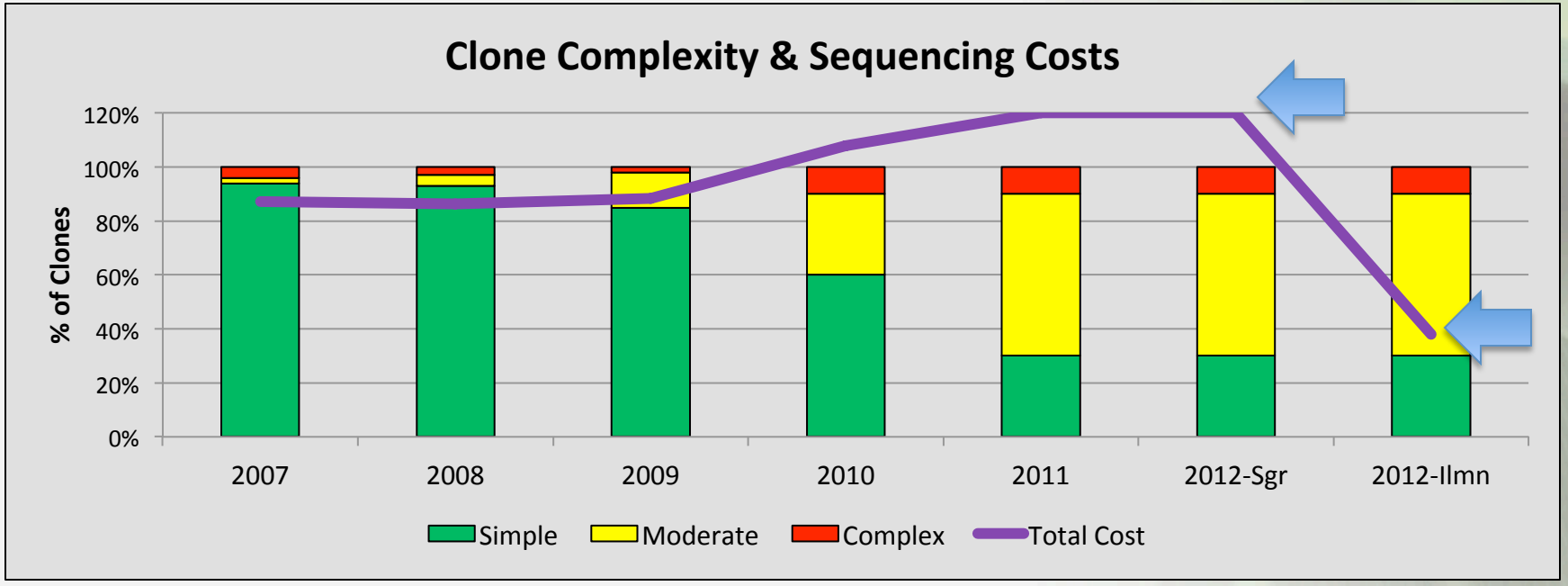
High Throughput Plasmid Sequencing with Illumina and CLC bio

Ajay Athavale
Monsanto Company
05 June 2012

Monsanto's Transgenic Plant Pipeline: Sequencing 1,000's of plasmids per year

Increasing Clone Complexity Required New Solutions for Plasmid Sequencing & Analysis



- NextGen sequencing offers significant savings & reduced processing time
- Illumina was demonstrated to be optimal NGS platform for plasmid sequencing
- Scalable and accurate finishing tools were needed for Illumina sequence reads

Requirements for HTP Plasmid Finishing Platform

- Performance, Accuracy & Scalability
  - Support structurally challenging plasmids
  - Support for multiple sequencing platforms
  - Hybrid assemblies (across platforms)
  - User configurable & flexible parameters
    - Read trimming
    - Mutation detection (SNPs)
    - Insertion & Deletion detection (DIPs)
- Usability & User Interface
  - Easily edit/manipulate assemblies
  - Execution from both UNIX and GUI environments
  - User friendly visualization tools
- Integration with in house corporate research data systems
- Customization of finishing workflows

