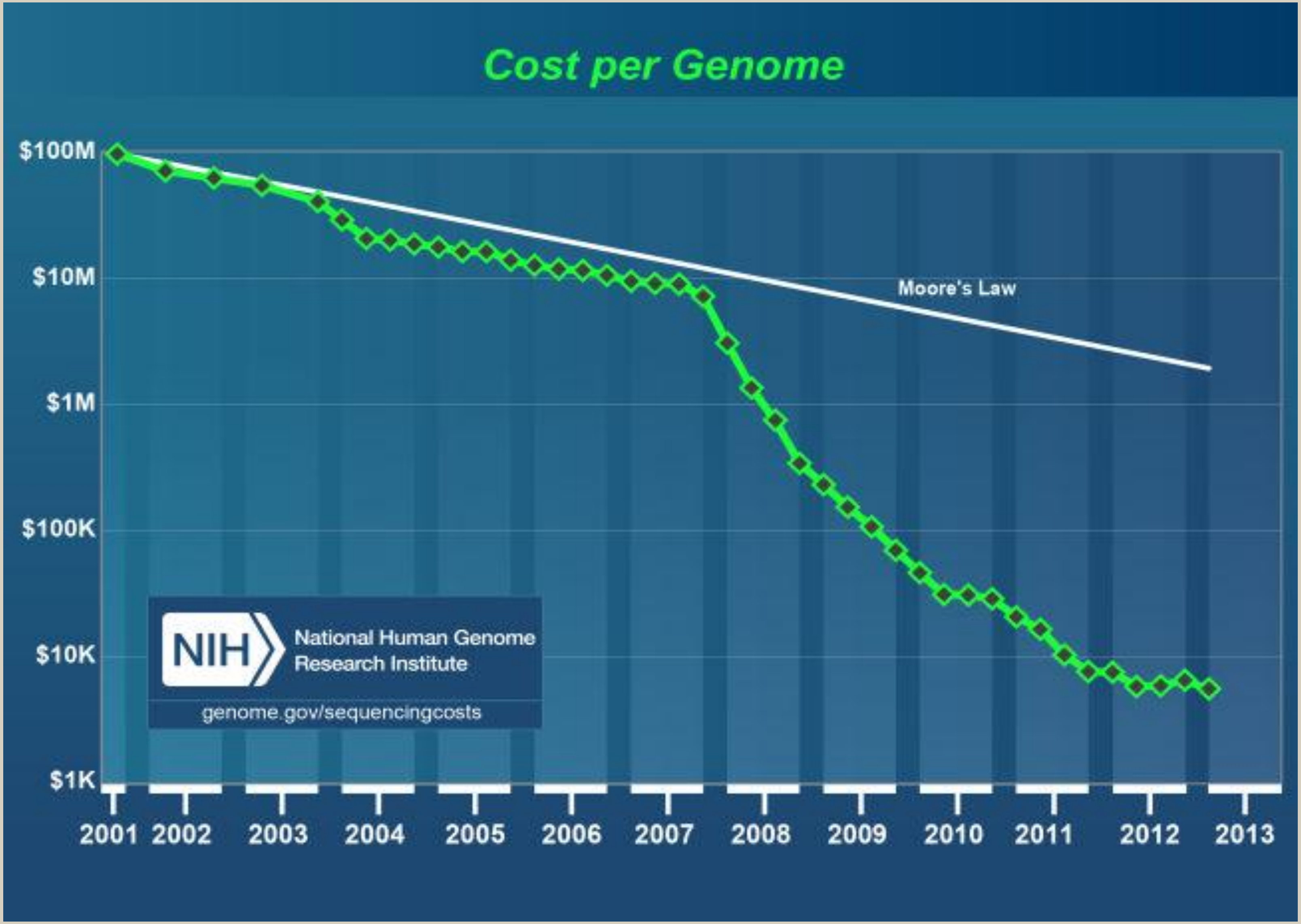
- DNA overview
- Background leading to problem
- Current Status in Assembly
- Methodology
- Results
- Conclusion/Future Work

