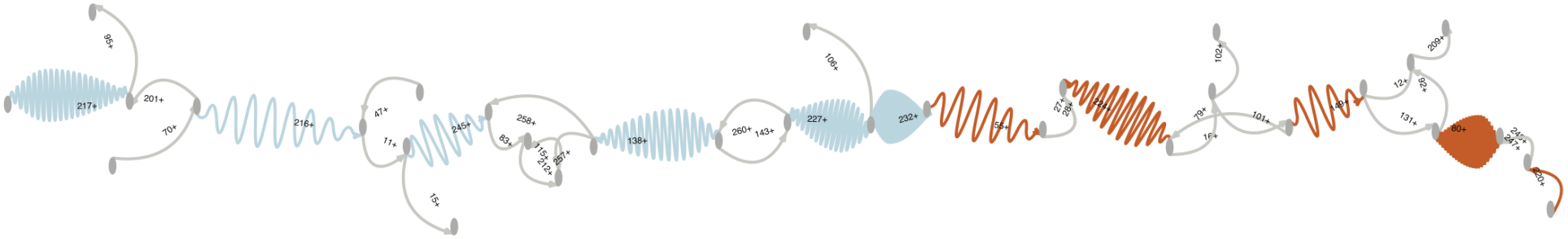
VISUALIZING ASSEMBLY FOR NGS

Cydney Nielsen

BC Cancer Agency, Genome Sciences Centre

Vancouver, Canada



Genome Sequencing

Sample preparation

↓

Physical sequencing

↓

Assembly

Sequencing Technologies

First Generation	Second Generation	Third Generation
 ABI 3730	 Roche/454	 Pacific Biosciences SMRT
	 Life Technologies SOLiD	 Oxford Nanopore MinION
	 Illumina HiSeq	
	. . .	. . .