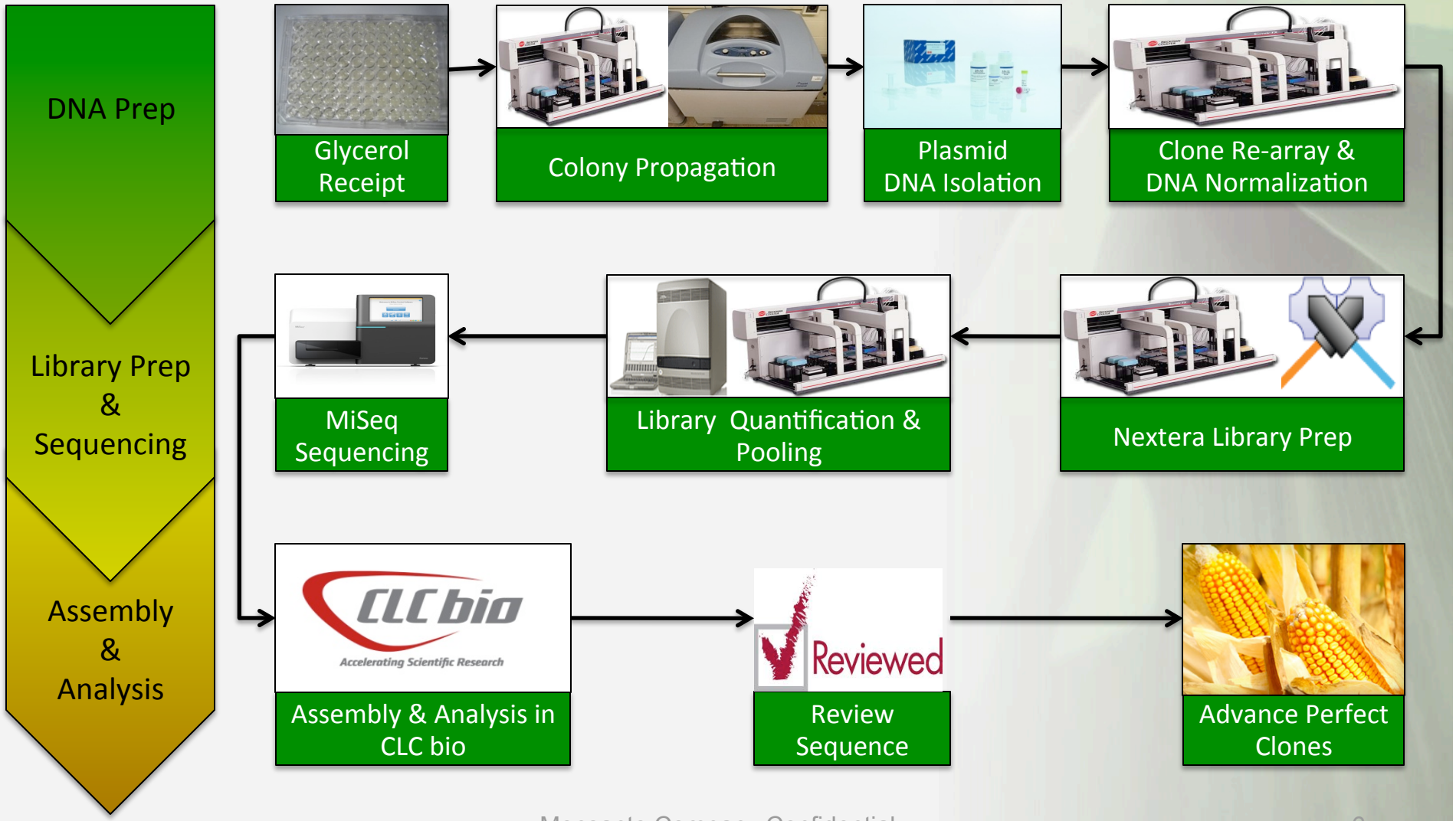
→ **CLC bio was chosen as next gen finishing tool, given these criteria**

Monsanto Company Confidential

Demonstrated identical finishing results for structurally complex plasmids using Illumina and CLC Bio