DNA

- What is a nucleotide?
- Nucleotide vs Base Pair



NGS data


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts

Paired-End Sequences

- Sequencing from both ends concurrently (by Illumina)
- Allows for detection of small frame shift mutations
- Paired information must be kept together, and in correct order

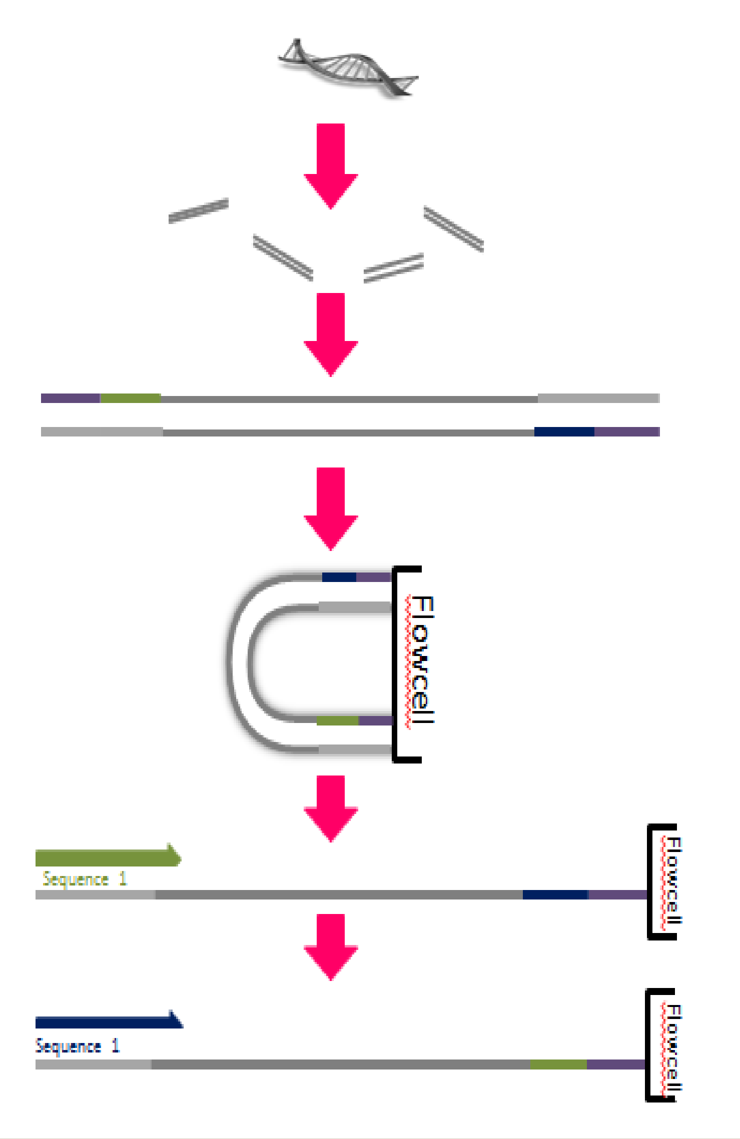


Diagram showing process of collecting paired-end reads. The genomic DNA is sequences into fragments which adaptors and primers are attached to (Green, Blue, and Purple ends). A cluster is formed and the sequences is read starting from both adaptors, producing the paired-end read.