Sequencing Technologies

Second Generation

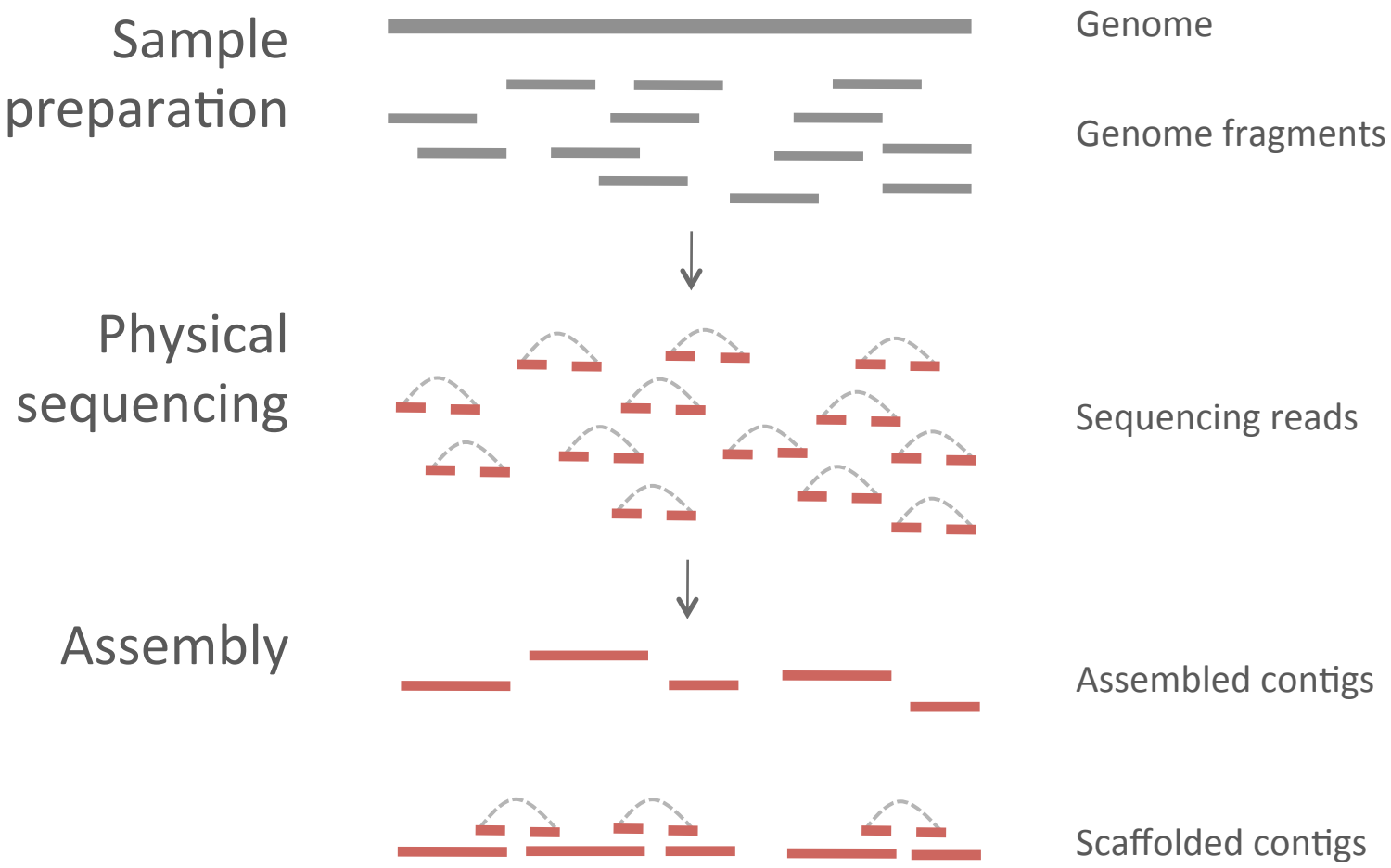
Roche/454

Life Technologies SOLiD

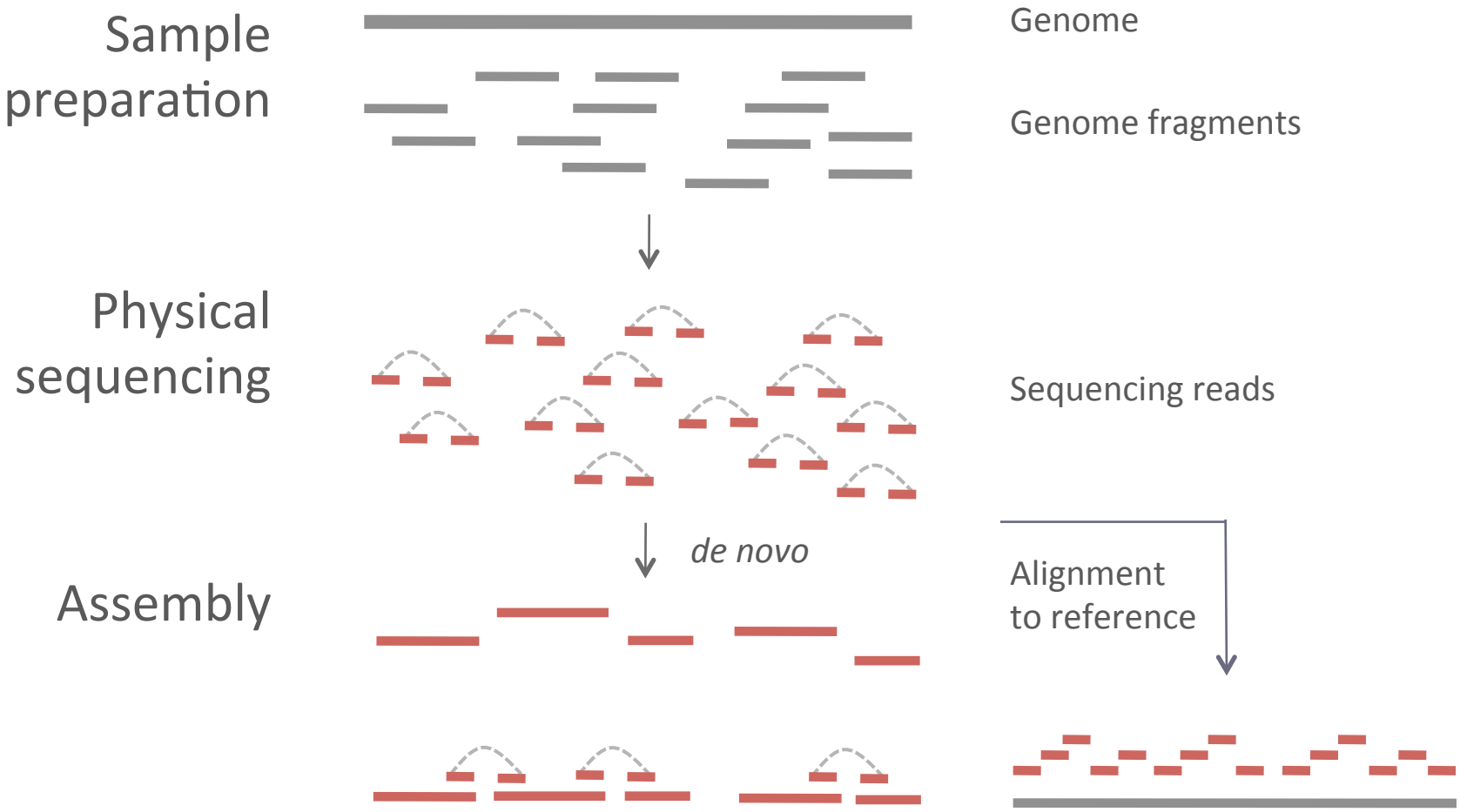
Illumina HiSeq

Produce sequencing reads that are tens to hundreds of bases long

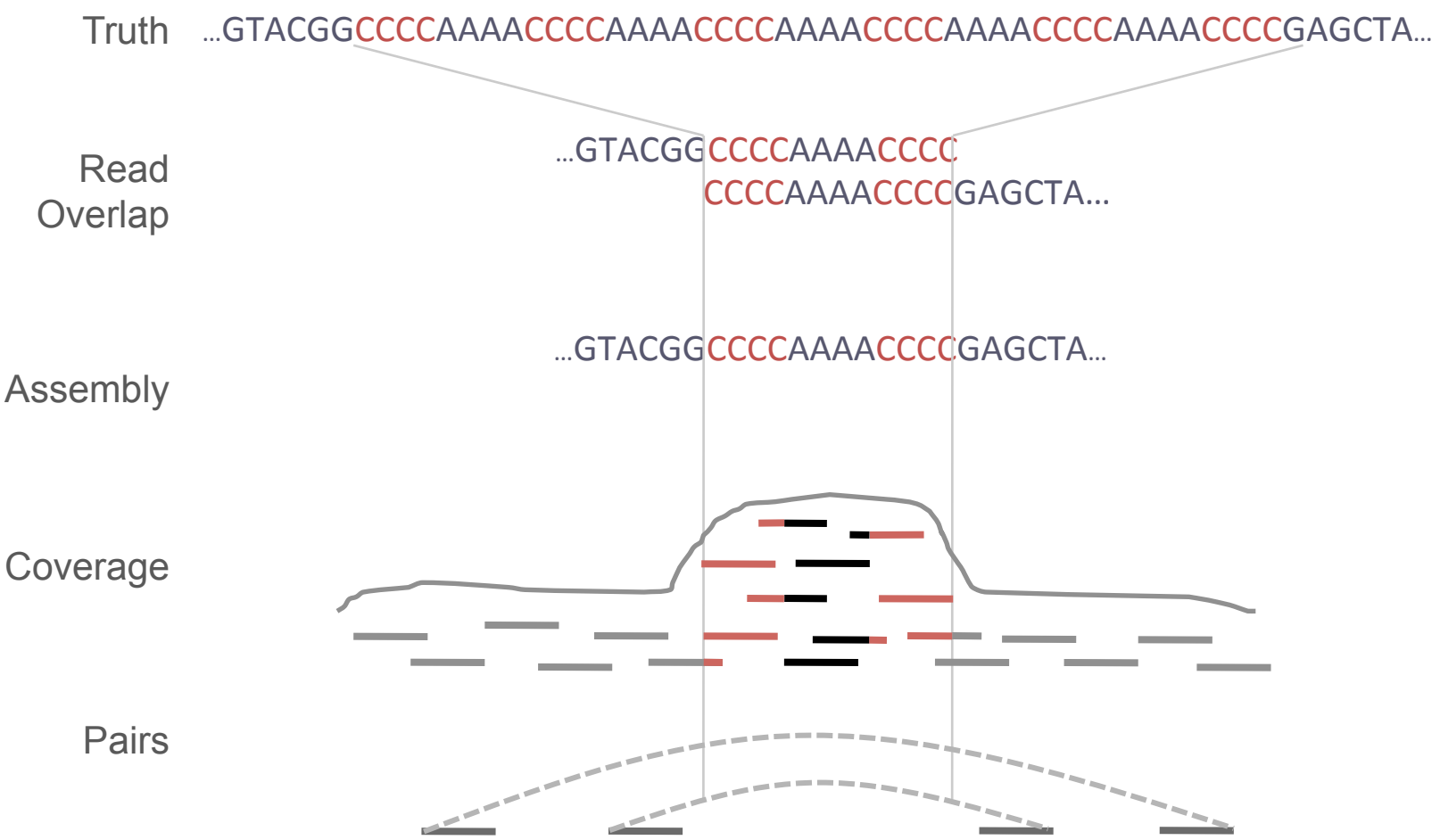
Genome Sequencing



Genome Sequencing



Genome Assembly Challenges



Adapted from Schatz *et al.*
Briefings in Bioinformatics, 2011

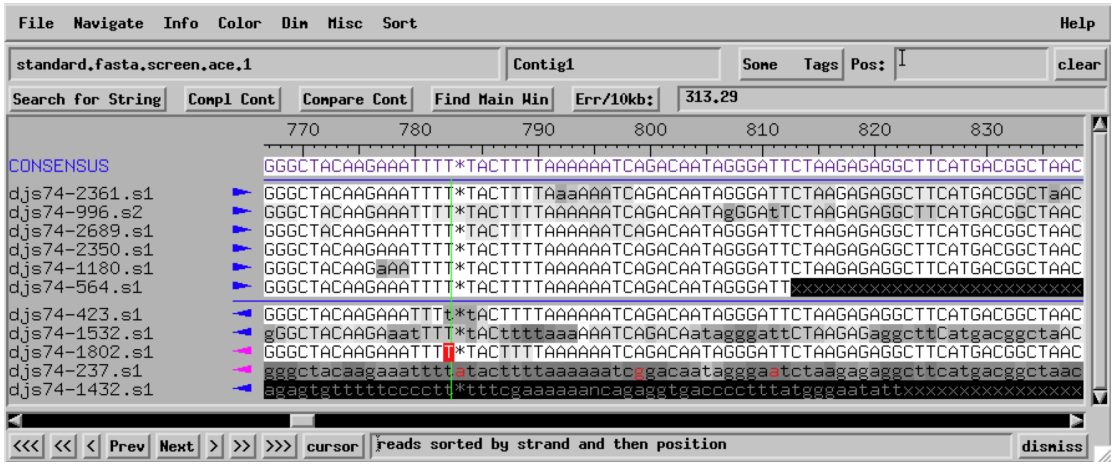
Assembly Visualization: Applications

- Finishing
  - involves closing gaps, correcting misassemblies and improving the error probabilities of consensus bases
  - often done manually; can be labour intensive and costly
- Algorithm development and iteration
  - often valuable to inspect potential assembly errors
- Investigation of structural variation
  - Detailed analysis of biologically relevant events in resequencing data

Task 1 | Examining Nucleotide-Level Discrepancies

Task 1 | Examining Nucleotide-Level Discrepancies

Aligned Reads Window



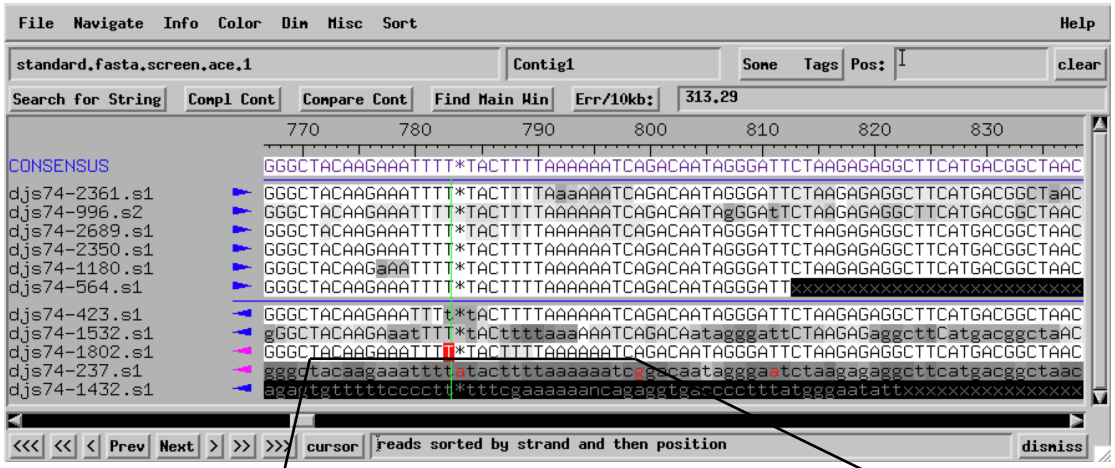
- Initially designed for Sanger sequencing (8-10x coverage of 500- to 1000-base reads)
- Introduction of quality values (log transformed error probabilities) was a significant contribution in providing an objective criterion to guide finishing

Consed

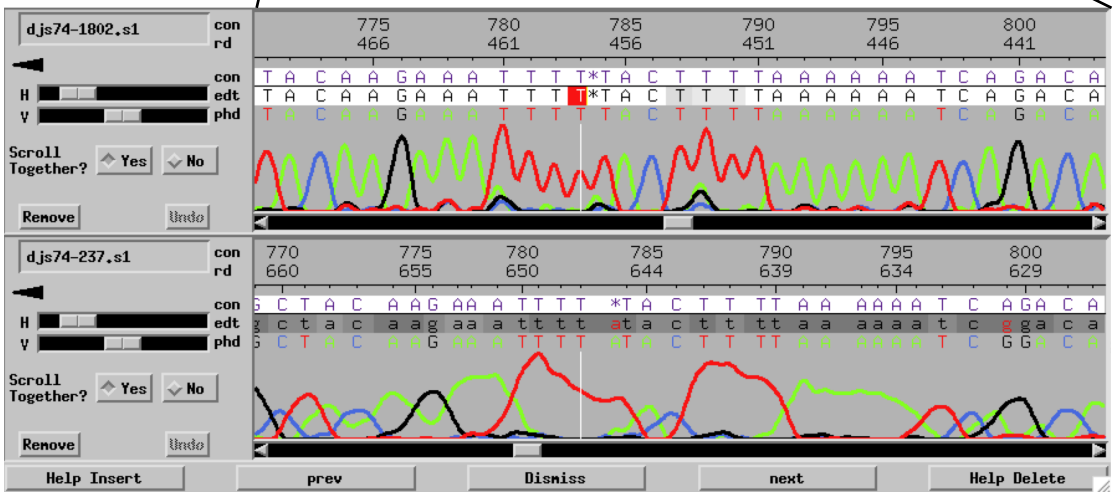
David Gordon and Phil Green

Task 1 | Examining Nucleotide-Level Discrepancies

Aligned Reads Window



Trace Window

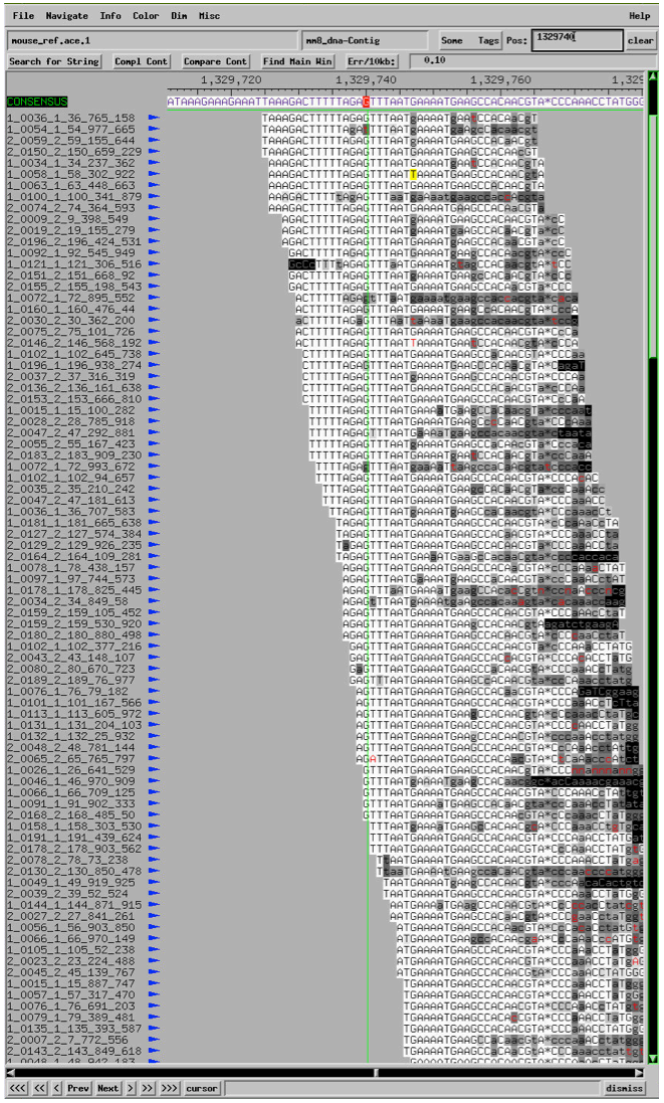


Consed

David Gordon and Phil Green

Task 1 | Examining Nucleotide-Level Discrepancies

Aligned Reads Window



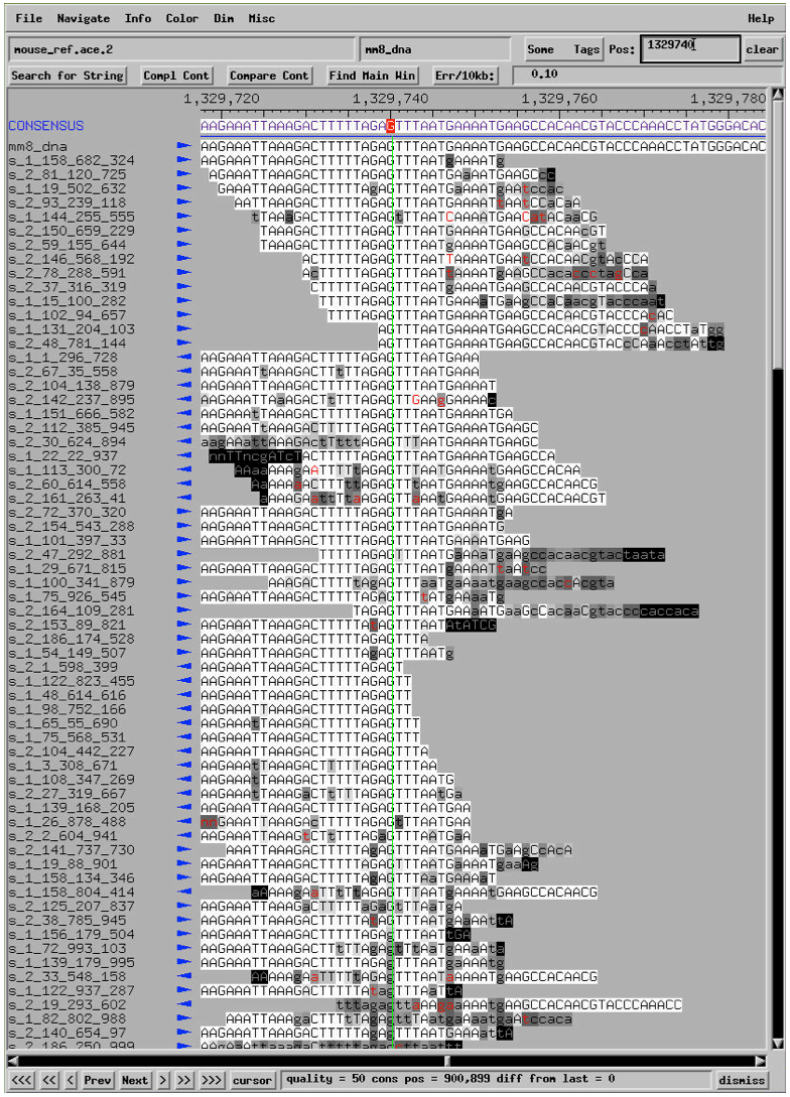
30-100x coverage of 50- to 100-base reads from second generation technologies pose a challenge