Plasmid	Known Complex Sequences	Observed Seq Sanger + CONSED	Observed Seq Illumina + CLC bio
Plasmid A	None	☑	☑
Plasmid B	Mixed Population	☑	☑
Plasmid C	Homopolymer tracts	☑	☑
Plasmid D	Homopolymer tracts Tandem Repeats	☑	☑
Plasmid E	Homopolymer tracts Inverted Repeats	☑	☑
Plasmid F	Homopolymer tracts Inverted Repeats Tandem Repeats	☑	☑
Plasmid G	Large Insertion	☑	☑
Plasmid H	Large Deletion	☑	☑

Overview of the Current Sequencing & Finishing Process with Illumina and CLC bio



Automated Assembly & Analysis Workflow in CLC bio streamlines workflow

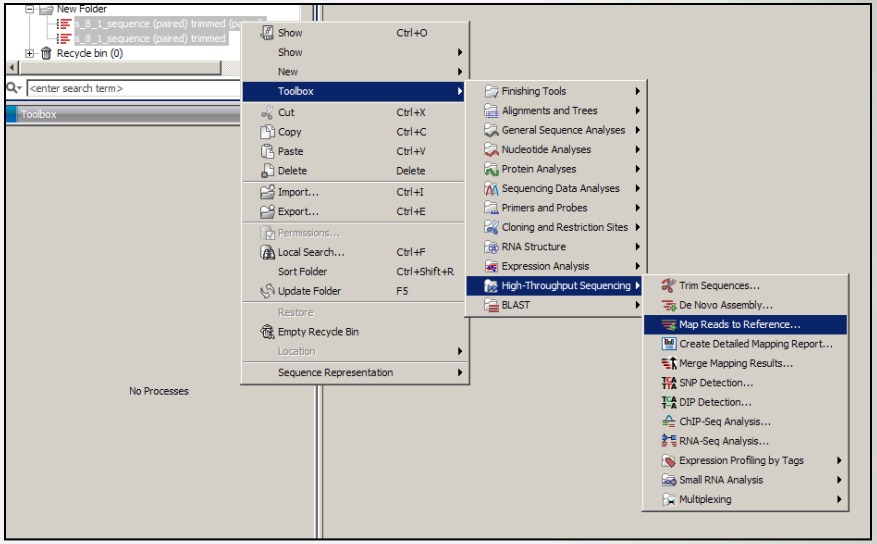
Assembly & Analysis Workflow

- Prepare Config file
- Import Illumina data
- Quality Trimming
- Read Mapping
- Coverage Analysis
- SNP Detection
- DIP Detection
- Review Assemblies

- The CLC bio Genomics Server and Command line tools enable efficient parallel processing of large batches of constructs

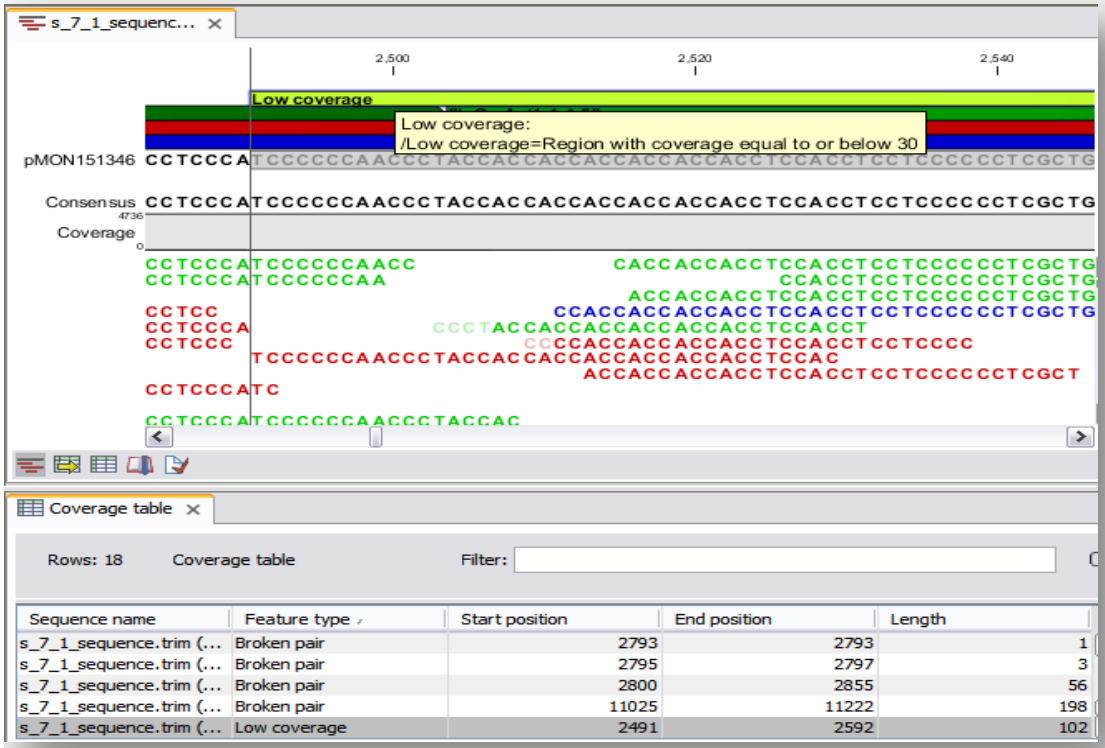
```
unix:> ./clcserver -S server -U user -W password -A
read_mapping -i input_reads --references reference
-d save_destination
```