Analysis Workflow

- Data is collected, now what?
  - Assembly
  - Analysis
  - Future Studies

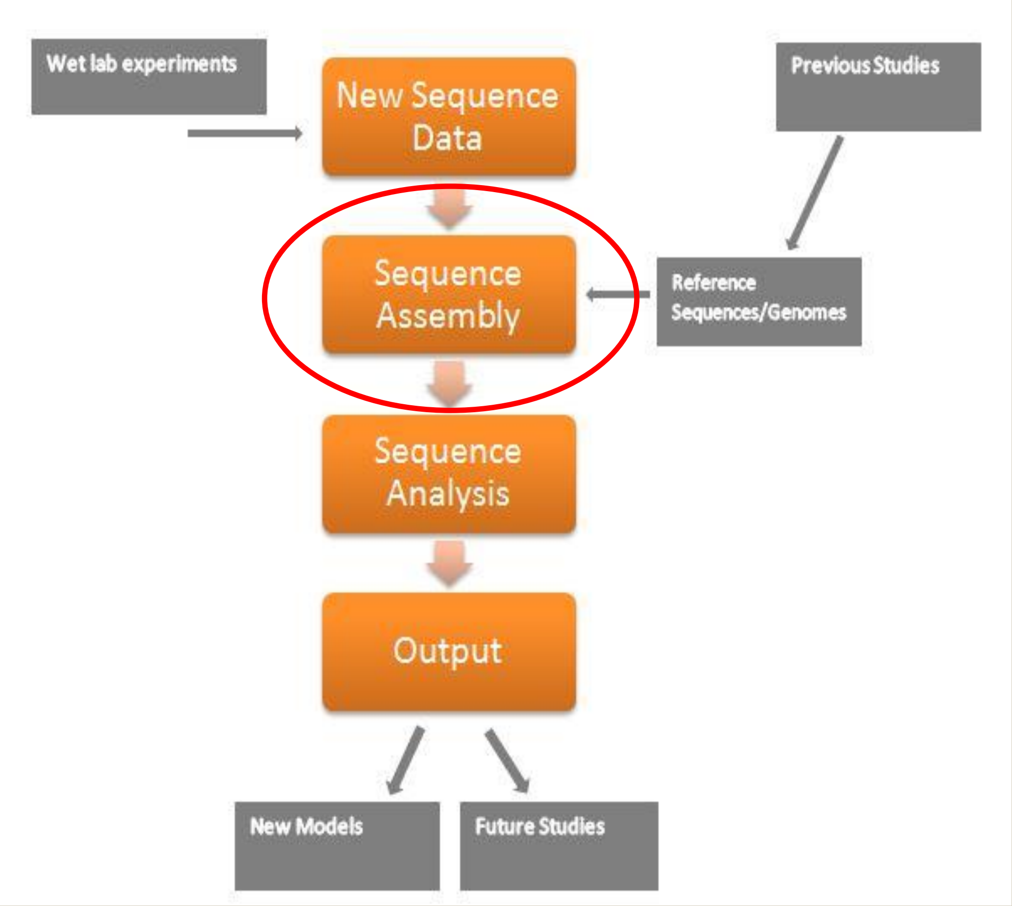


Diagram for the complete data analysis process. Orange rectangles are the actual analysis steps while the gray rectangles represent input from outside sources.