Consed

David Gordon and Phil Green

Task 1 | Examining Nucleotide-Level Discrepancies

Aligned Reads Window



Sorting by quality value is a useful guide

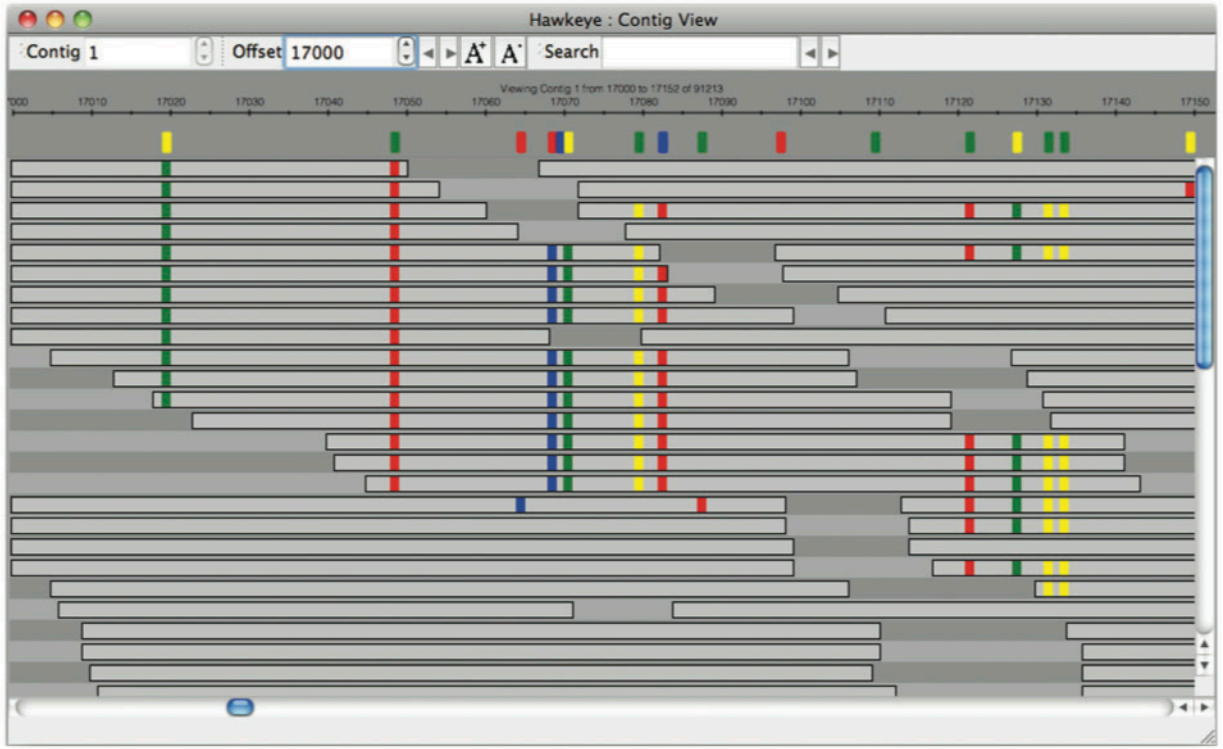
No longer need to inspect raw data underlying an individual read

Consed

David Gordon and Phil Green

Task 1 | Examining Nucleotide-Level Discrepancies

Contig View

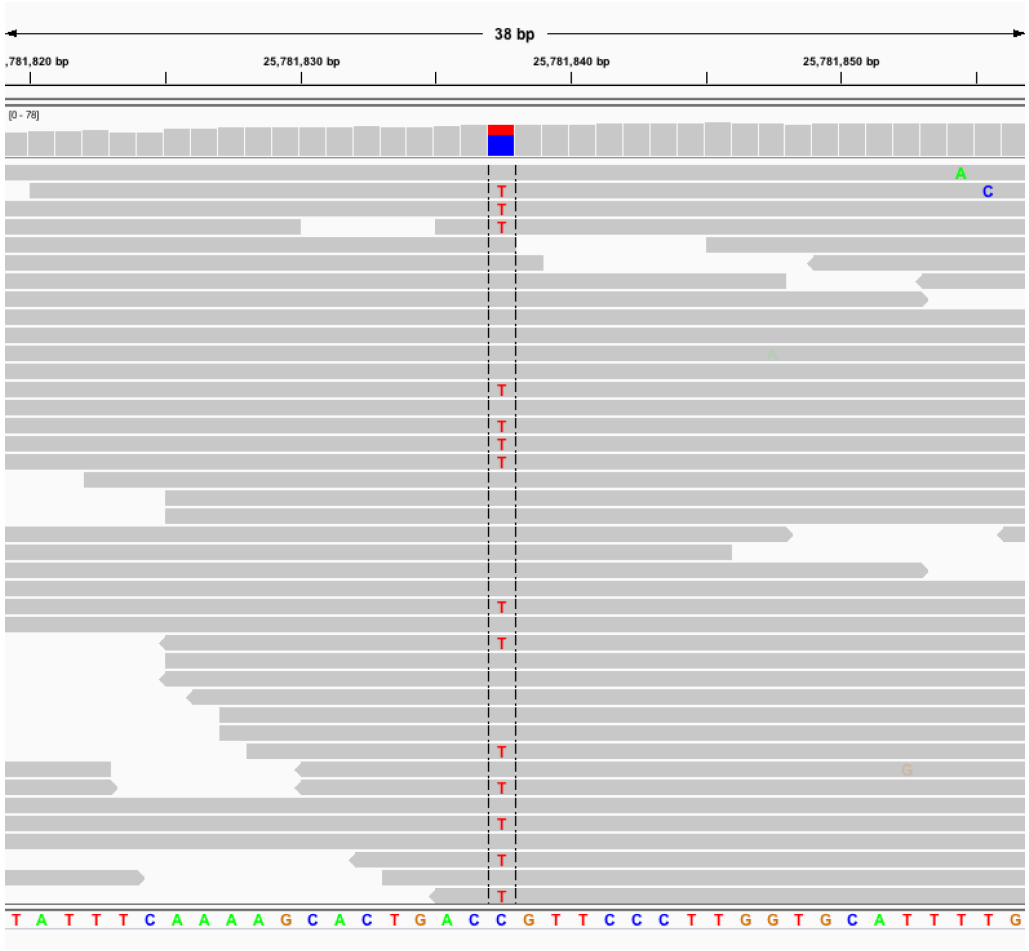


- High quality discrepancies can be indicative of misassemblies or nucleotide variants

Hawkeye

Schatz *et al.*, 2007; 2011

Task 1 | Examining Nucleotide-Level Discrepancies



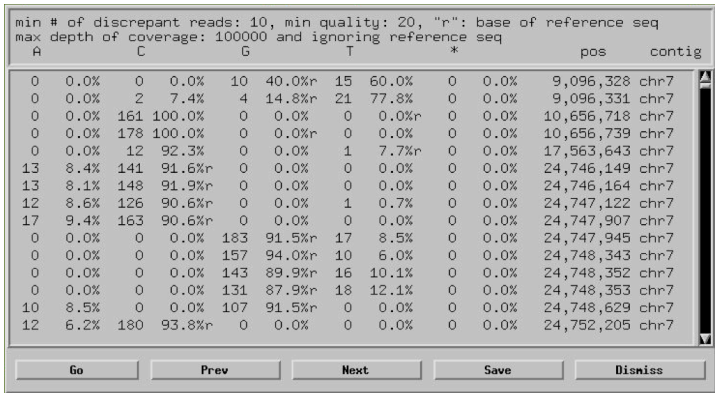
Integrative genomics viewer (IGV)

Robinson *et al.*, 2011

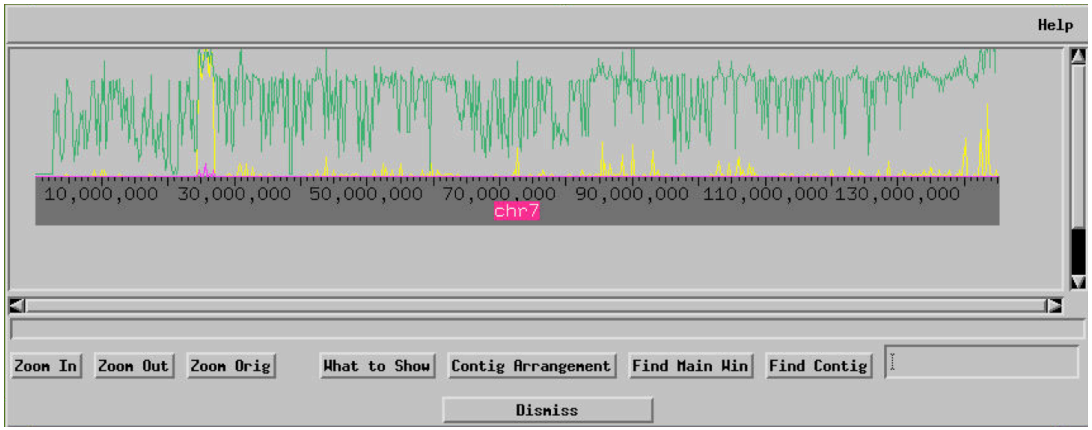
Task 1 | Examining Nucleotide-Level Discrepancies

Navigating to discrepant positions

Highly Discrepant Positions Table



Assembly View



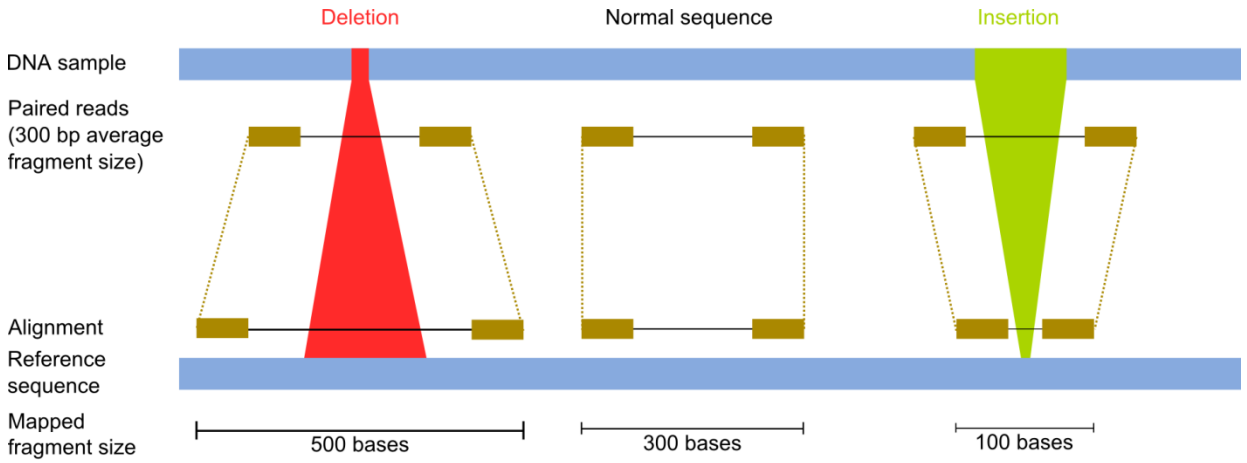
Coverage (green) Indels (magenta) Non-indel discrepancies (yellow)

