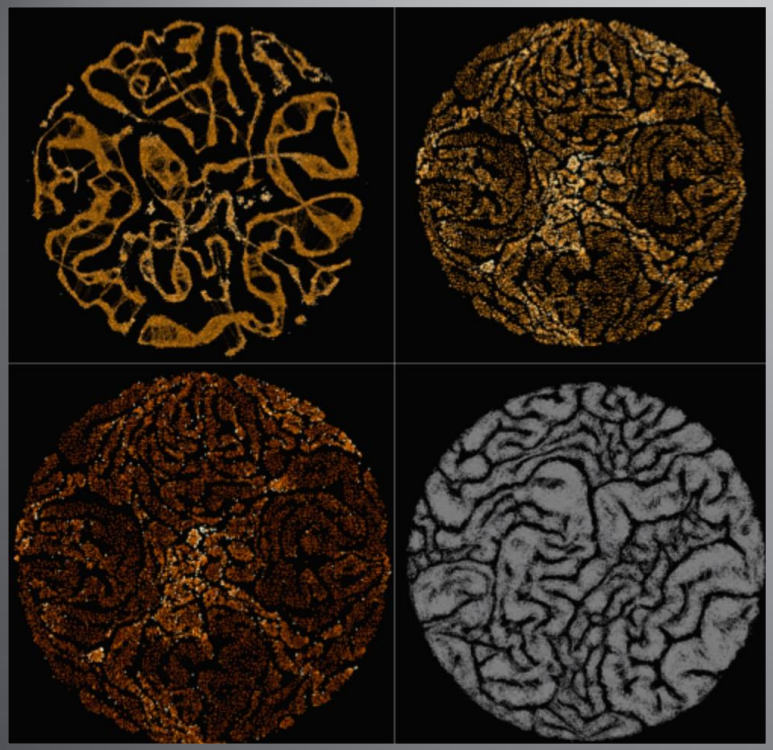


The “Art” of Shotgun Assembly

PacBio Bioinformatics Developer Meeting, Aug 26, 2015

FIND MEANING IN COMPLEXITY

Large Genome Assembly: A Bit of Every Everything

Information Technology

- Grid computing
- Cloud computing
- Distributed file system
- Data management

Software Engineering

- Open source development
- Modularization
- Workflow management
- yment

Sequence Data Analytics

- Data quality assessment
- Genome repeat content estimation
- Algorithm parameter optimization
- Effective data reduction

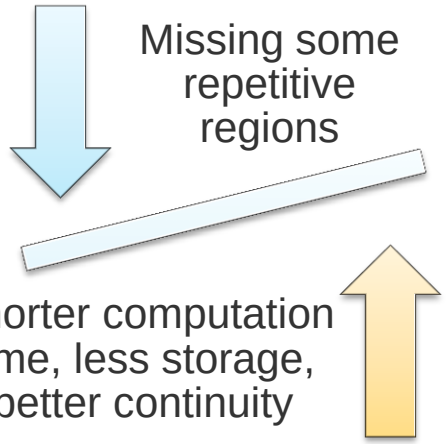
natics Algorithms

- overlappers
- Assembly graph construction and reduction
- Metadata tracking and management
- Understand the trade-off between various factors

Current Challenges And Solutions (for Falcon)

- Reduce false overlapping caused by repeats
  - What is the right amount of data for XYZ?
  - Read length cutoff?
  - Sensitivity and specificity?
  - One month or 48 hours computation?
- Find and understand “good rules” for assembly graph reduction
  - Overlap filtering
  - Remove “repeat induced bridging unitigs”

Main challenge: balance the trade-off



A better understanding of genome **repeat structures** is the key!!

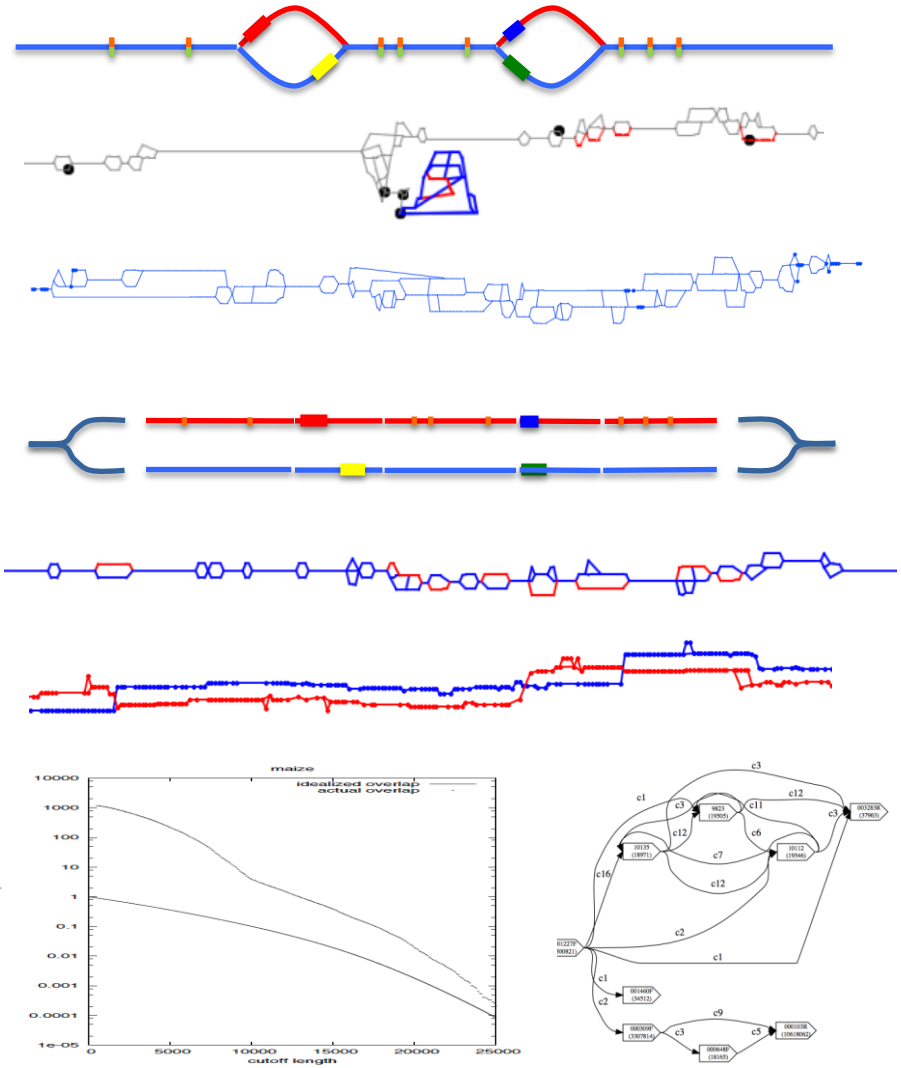
- Diploid genome “unzipping”
  - New concept to adapt (`https://speakerdeck.com/jchin`)
  - Need to track every raw read
  - “Augmented alignment” needed for Quiver polishing

Challenges

- More complicated data model
- Evaluation and validation can be tricky

Near Future Opportunities

- Assembly graph visualization GUI for manual refinement or assembly error detection
- Estimate optimized assembly algorithm parameters by analyzing small sample for “**REPEATREPEAT**”
- Multi-step assembly to provide full pictures including all repeat levels
- Algorithm and efficiency improvement for polyploid genomes
- Combination of other types of data tc improve the final assembly quality



PACIFIC
BIOSCIENCES®