PACIFIC BIOSCIENCES®

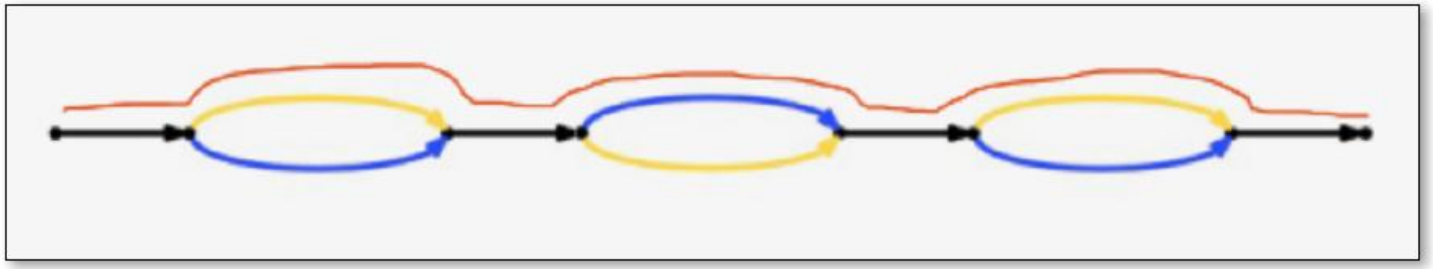
# FALCON-Phase: a diploid assembler for PacBio and Hi-C data

Collaboration between PacBio and Phase Genomics, 2018

# THE NEED FOR HAPLOTYPE-RESOLVED ASSEMBLIES

## Persistent Paradigm in Genome Assembly

- Choose an inbred individual to make assembly simpler and more attainable
- When encountering heterozygosity in genome, “collapse” haplotypes into a single mosaic consensus representation of the genome



Longest path mixing haplotypes (blue and yellow) to generate a “pseudohaplotype”

Image from Weisenfeld et al. 2017

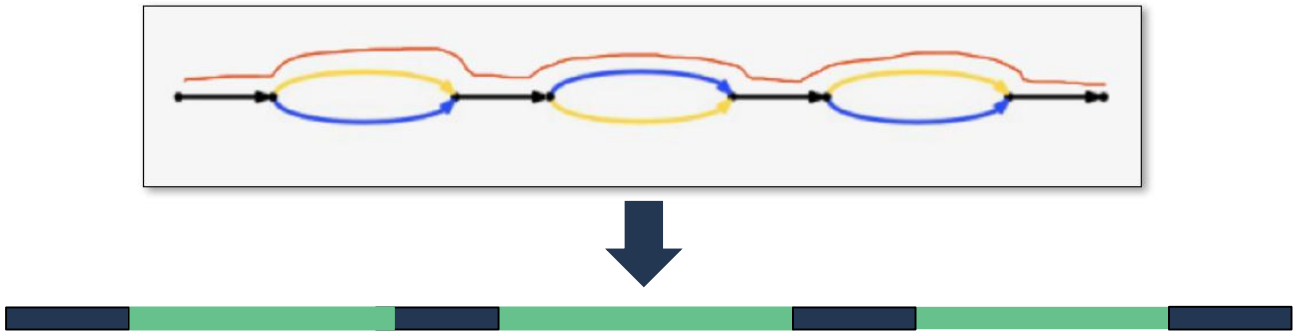
# THE NEED FOR HAPLOTYPE-RESOLVED ASSEMBLIES

## Persistent Paradigm in Genome Assembly

- Choose an inbred individual to make assembly simpler and more attainable
- When encountering heterozygosity in genome, “collapse” haplotypes into a single mosaic consensus representation of the genome

## Issues with Paradigm

- Choosing an inbred individual misrepresents the genetic diversity present within the organism and/or species
- Collapsed haplotype assemblies result in “Franken-haplotypes” that do not accurately represent the genome of interest



Resulting linear genome of haplotype collapsing approach to assembly

Image from Rhie, Koren, Phillippy PAG 2018 and 10X Genomics

# THE NEED FOR HAPLOTYPE-RESOLVED ASSEMBLIES

## Persistent Paradigm in Genome Assembly

- Choose an inbred individual to make assembly simpler and more attainable
- When encountering heterozygosity in genome, “collapse” haplotypes into a single mosaic consensus representation of the genome

