Towards Copy-Aware Assembly of the Sugarcane Genome

Gabriel R. A. Margarido
Cristina Pop
Bob Davidson
Glaucia M. Souza
David Heckerman

Outline

- Initial efforts
- Initial findings
- Directions we are headed

Introduction

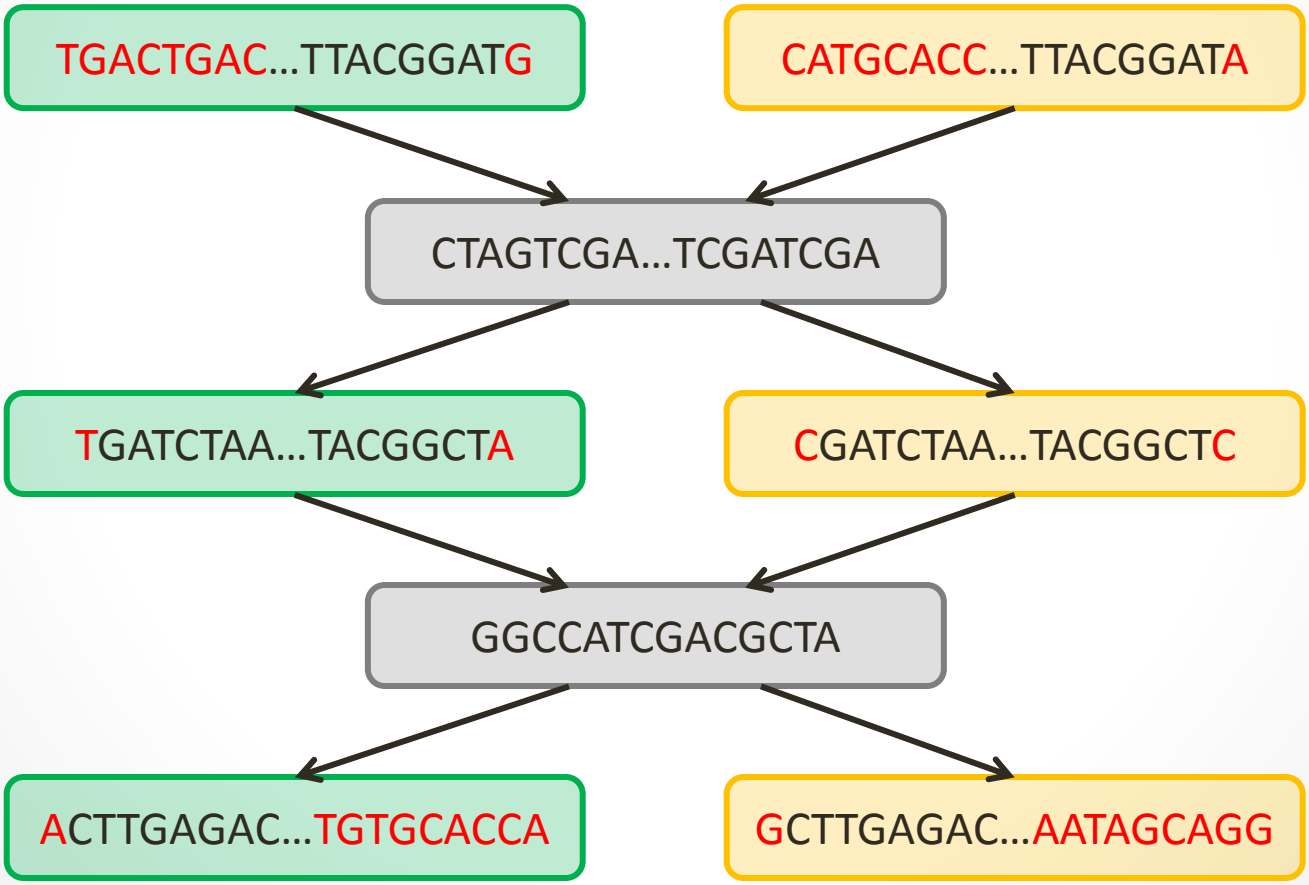
- Traditional breeding facilitated by cloning