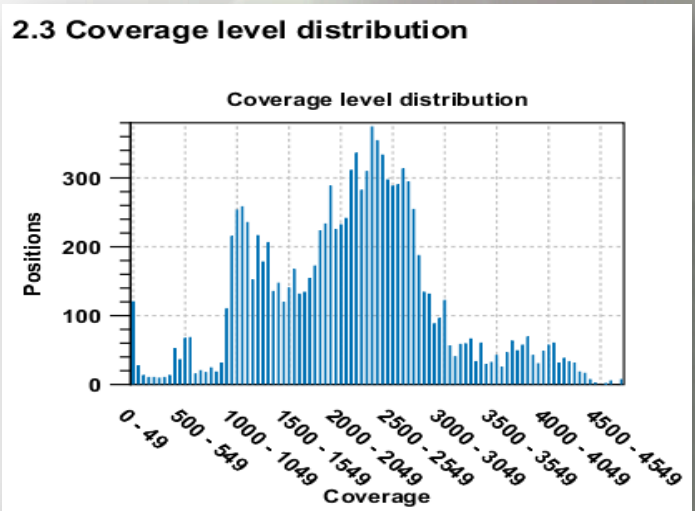
- For the non-UNIX user, command line tools can be easily run in the GUI (CLC Genomics workbench)