Assembly

- Quality Control
- Assembly
- Assembly Verification

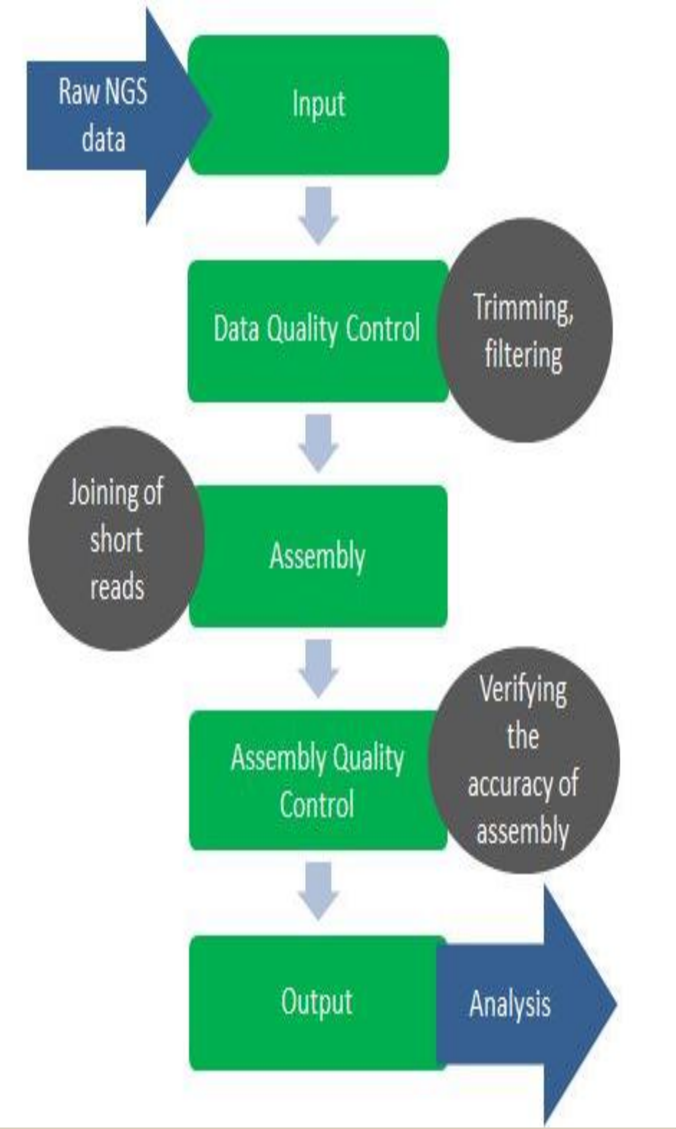
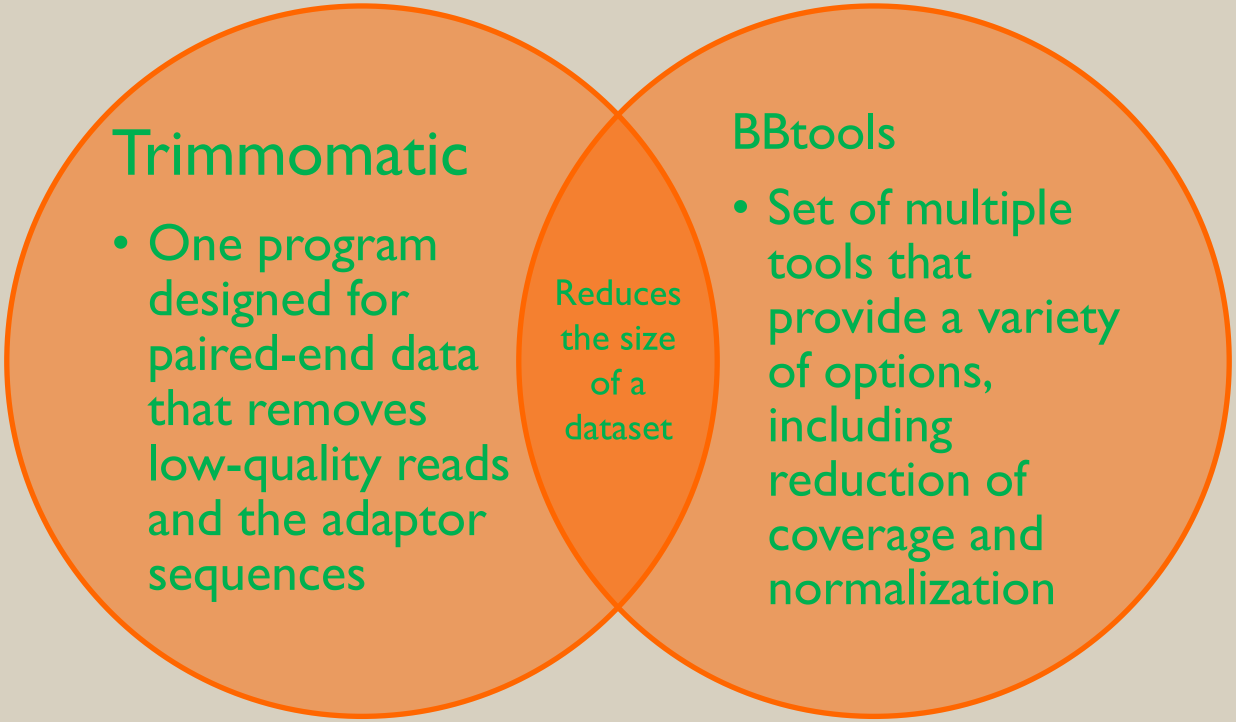


Diagram for the complete assembly process, beginning with raw sequence data. The assembled sequences must be checked for accuracy– a difficult step. Green rectangles are the steps, gray circles a short description. And blue arrows are steps that have their own process.