- Complex aneuploid and polyploid genome
- 10 Gb in 100-130 chromosomes
  - 1 Gb monoploid genome
  - 6 to 12 copies each

- Interest in difference between homoeologues
  - Assembly software collapses SNP's

Motivation

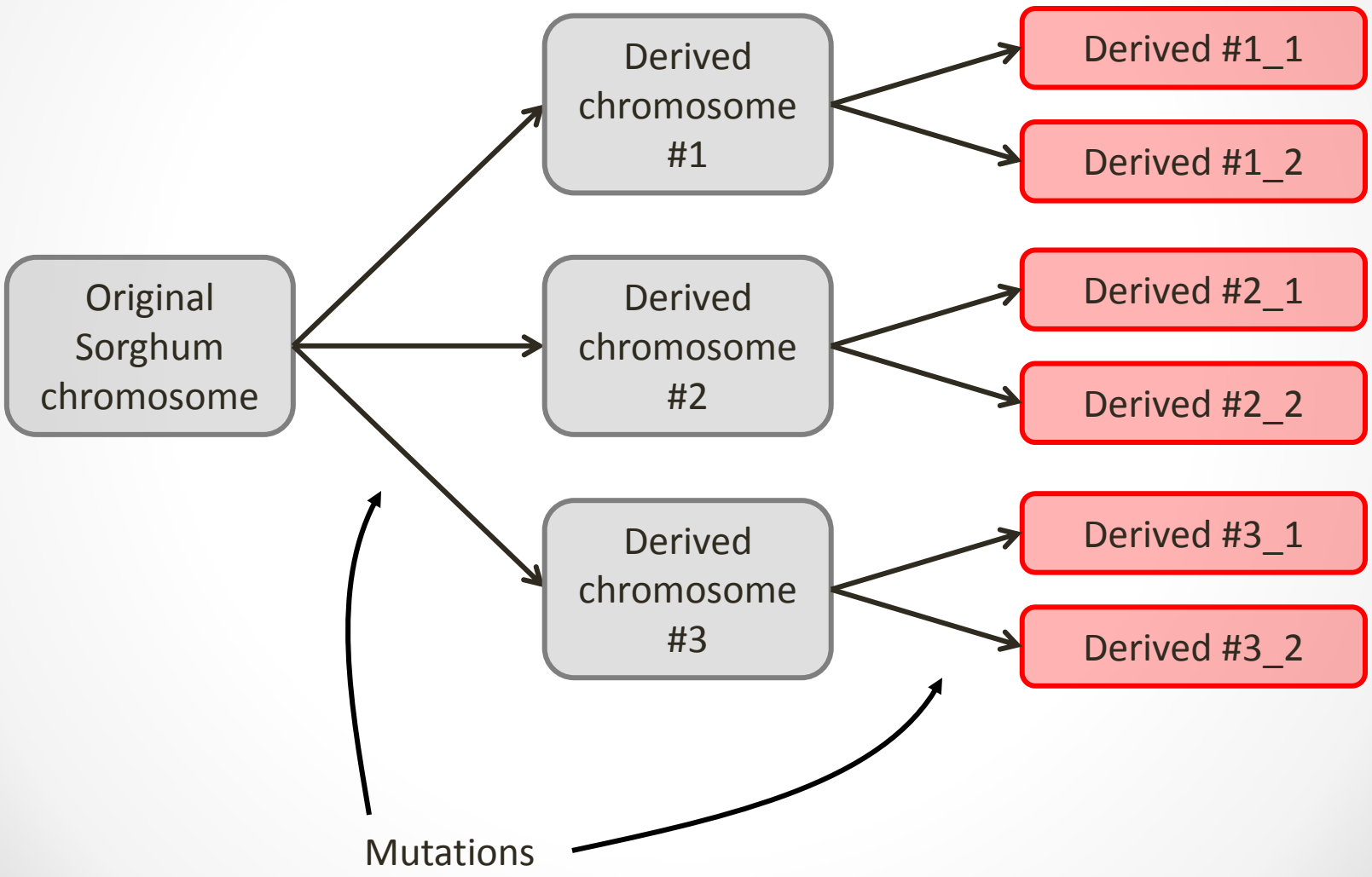
- Goal: assembly with chromosome copy sorting