Consed

David Gordon and Phil Green

Task 2 | Inspecting Deviant Read Pairs

Task 2 | Inspecting Deviant Read Pairs



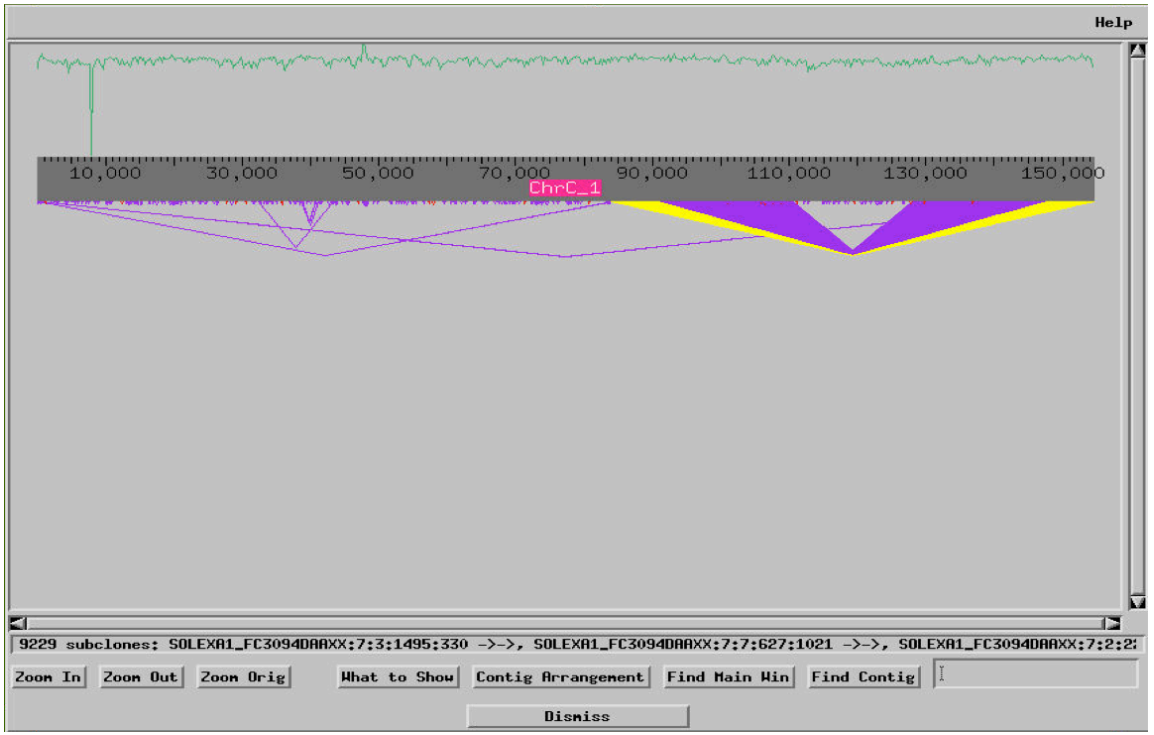
From Manske and Kwiatkowski,
Genome Research, 2009

Task 2 | Inspecting Deviant Read Pairs

Indicate deviations as overlaid arcs

Assembly View

Filtered inconsistent pairs shown



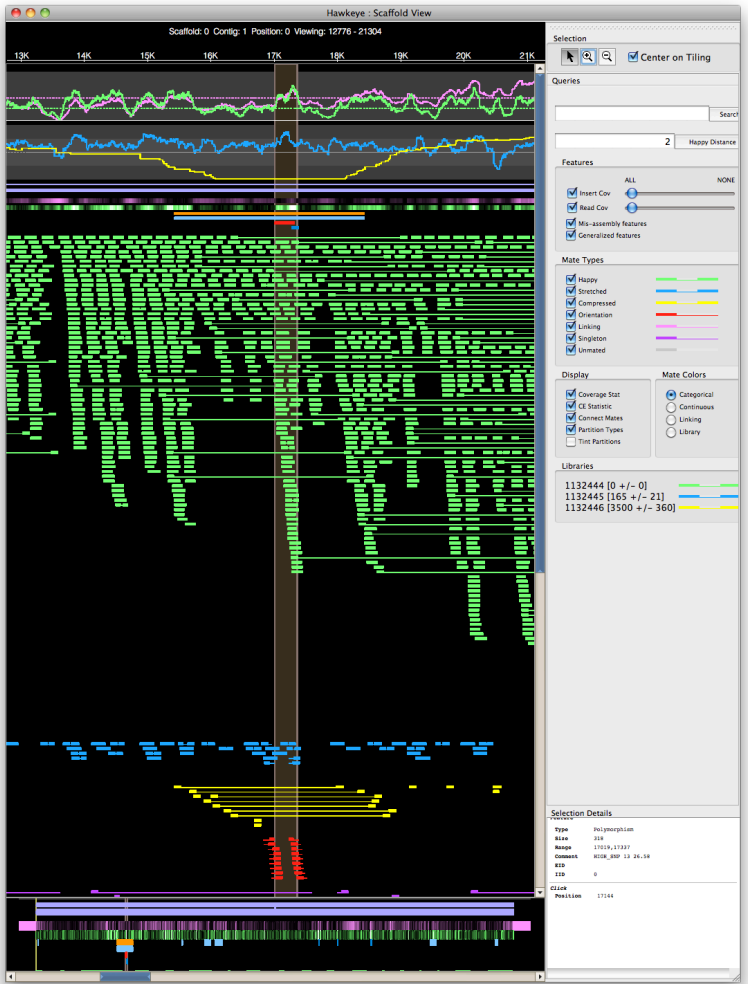
Consed

David Gordon
and Phil Green

Task 2 | Inspecting Deviant Read Pairs

Indicate deviations with colour and clustering

Scaffold View



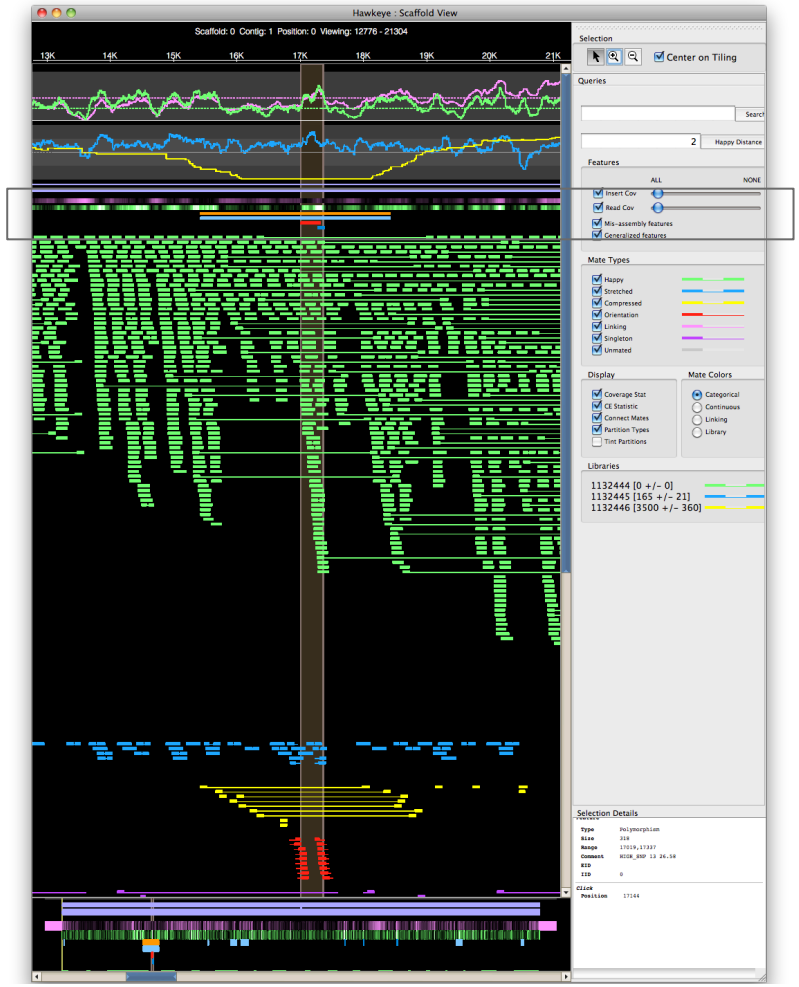
Hawkeye

Schatz *et al.*, 2011

Task 2 | Inspecting Deviant Read Pairs

Indicate deviations with colour and clustering

Scaffold View



Evidence for a misassembly:

Orange bar: suspicious region (by AMOSvalidate)