Trimmomatic vs. BBTools

- Quality Control
- Assembly
- Assembly Quality Control



Trimmomatic

- One program designed for paired-end data that removes low-quality reads and the adaptor sequences

Reduces the size of a dataset

BBtools

- Set of multiple tools that provide a variety of options, including reduction of coverage and normalization

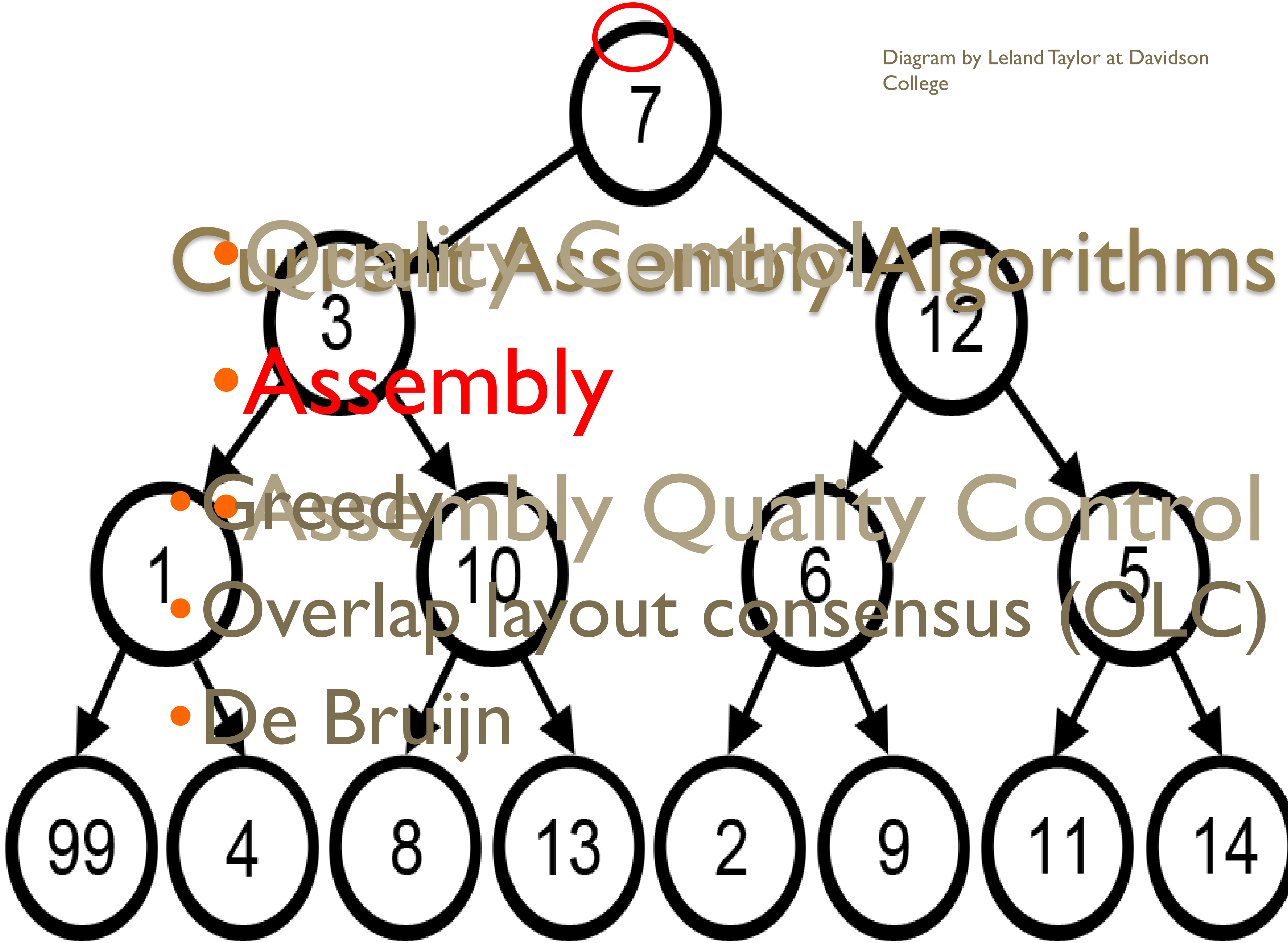


Diagram by Leland Taylor at Davidson College