Synthetic Genome

- *Sorghum bicolor*
  - Closest diploid species to sugarcane
    - 95% similarity
  - Sequenced genome

Synthetic Genome



Synthetic Genome

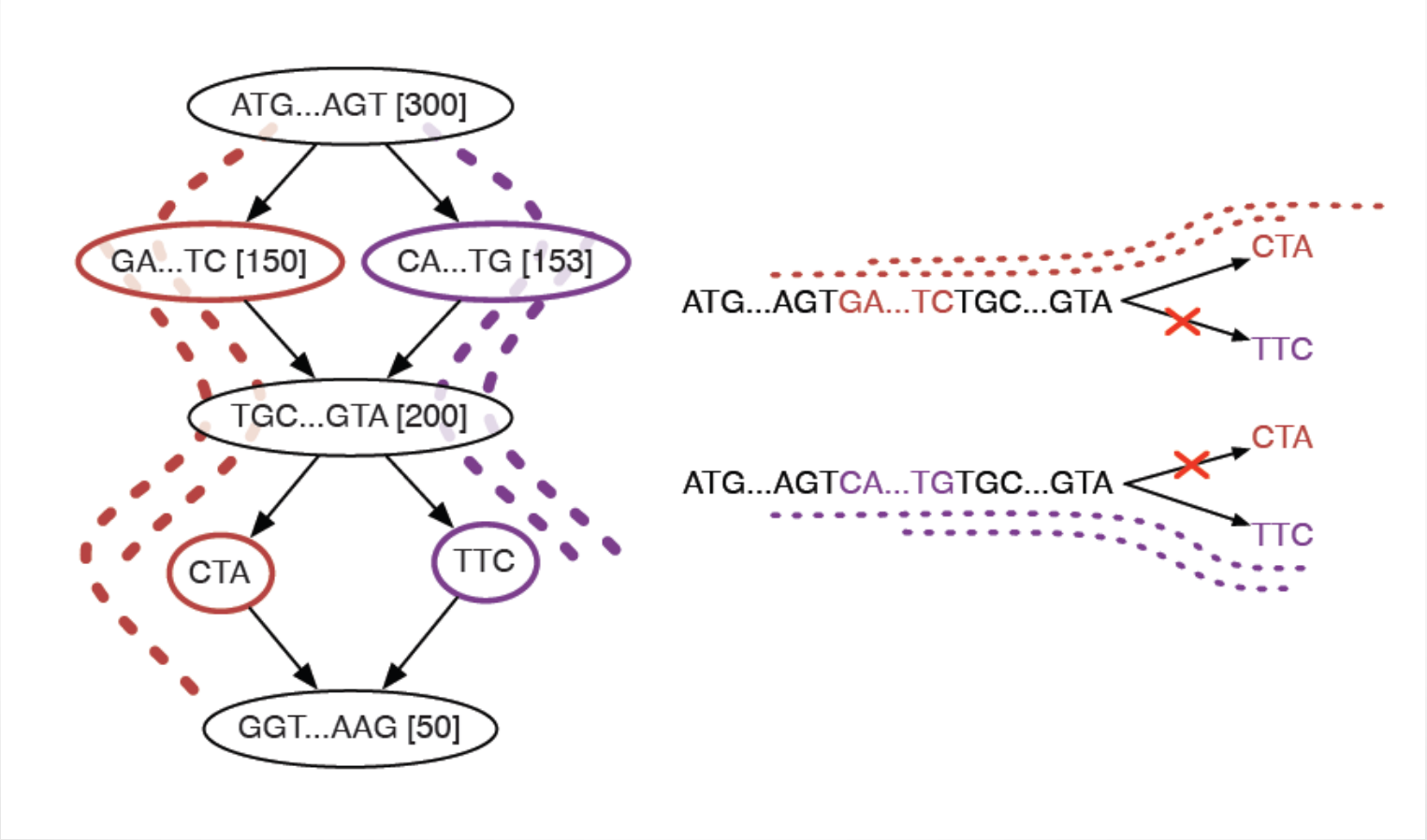
- Rearrangements
  - Fusions
  - Duplications
  - Inversions
  - (Reciprocal) Translocations
- Hypothetical polyploid genome
  - 96 chromosomes
- Read simulation
  - Any desirable error model