Light blue bar: compression

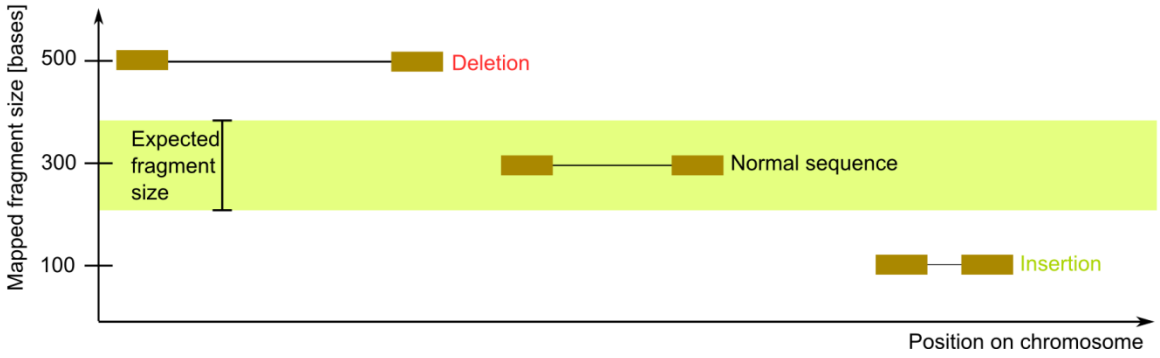
Red: high density of heterogeneous SNPs

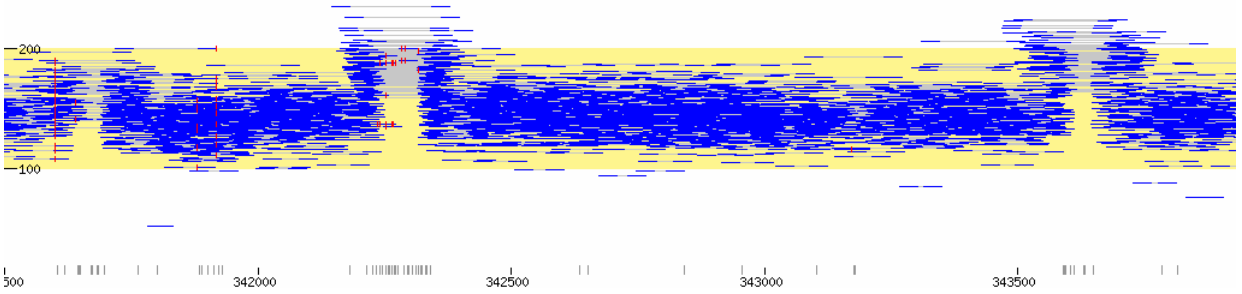
Hawkeye

Schatz *et al.*, 2011

Task 2 | Inspecting Deviant Read Pairs

Use the y-axis to indicate insert size





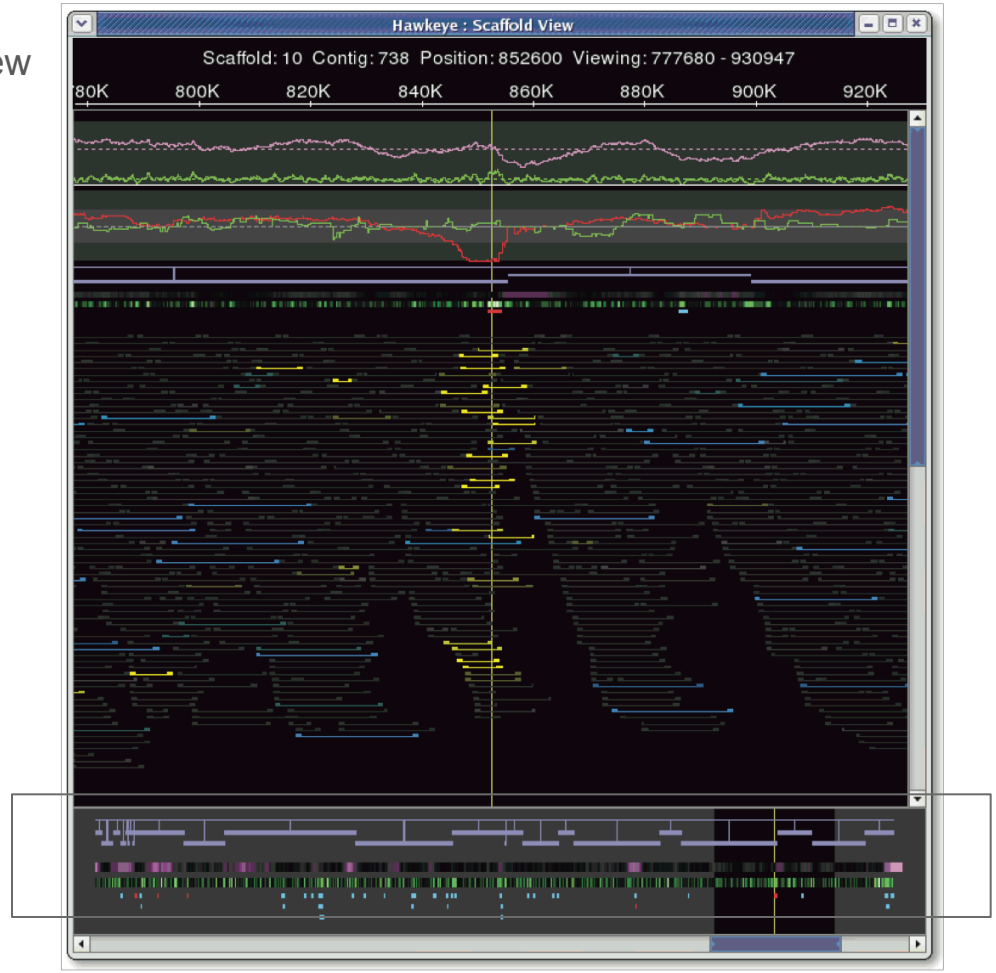
LookSeq

Manske and Kwiatkowski, 2009

Task 3 | Investigating Contig Connectivity

Task 3 | Investigating Contig Connectivity

Scaffold View



Contig order in the scaffold displayed in overview panel

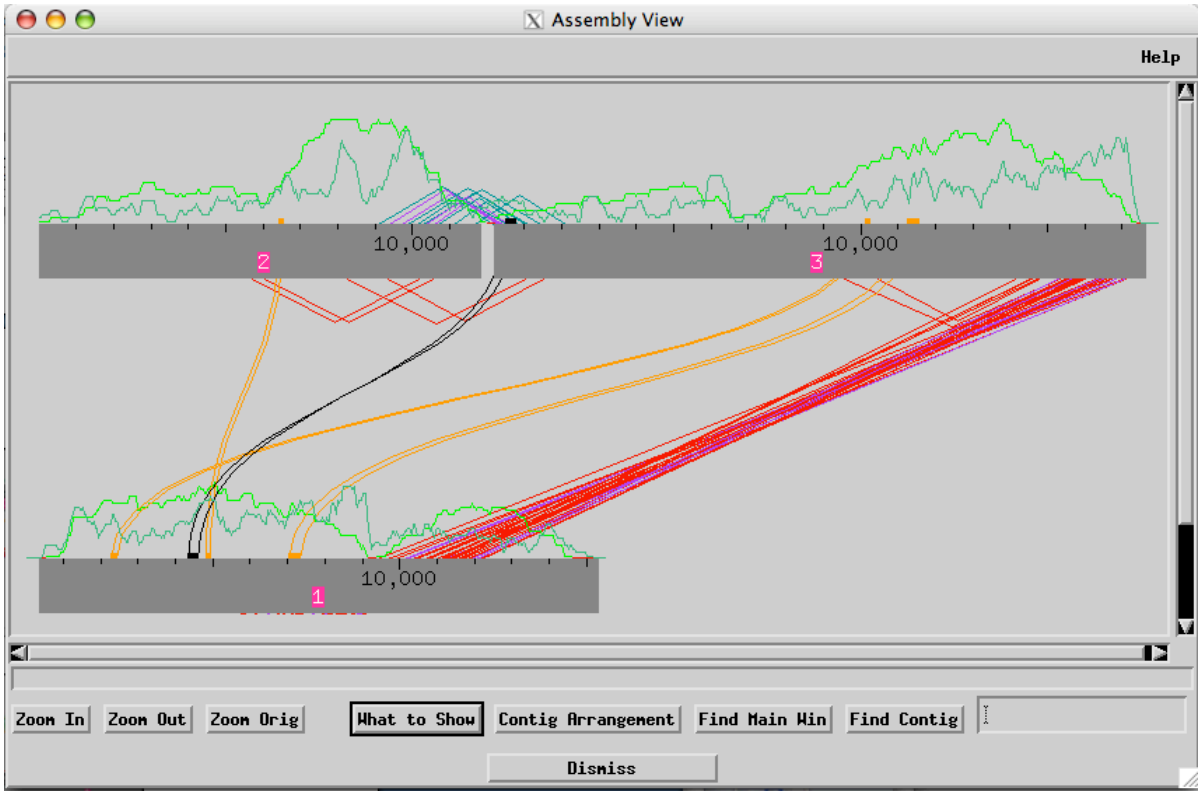
Hawkeye

Schatz *et al.*, 2007

Task 3 | Investigating Contig Connectivity

Inconsistent read pairs (red) indicate a misassembly

Assembly View



Read coverage: line plots

Read pairs: angled lines

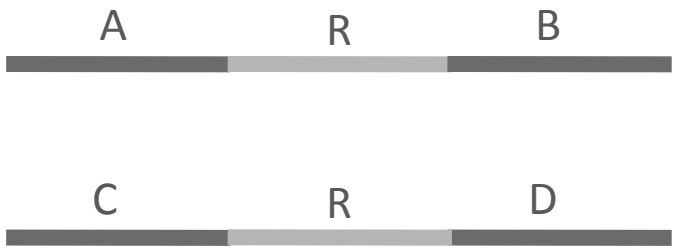
Sequence similarity: curved lines

Consed

David Gordon
and Phil Green

Task 3 | Investigating Contig Connectivity

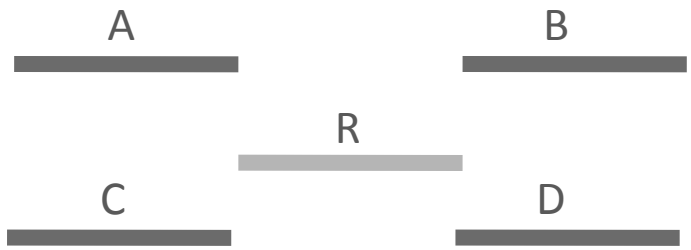
Source DNA



Repeat R

Task 3 | Investigating Contig Connectivity

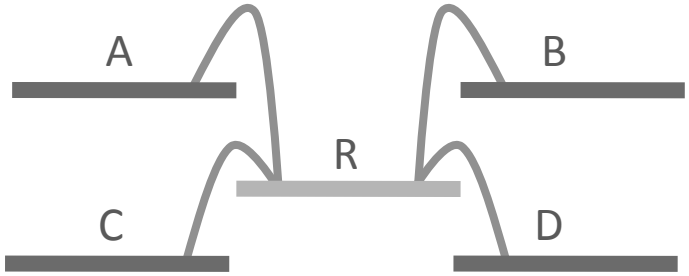
Assembled contigs



Collapsed repeat R

Task 3 | Investigating Contig Connectivity

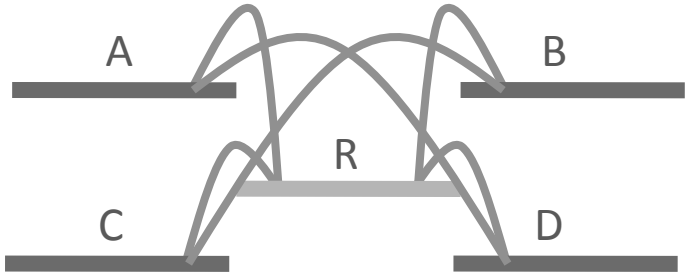
Assembled contigs



Collapsed repeat R
Arcs = Read Pairs

Task 3 | Investigating Contig Connectivity

Assembled contigs



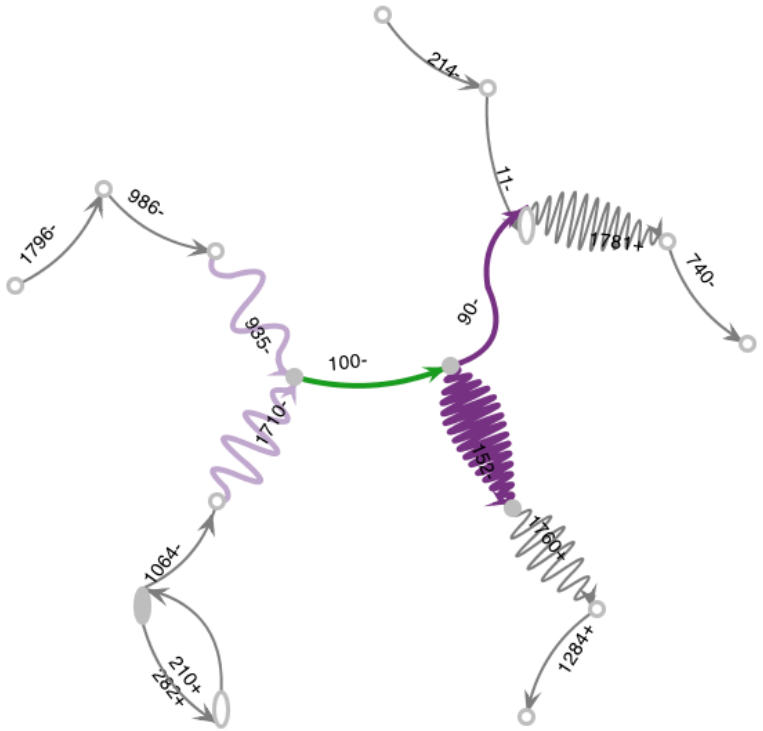
Collapsed repeat R
Arcs = Read Pairs

Arcs + linear ordering : gets complicated fast

Task 3 | Investigating Contig Connectivity

ABySS-Explorer

Nielsen *et al.*, InfoVis 2009

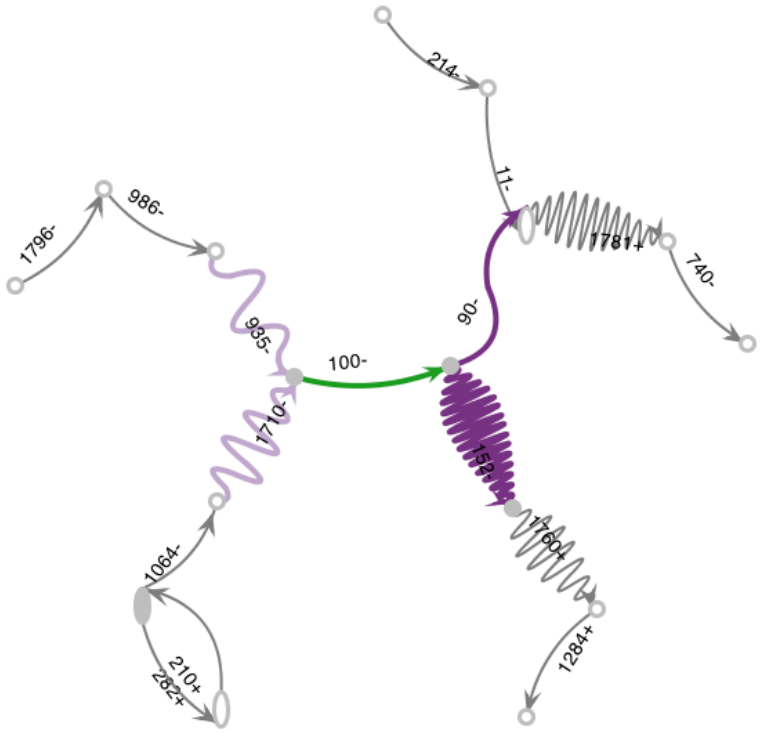


Edge = contig

Vertex = overlap

Squiggle = contig length (one oscillation per 1000 bps)