Identification of low coverage regions with the "Contig Analysis" tool

Easy identification of low coverage areas

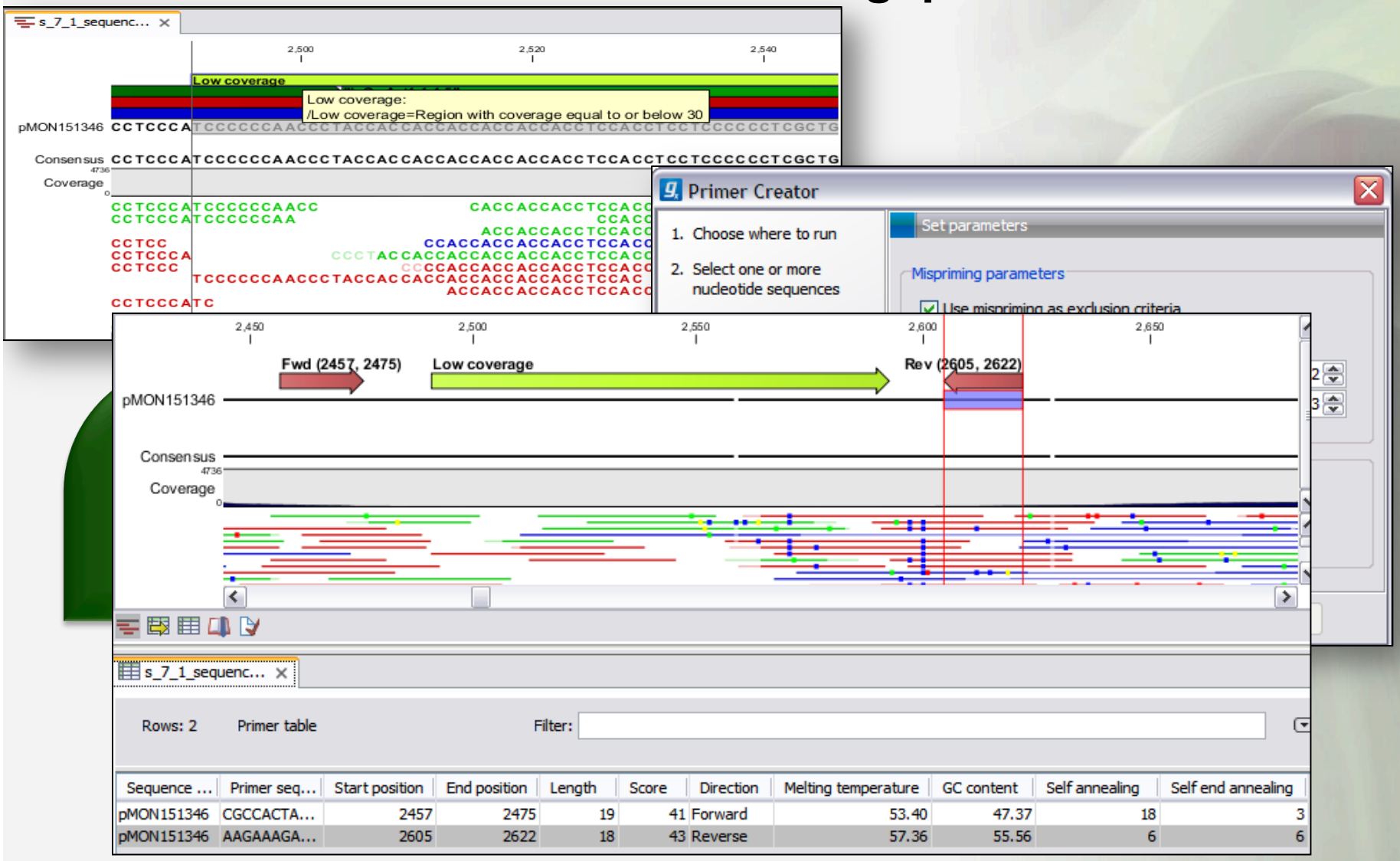




2.2 Coverage statistics