Comparative Approach

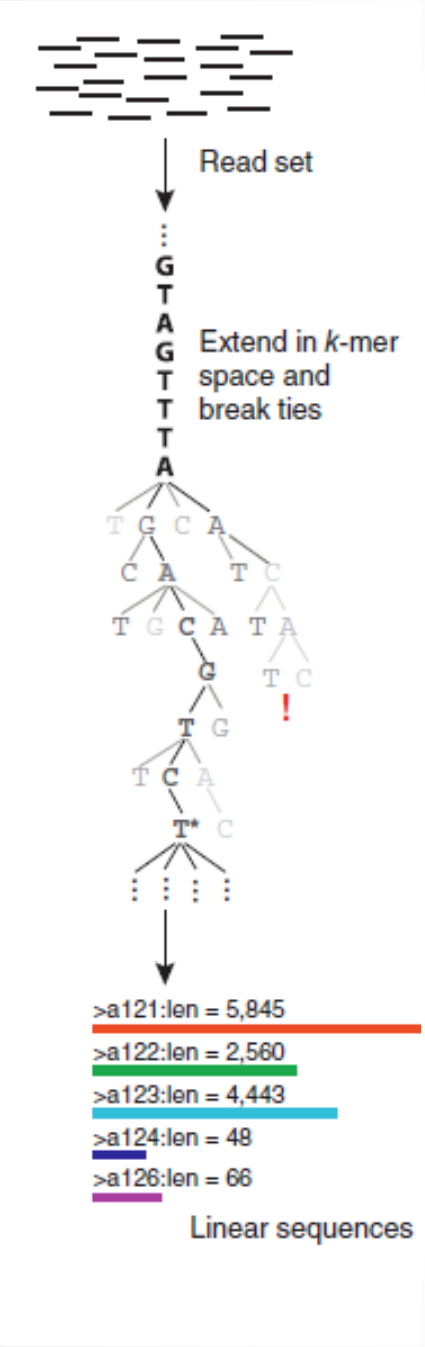
- Comparative genome assembly
  - Alignment against reference
  - Layout identification
  - Contig formation
- Results (projected third-generation technology)
  - Broken and collapsed assembly
  - Short contigs