Task 3 | Investigating Contig Connectivity



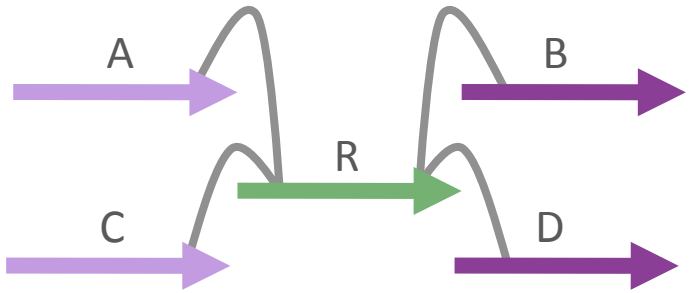
Edge = contig

Vertex = overlap

Squiggle = contig length (one oscillation per 1000 bps)

ABySS-Explorer

Nielsen *et al.*, InfoVis 2009

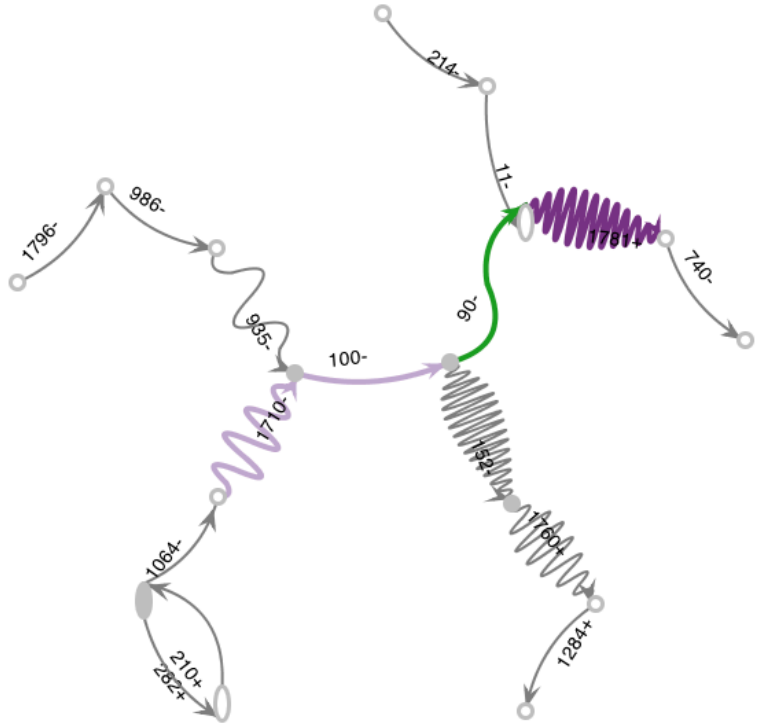


Green = selected contig

Light Purple = selected contig has upstream paired reads in this contig

Dark Purple = selected contig has downstream paired reads in this contig

Task 3 | Investigating Contig Connectivity



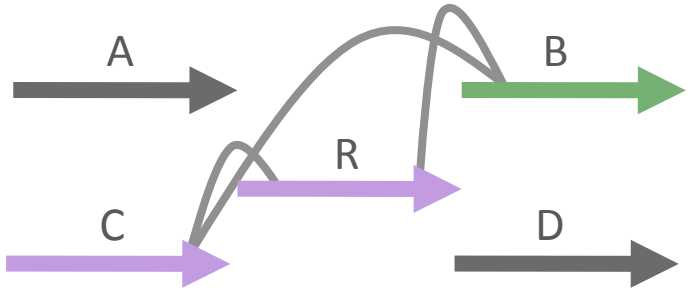
Edge = contig

Vertex = overlap

Squiggle = contig length (one oscillation per 1000 bps)

ABySS-Explorer

Nielsen *et al.*, InfoVis 2009

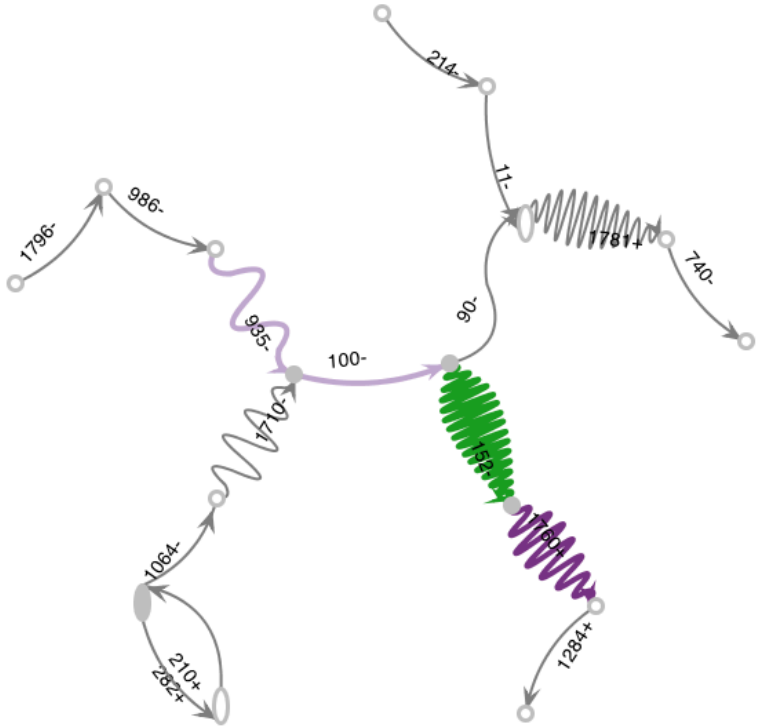


Green = selected contig

Light Purple = selected contig has upstream paired reads in this contig

Dark Purple = selected contig has downstream paired reads in this contig

Task 3 | Investigating Contig Connectivity



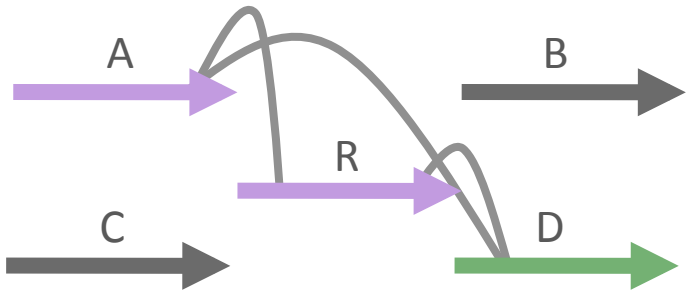
Edge = contig

Vertex = overlap

Squiggle = contig length (one oscillation per 1000 bps)

ABySS-Explorer

Nielsen *et al.*, InfoVis 2009



Green = selected contig

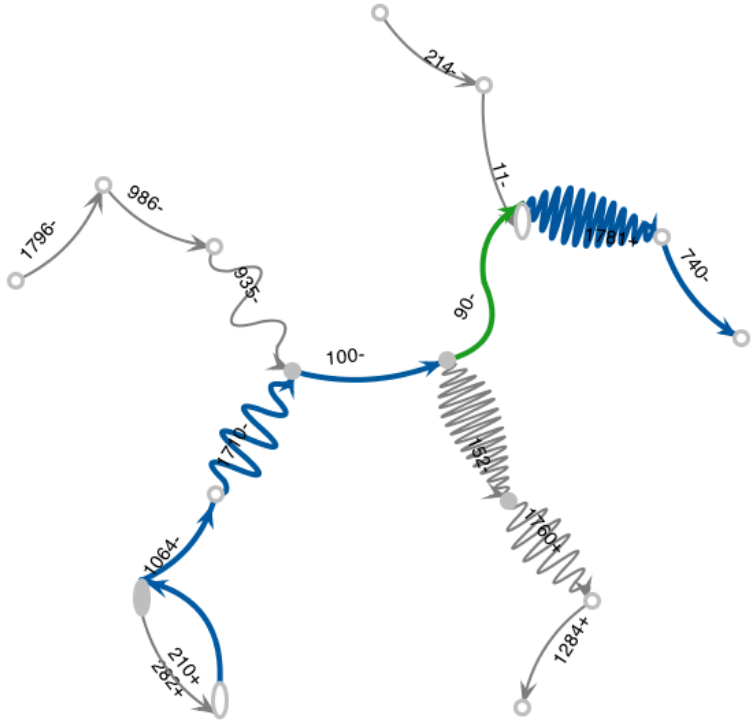
Light Purple = selected contig has upstream paired reads in this contig

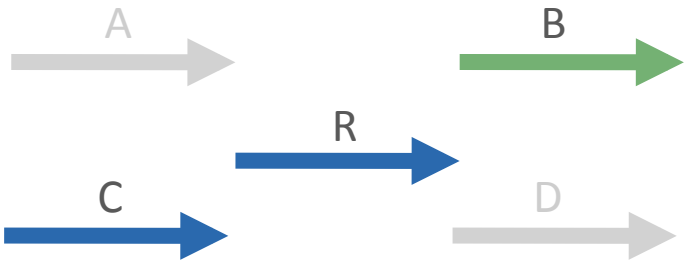
Dark Purple = selected contig has downstream paired reads in this contig

Task 3 | Investigating Contig Connectivity

ABySS-Explorer

Nielsen *et al.*, InfoVis 2009





Edge = contig

Vertex = overlap

Squiggle = contig length (one oscillation per 1000 bps)