Current Assembly Algorithms

- Quality Control
- Assembly
- Greedy
- Assembly Quality Control
- Overlap layout consensus (OLC)
- De Bruijn

SPAdes

MaSuRCA

Velvet

SOAPdenovo

ABySS

SPAdes

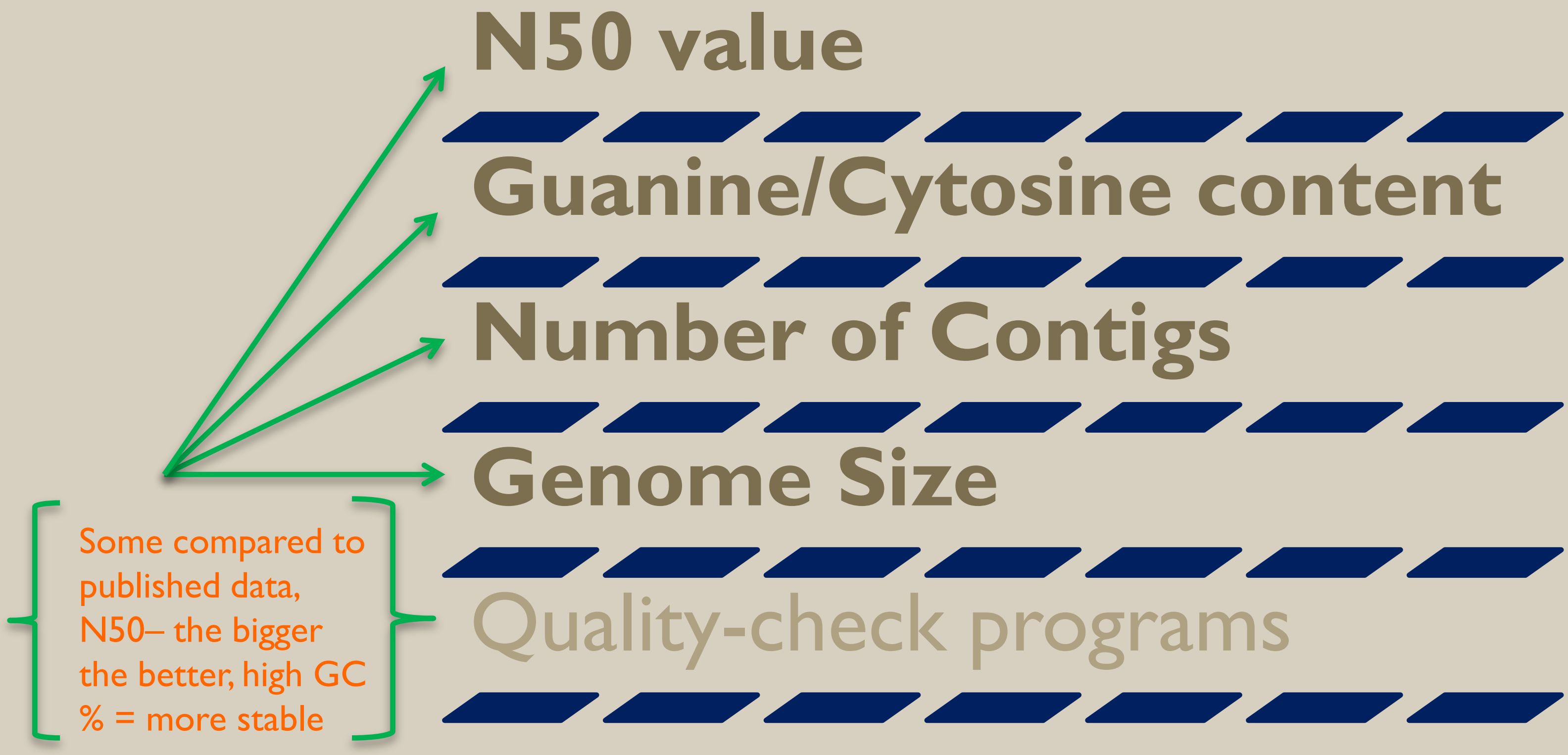
- New form of de Bruijn graph– *Multisized de Bruijn*
  - *Implements new "error correction" methods*
  - *Allows user to backtrack over graph construction process*
- Can detect "best" k-mer size (if desired)