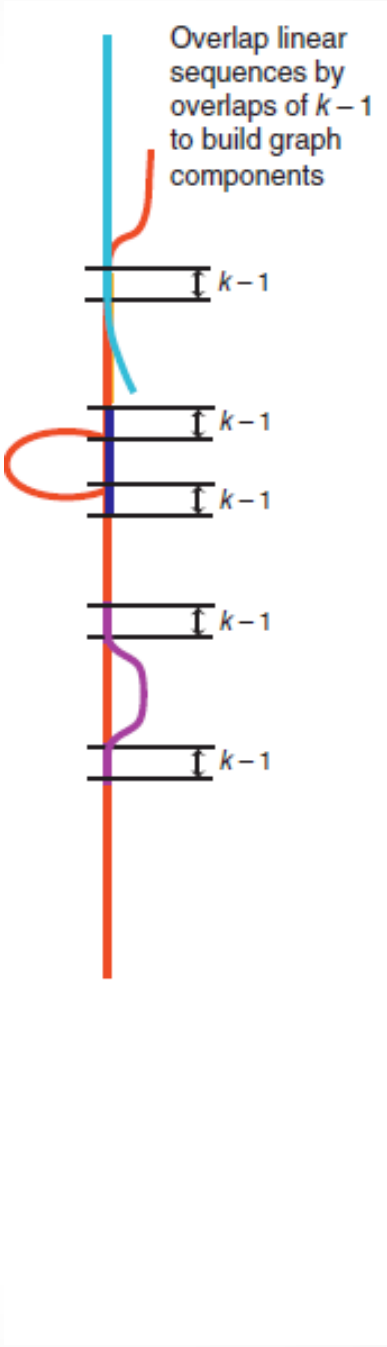
Ideas from RNA-Seq

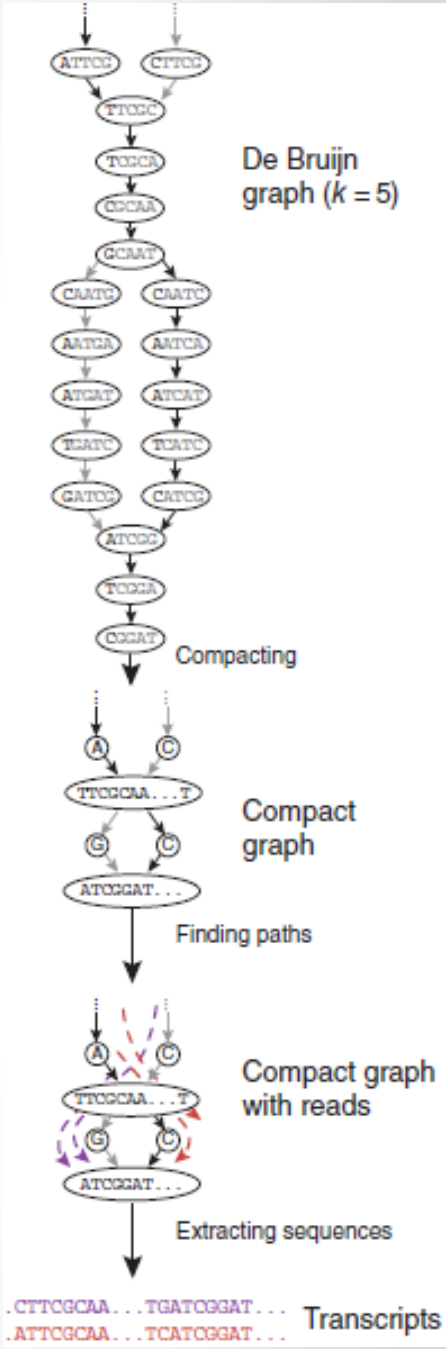
- Alternative splicing
- Expression of both alleles in diploids



Trinity


Read set
GTAGTTTA
Extend in k-mer space and break ties
>a121:len = 5,845
>a122:len = 2,560
>a123:len = 4,443
>a124:len = 48
>a126:len = 66
Linear sequences


Overlap linear sequences by overlaps of k – 1 to build graph components
k – 1
k – 1
k – 1
k – 1
k – 1


De Bruijn graph (k = 5)
Compacting
Compact graph
Finding paths
Compact graph with reads
Extracting sequences
Transcripts

Trinity Results

- Extremely short contigs
  - Smaller than average read length
- Theoretically promising method
- In current state, not appropriate