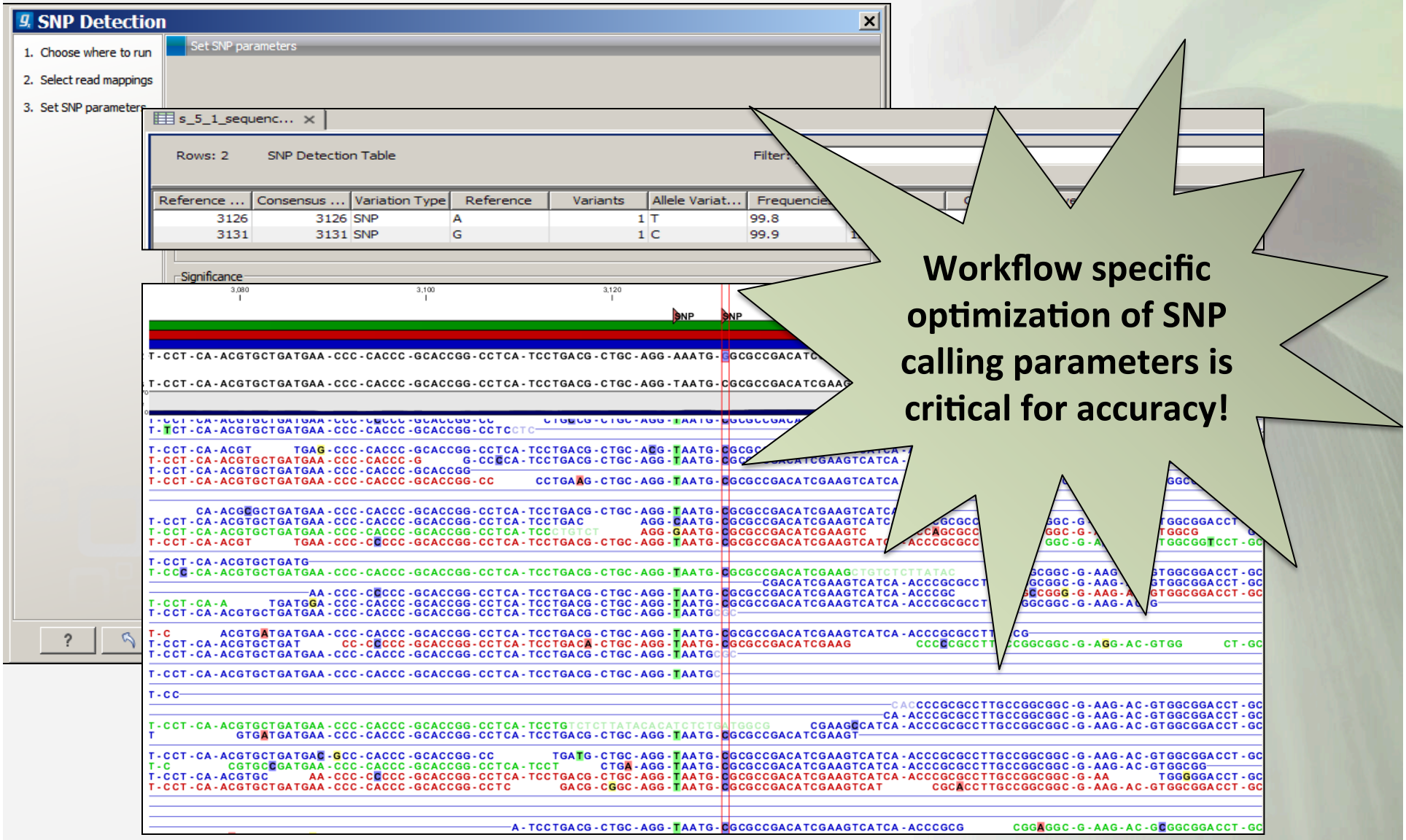
Total reference length	11,222
Minimum coverage	6
Maximum coverage	4,736
Average coverage	2,118.12
Standard deviation	855.63