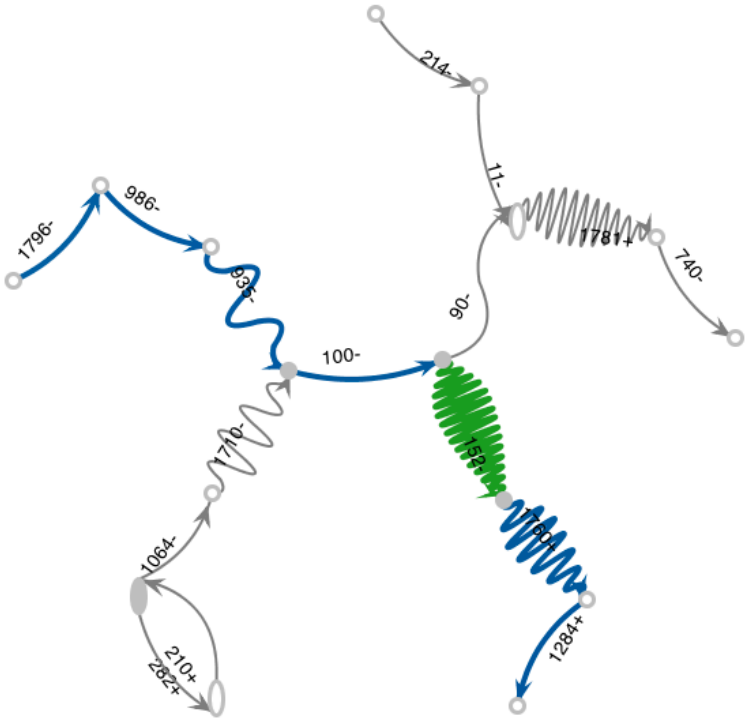
Green = selected contig

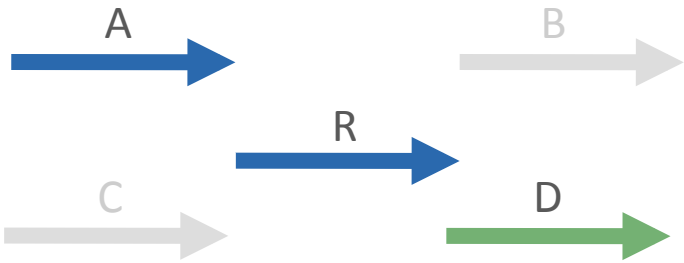
Blue = predicted scaffold

Task 3 | Investigating Contig Connectivity

ABySS-Explorer

Nielsen *et al.*, InfoVis 2009





Edge = contig

Vertex = overlap

Squiggle = contig length (one oscillation per 1000 bps)

Green = selected contig

Blue = predicted scaffold

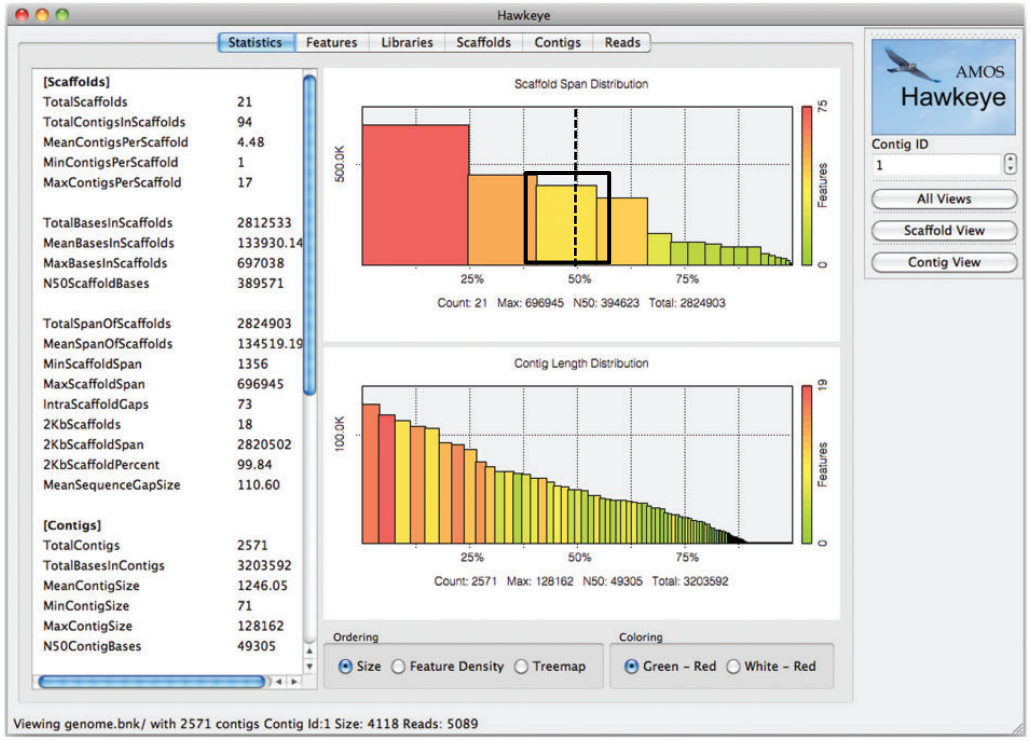
Task 4 | Assembly Evaluation

Task 4 | Assembly Evaluation

Contig size is a common metric of assembly quality

Scaffold N50

Half of the genome has been assembled into scaffolds larger than the N50 value

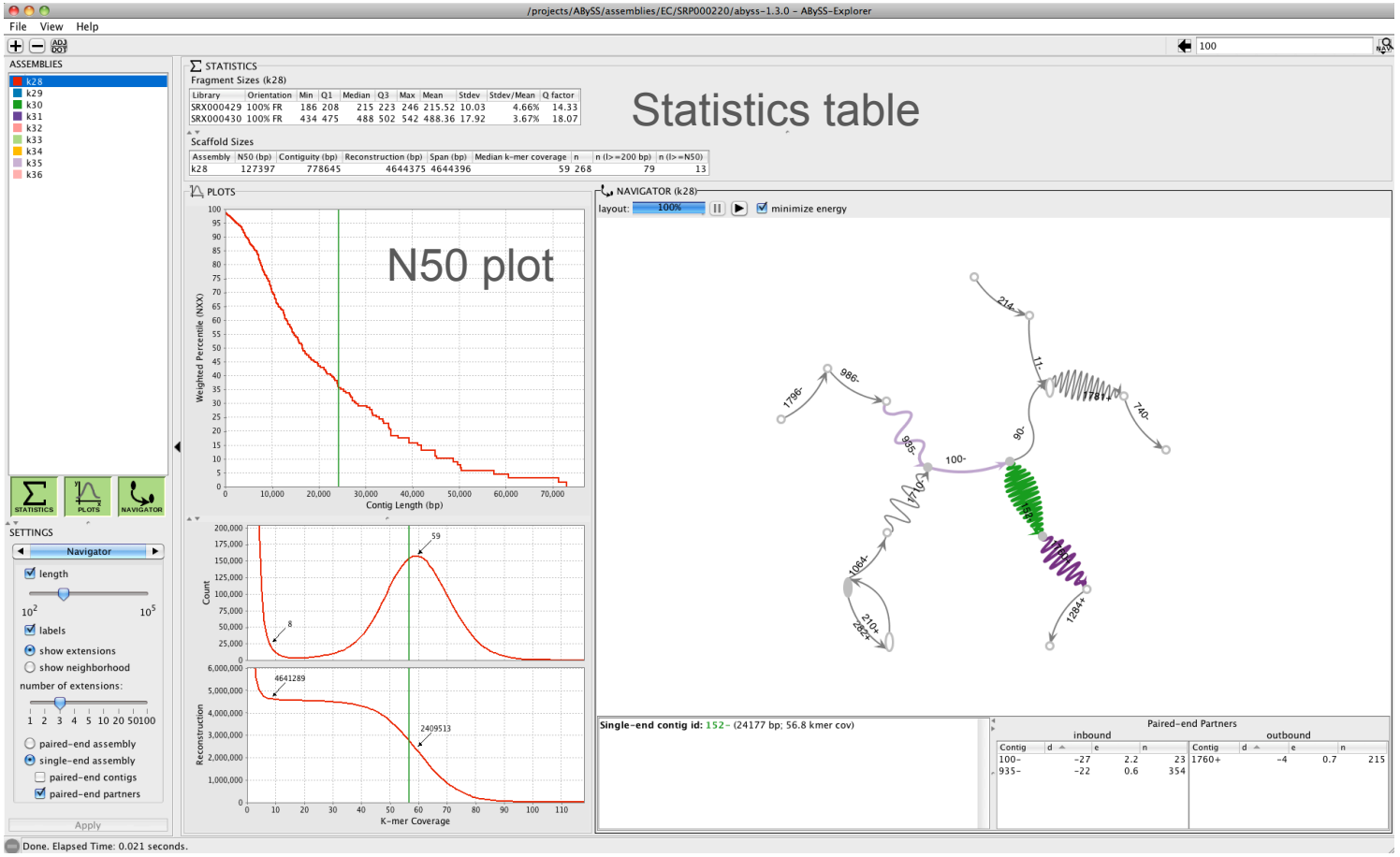


Bar height = contig size
Bar width = relative fraction of the genome size

Hawkeye

Schatz *et al.*, 2007; 2011

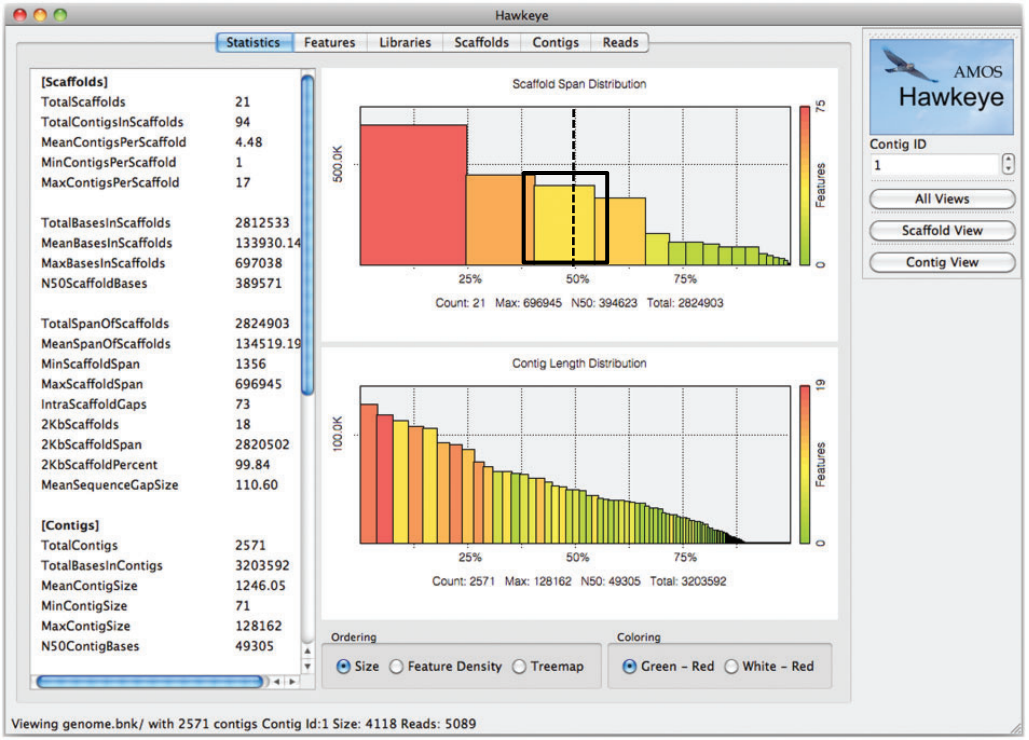
Task 4 | Assembly Evaluation



Task 4 | Assembly Evaluation

Contig quality also an important consideration

Bar colour indicates number of misassembly features discovered by AMOSvalidate



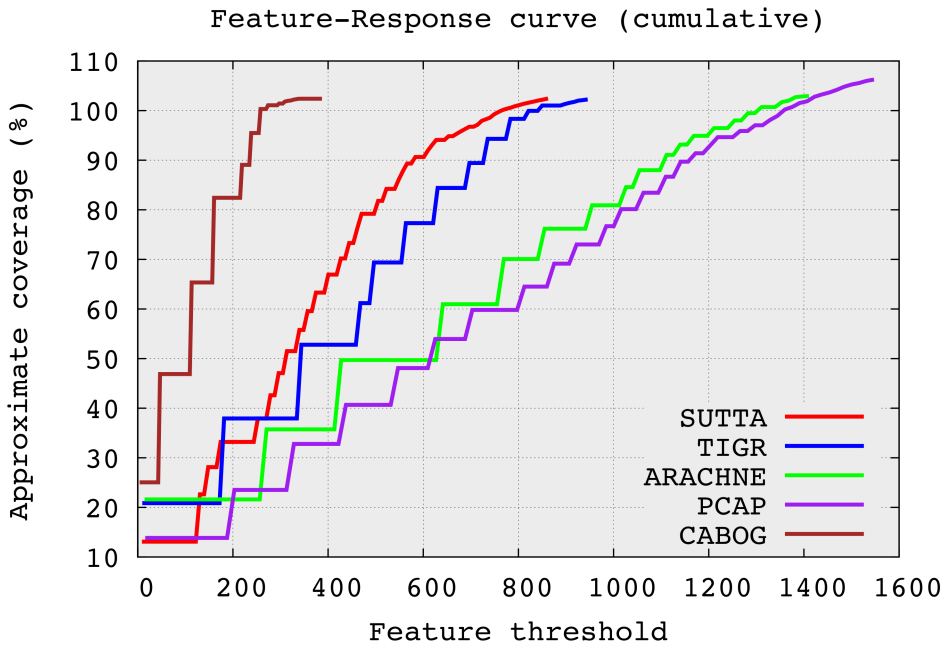
Bar height = contig size
Bar width = relative fraction of the genome size