Current Technologies

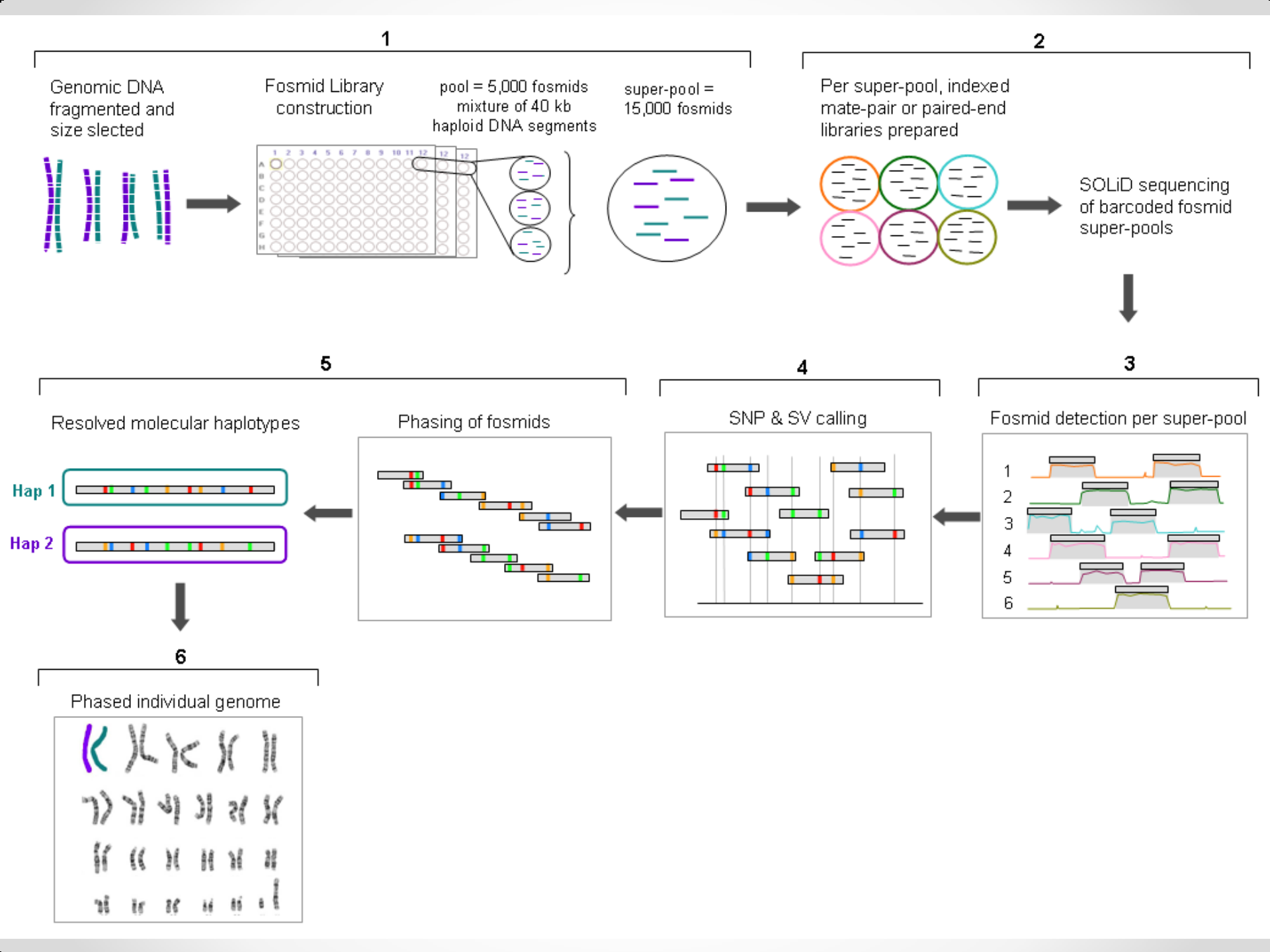
- Read length
- Sequencing errors
  - 454 reads for initial assembly
  - Deep coverage (> 50X) Illumina data for SNP calling