The “Primer Creator” tool helps to streamline efforts to close gaps

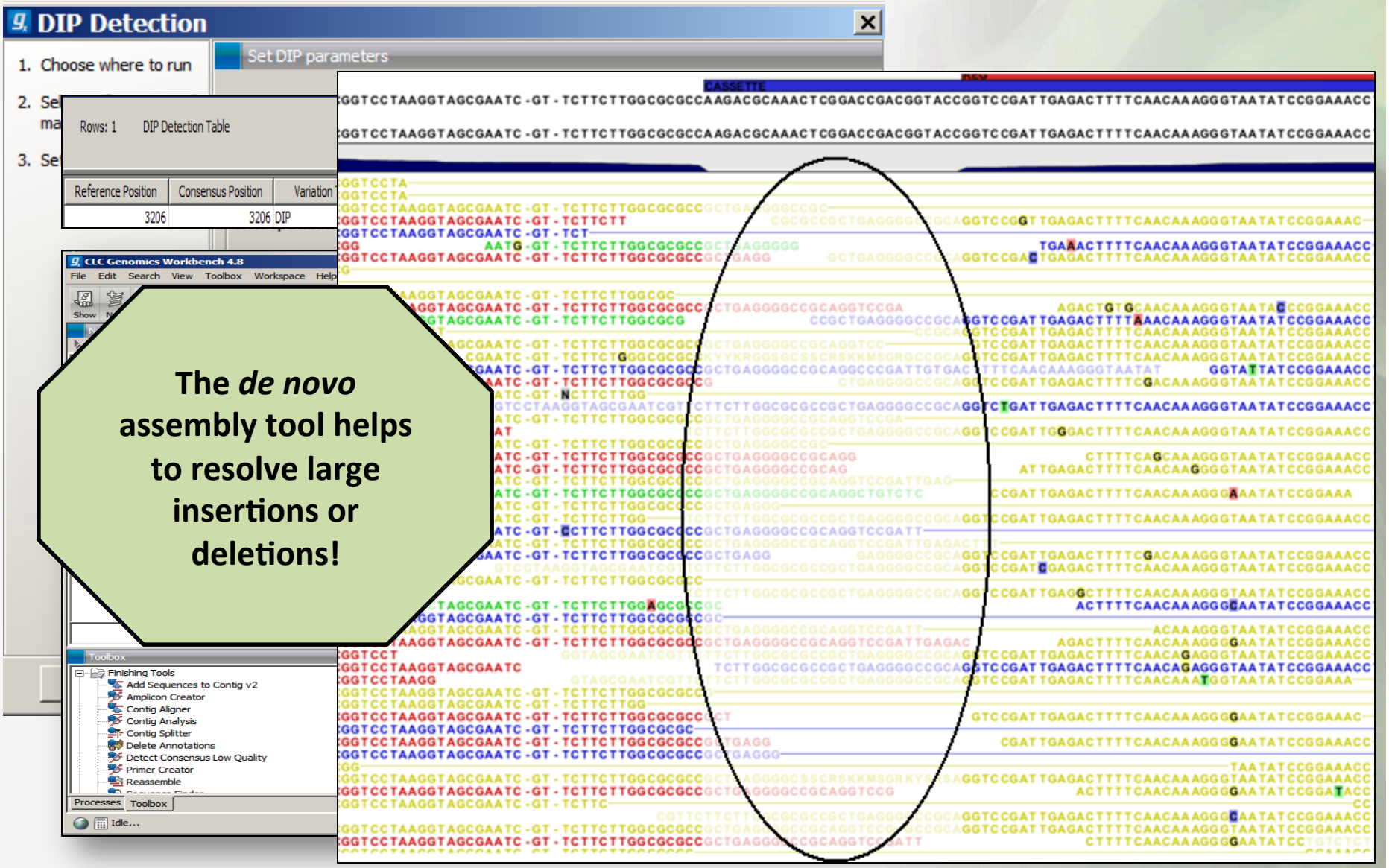


Sequence ...	Primer seq...	Start position	End position	Length	Score	Direction	Melting temperature	GC content	Self annealing	Self end annealing
pMON151346	CGCCACTA...	2457	2475	19	41	Forward	53.40	47.37	18	3
pMON151346	AAGAAAGA...	2605	2622	18	43	Reverse	57.36	55.56	6	6