SOAPdenovo

- *De novo* assembly of large, mammalian genomes
- Uses de Bruijn graph algorithm
  - Edges must be linked to existing sequence

QUAST and Statistics

- Quality Control
- Assembly
- Assembly Quality Control



N50 value

Guanine/Cytosine content

Number of Contigs

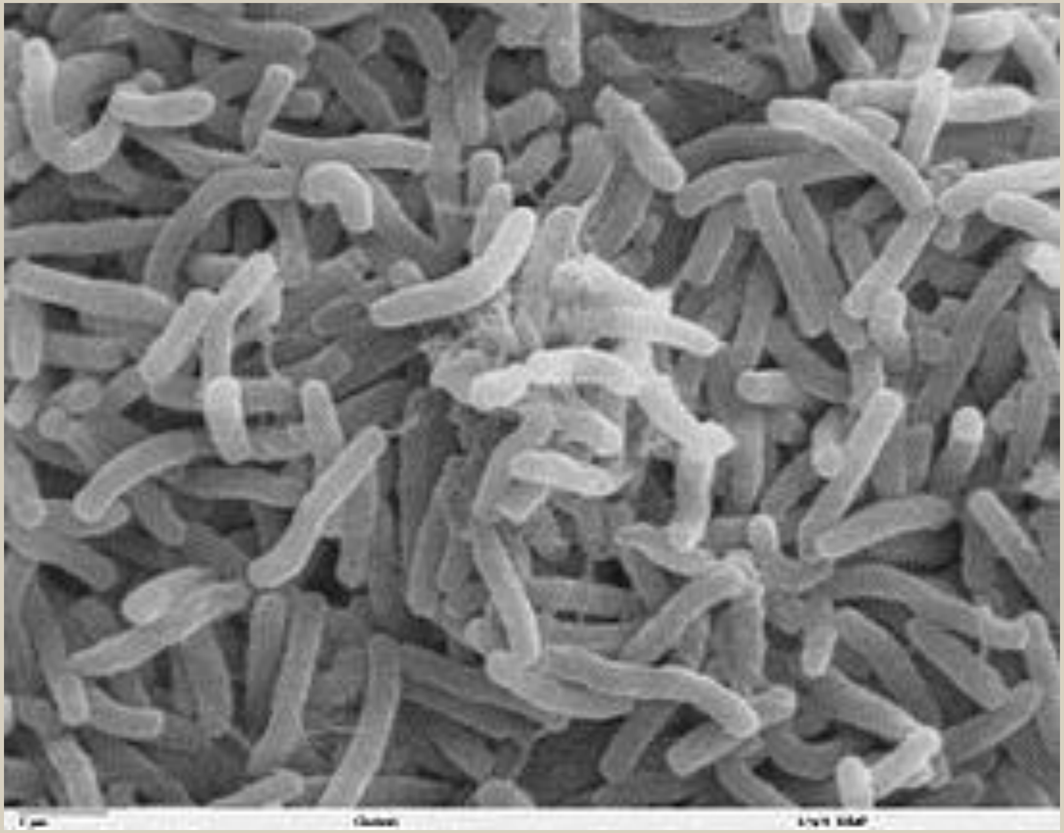
Genome Size

Quality-check programs

Some compared to published data, N50– the bigger the better, high GC % = more stable

Vibrio gazogenes

- 36 chromosomes
- Genome size?