Phasing in Humans

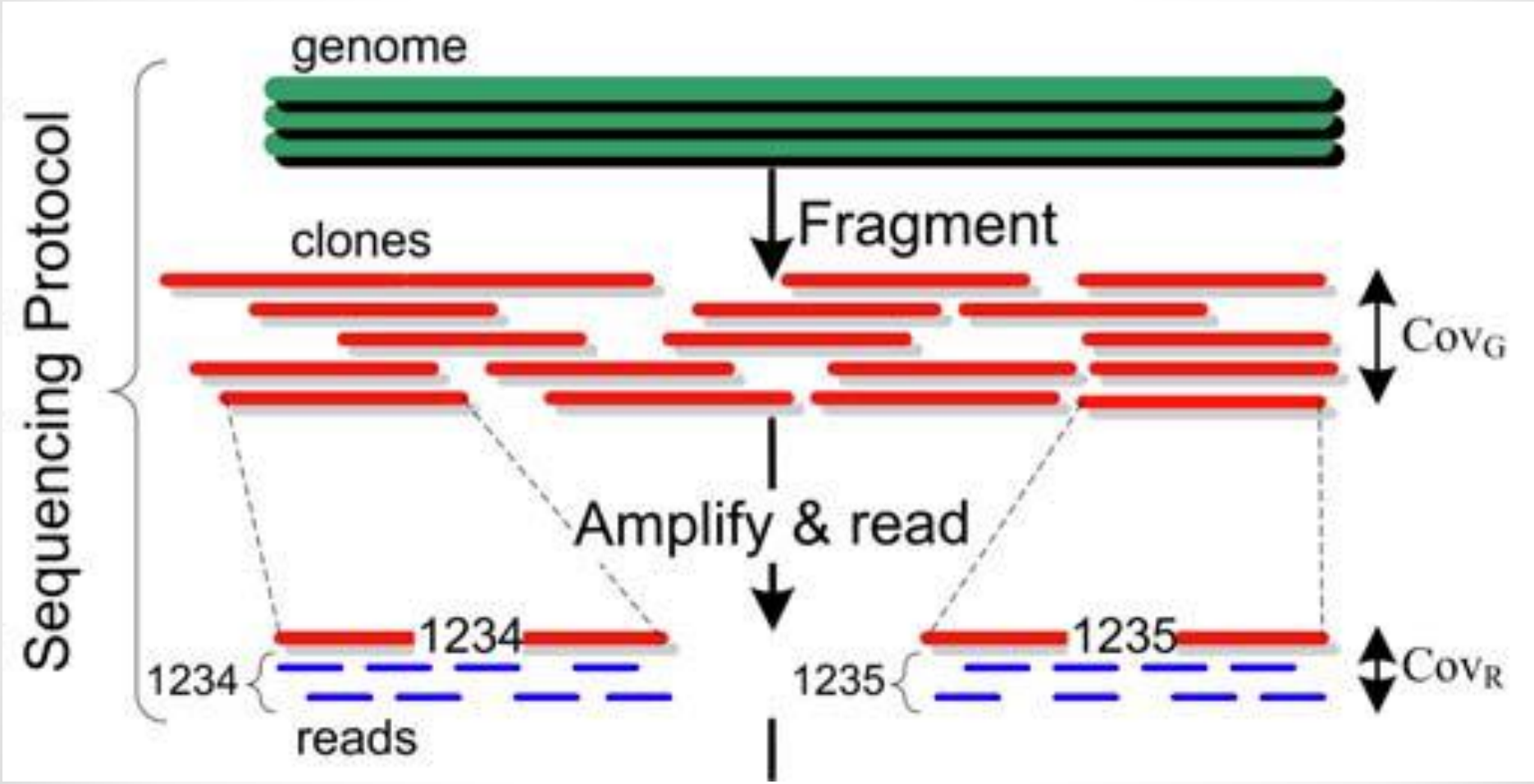
- Most assemblies are haploid
- SNP calling

- Phasing = chromosome sorting in sugarcane
- Current technologies are not enough for reliable phasing

- Current trend
  - Use of fosmids for individual sequencing of haploid segments


1
Genomic DNA fragmented and size slected
Fosmid Library construction
pool = 5,000 fosmids mixture of 40 kb haploid DNA segments
super-pool = 15,000 fosmids
2
Per super-pool, indexed mate-pair or paired-end libraries prepared
SOLiD sequencing of barcoded fosmid super-pools
3
Fosmid detection per super-pool
1
2
3
4
5
6
4
SNP & SV calling
5
Phasing of fosmids
Resolved molecular haplotypes
Hap 1
Hap 2
6
Phased individual genome

Hierarchical Assembly



- High fragment coverage and low read coverage

Acknowledgements

- Manju Shivanna
- Nir Friedman
- Carlos Hotta
- Jacob Kitzman