Simplified variant detection with the “SNP Detection” tool

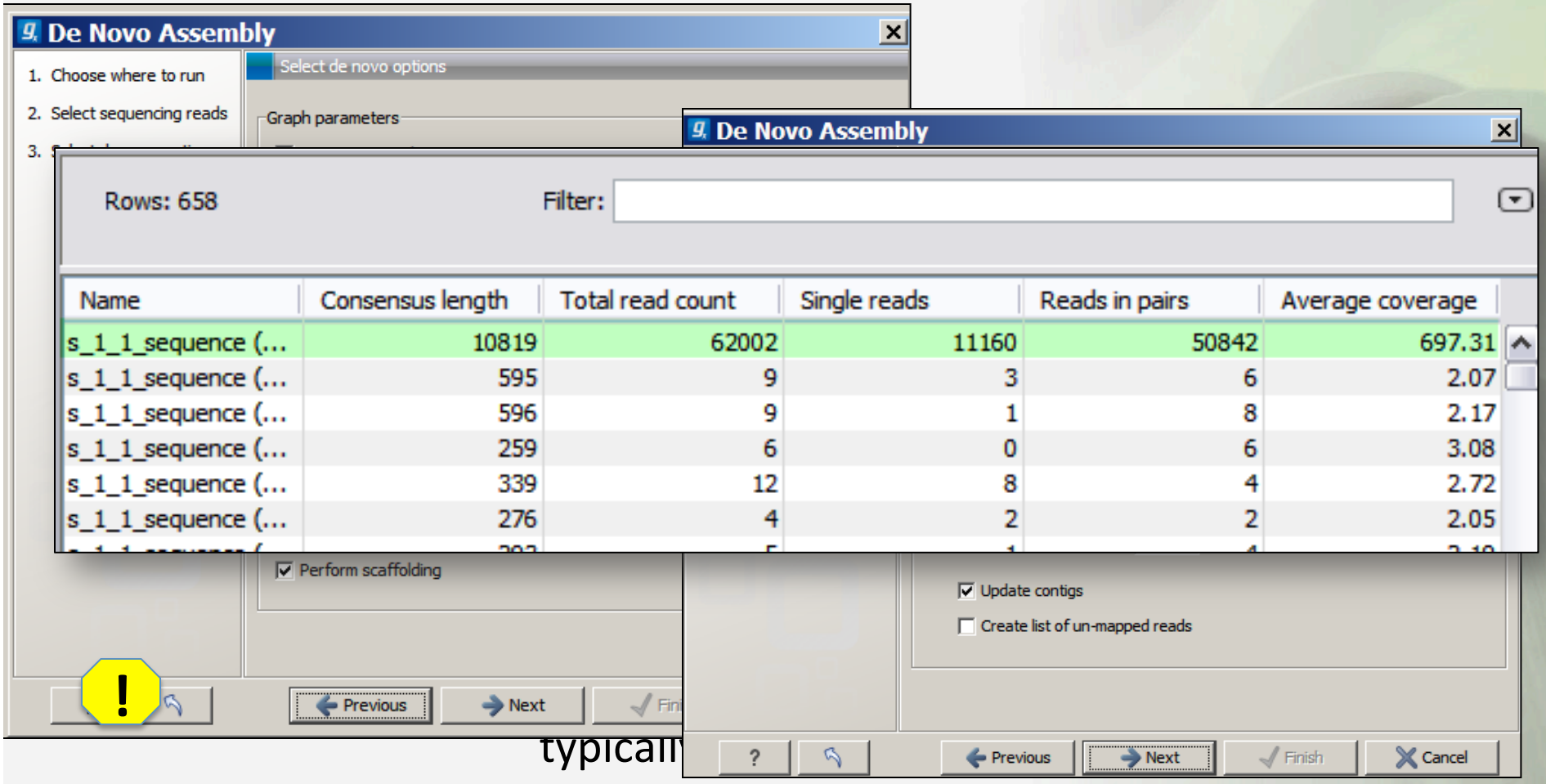


“DIP Detection” for Deletions & Insertions



The *de novo* assembly tool helps to resolve large insertions or deletions!

Denovo assembly used to resolve Indels



Name	Consensus length	Total read count	Single reads	Reads in pairs	Average coverage
s_1_1_sequence (...	10819	62002	11160	50842	697.31
s_1_1_sequence (...	595	9	3	6	2.07
s_1_1_sequence (...	596	9	1	8	2.17
s_1_1_sequence (...	259	6	0	6	3.08
s_1_1_sequence (...	339	12	8	4	2.72
s_1_1_sequence (...	276	4	2	2	2.05

typicall

CLC Bio Meets Requirements for Monsanto HTP Plasmid Finishing

- ☑ Performance, Accuracy & Scalability
  - ☑ Support structurally challenging plasmids
  - ☑ Support for multiple sequencing platforms
  - ☑ Hybrid assemblies (across platforms)
  - ☑ User configurable & flexible parameters
    - Read trimming
    - Mutation detection (SNPs)
    - Insertion & Deletion detection (DIPs)
- ☑ Usability & User Interface
  - ☑ Easily edit/manipulate assemblies
  - ☑ Execution from both UNIX and GUI environments
  - ☑ User friendly visualization tools
- ☑ Integration with in house corporate research data systems
- ☑ Customization of finishing workflows

Additional CLC bio capabilities are also used across Monsanto

- In silico cloning & plasmid design
- Sanger sequence review/assembly
- Plasmid sequencing & finishing via Sanger & Illumina
- Genome assembly for microbes and plants
- DNA and Protein alignment for research teams
- … and much more!

Acknowledgements

- CLC Bio
  - Jannick Bendtsen
  - Henrik Sandmann
  - Joe Salvatore

- Monsanto
  - Todd Michael
  - Dan Ader
  - Amber Ford
  - Susan Johnson
  - Cynthia LaBanca
  - Kim Lawry
  - Jing Lu
  - Karen Martin
  - Tim Mitsky
  - Stacie Norton