Picture of *V. cholera* bacteria. Closely related to *V. gazogenes*

Results

- Using Trimmomatic (quality of read)

SPAdes (Trimmomatic)				
Kmer size	# of Contigs	Genome Size	N50	GC %
21	514	4,430,394	17,374	45.27
33	282	4,467,765	54,782	45.27
55	215	4,496,327	68,126	45.27
71	120	4,555,395	246,573	45.32
Subset 51	201	4,468,133	61,386	45.30
Subset 61	193	4,485,523	68,843	45.31
Subset 71	180	4,499,332	79,631	45.32
Subset 81	173	4,510,565	88,093	45.33
Subset 91	88	4,545,153	262,031	45.36

Table for the assembly of Trimmomatic trimmed data through SPAdes showing number of contigs , genome size , N50, and GC content statistics for k-mer sizes 21,33,55,71 and a random 50% subset of data's statistics for k-mer sizes 51,61,71,81, and 91.

SOAPdenovo2 (Trimmomatic)				
Kmer Size	# of Contigs	Genome Size	N50	GC %
21	16	11,398	690	42.96
33	17	11,766	690	41.00
55	1,385	968,669	685	46.87
71	444	4,448,857	18,563	45.33
Subset 51	1,481	4,321,140	4,296	45.39
Subset 61	309	4,459,372	29,329	45.30
Subset 71	206	4,481,934	55,249	45.30
Subset 81	172	4,499,317	75,768	45.32
Subset 91	159	4,519,076	100,098	45.34

Table for the assembly of Trimmomatic trimmed data using SOAPdenovo2. showing number of contigs , genome size , N50, and GC content statistics for k-mer sizes 21,33,55,71 and a random 50% subset of data's statistics for k-mer sizes 51,61,71,81, and 91.

Results

- Using BBtools (bbnorm and bbtrim)

SPAdes (BBtools)				
Kmer Size	# of Contigs	Genome Size	N50	GC %
21	506	4,409,861	17,893	45.29
33	263	4,445,712	49,223	45.28
55	190	4,474,737	65,281	45.30
71	106	4,532,943	167,499	45.31

Table for the assembly of Bbtool trimmed data through SPAdes showing number of contigs, genome size, N50, and GC content statistics for k-mer sizes 21, 33, 55, and 71.

SOAPdenovo2 (BBtools)				
Kmer Size	# of Contigs	Genome Size	N50	GC %
21	770	4,389,210	9,940	45.29
33	379	4,430,953	24,090	45.30
55	202	4,467,392	62,696	45.30
71	169	4,488,672	81,399	45.35

Table for the assembly of Bbtool trimmed data through SOAPdenovo2. show number of contigs, genome size, N50, and GC content statistics for k-mer sizes 21, 33, 55, and 71.

Conclusions

- Trimmomatic: no negative effect on assembly process
- Genome size ~4.5 million bp

Future Goals

- Collective scripts for all four aspects of NGS pipeline project
  - Genome assembly
  - Genome annotation
  - RNA-seq
  - Variant calling
- Collective script for all steps of assembly
- Web Interface for ease of access

References

- Bankevich, Anton, et al. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology* 19.5 (2012): 455-77. Web.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.
- Dikow, Rebecca B., and William Leo Smith. "Genome-level Homology and Phylogeny of Vibrionaceae (Gammaproteobacteria: Vibrionales) with Three New Complete Genome Sequences." *BMC Microbiology* 13.1 (2013): 80. Web.
- Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. "QUAST: Quality assessment Tool for Genome Assemblies." *Bioinformatics* 29.8 (2013): 1072-075. Web.
- Li, Yingrui, Yujie Hu, Lars Bolund, and Jun Wang. "State of the Art De Novo Assembly of Human Genomes from Massively Parallel Sequencing Data." *Human Genomics* 4.4 (2010): 271. Web.
- Luo et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 2012 1:18

Acknowledgments

I would like to thank the CSURE program funded by the NSF and partnered with JICS, ORNL, and the University of Tennessee for the opportunity to work on this project.

Questions