Hawkeye

Schatz *et al.*, 2007; 2011

Task 4 | Assembly Evaluation

FRCurve - simultaneously measure connectivity and quality

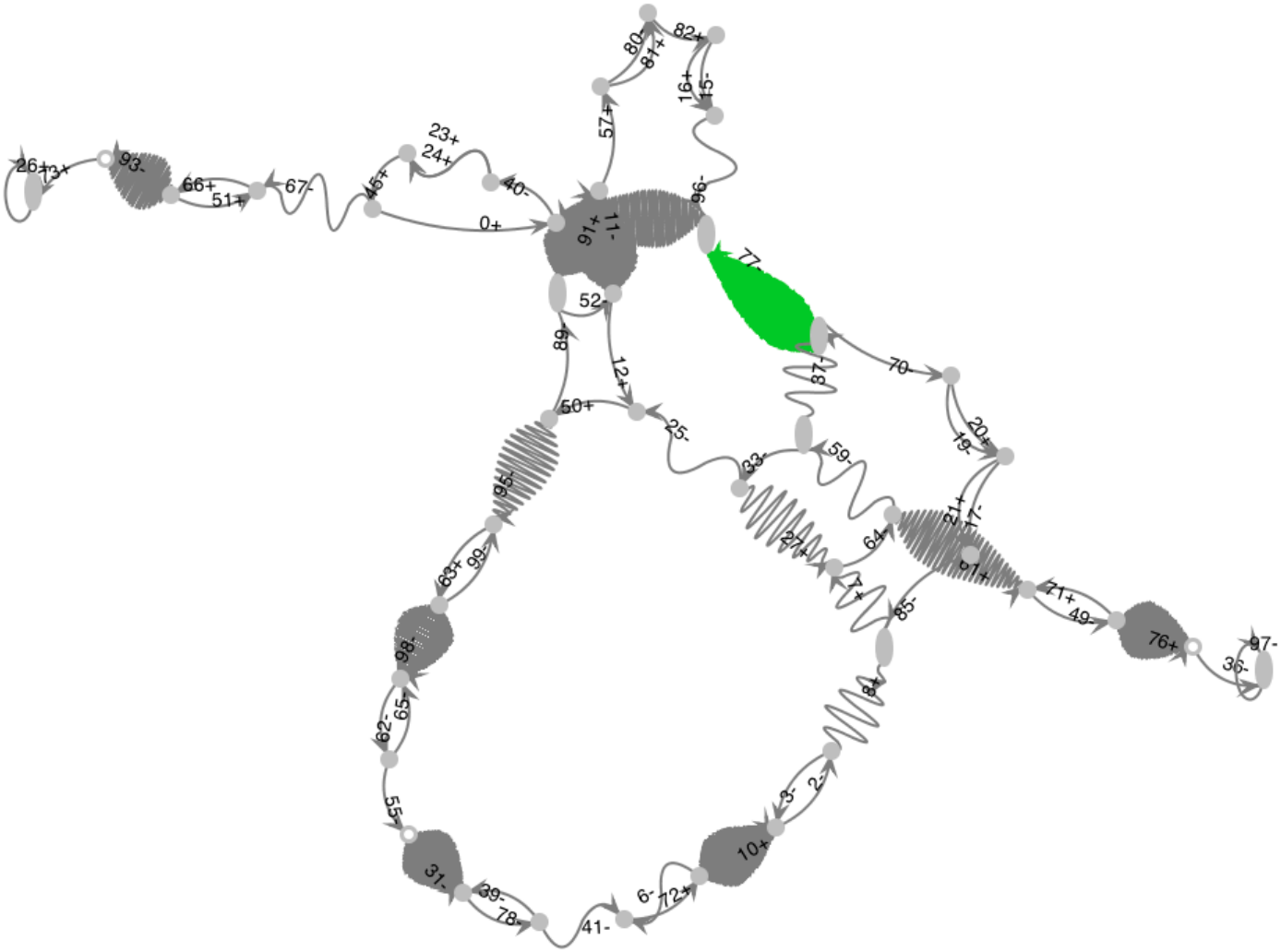


- Each contig has a number of misassembly features (detected by AMOSvalidate)
- Sort contigs by size (largest first) and tally genome coverage for contigs with < threshold missassembly feature counts

From Schatz *et al*.
Briefings in Bioinformatics, 2011

Task 5 | Making Sense of Complex Structures

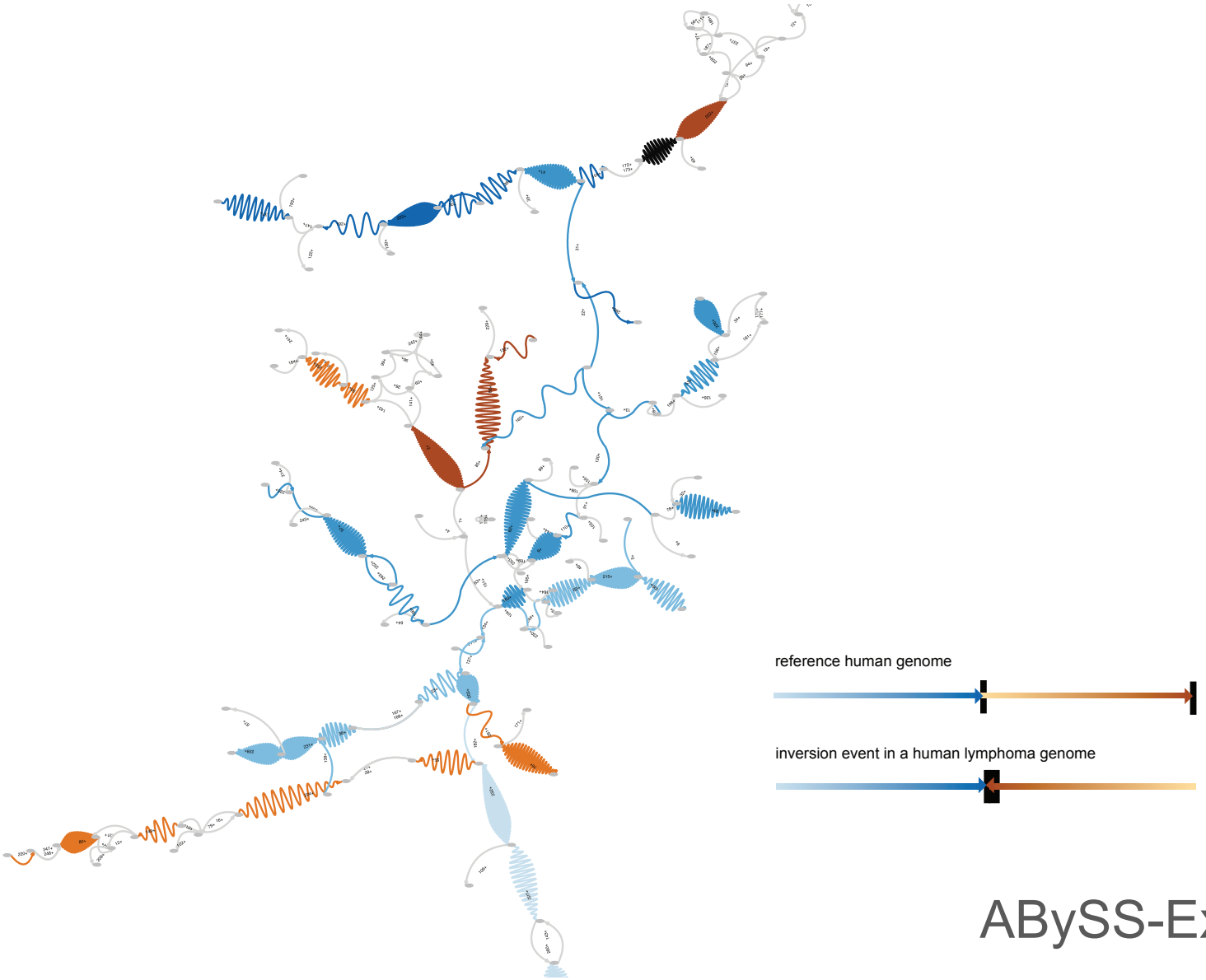
Task 5 | Making Sense of Complex Structures



1 Mbp *Mycoplasma capricolum* genome

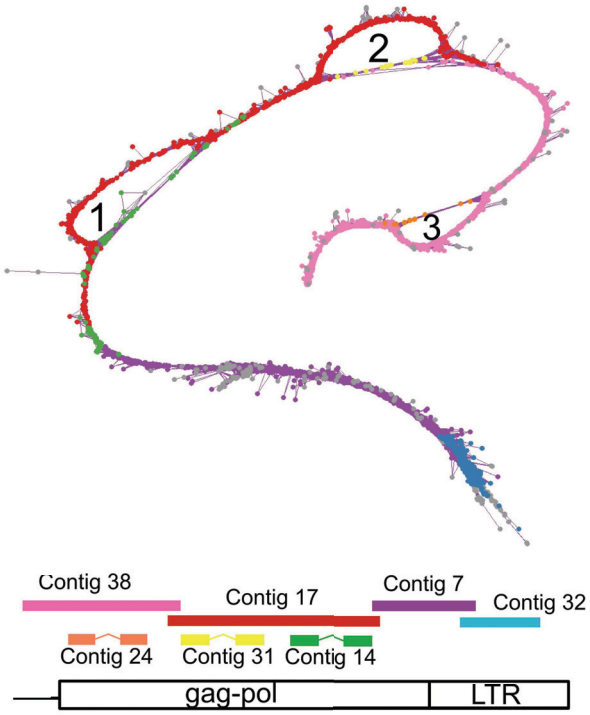
ABySS-Explorer

Task 5 | Making Sense of Complex Structures



ABySS-Explorer

Task 5 | Making Sense of Complex Structures



Vertex = sequencing read

Edge = overlap

- Use read clustering (not assembly *per se*) based on sequence similarity to identify structures of repeat families
- 5320 reads from the *Pisum sativum* (pea) genome representing the Ty1/copia LTR-retrotransposon Angela (CAP3 assembled contigs shown below)
- Enabled identification of the most common form of the Angela element and three less frequent deletions

From Novák *et al*.
BMC Bioinformatics, 2010

Summary | Where visualization is used

1. Nucleotide-Level Discrepancies
2. Deviant Read Pairs
3. Contig Connectivity
4. Assembly Evaluation
5. Complex Structures

Summary | Challenges

1. Must address multiple levels of resolution
2. Large data sets pose computational and performance challenges
3. Rapidly changing field directly affected by innovations in sequencing technology

Acknowledgements

BCGSC Vancouver, Canada

ABySS Team:

Ka Ming Nip

Shaun Jackman

Karen Mungall

İnanç Birol

Martin Krzywinski

Steven Jones

Assembly Visualization

David Gordon

Mick Watson

Simon Andrews

Mark Blaxter

Peter Cock

Tom Freeman

Sujai Kumar

Giuseppe Narzisi

Michael Schatz

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

Michael Smith Foundation for Health Research