## Issues with Paradigm

- Choosing an inbred individual misrepresents the genetic diversity present within the organism and/or species
- Collapsed haplotype assemblies result in “Franken-haplotypes” that do not accurately represent the genome of interest

## Advantages to Phased Diploid Assembly

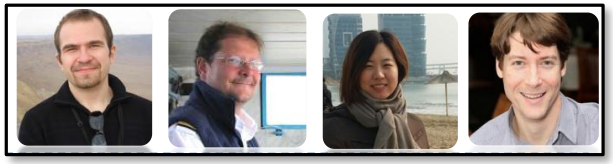
- SNP and SV characterization can be done in context with haplotype
- Gene models and annotations will be more accurate
- Targeted use of the genome will be less error-prone

Phased diploid assembly

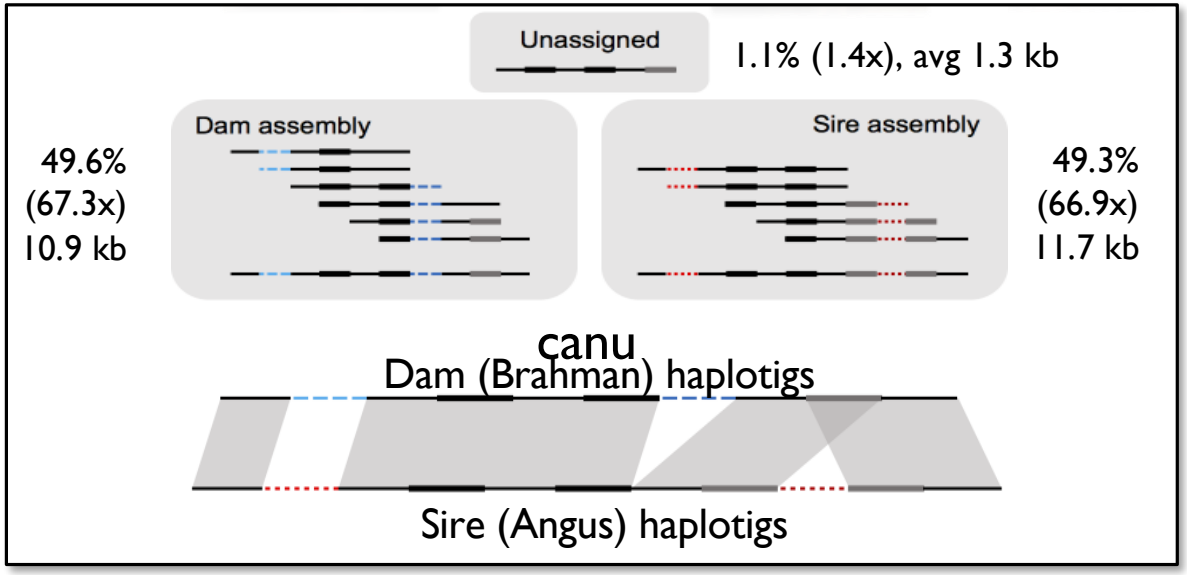
# CURRENT APPROACHES TO LONG READ DIPLOID ASSEMBLY

## 1. Trio Binning with TrioCanu

- PacBio data for F1
- ILMN data for parents
- Identify parent-specific markers to bin PacBio reads
- Perform two haploid Canu assemblies



Example on F1 cattle hybrid




Koren, S. et al. (2018). Complete assembly of parental haplotypes with trio binning. bioRxiv. Available from: https://doi.org/10.1101/271486

# CURRENT APPROACHES TO LONG READ DIPLOID ASSEMBLY

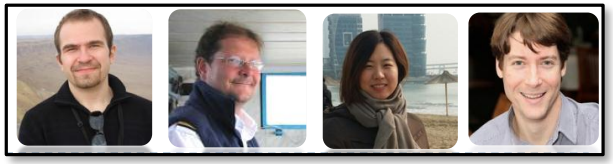
## 1. Trio Binning with TrioCanu

- PacBio data for F1
- ILMN data for parents
- Identify parent-specific markers to bin PacBio reads
- Perform two haploid Canu assemblies

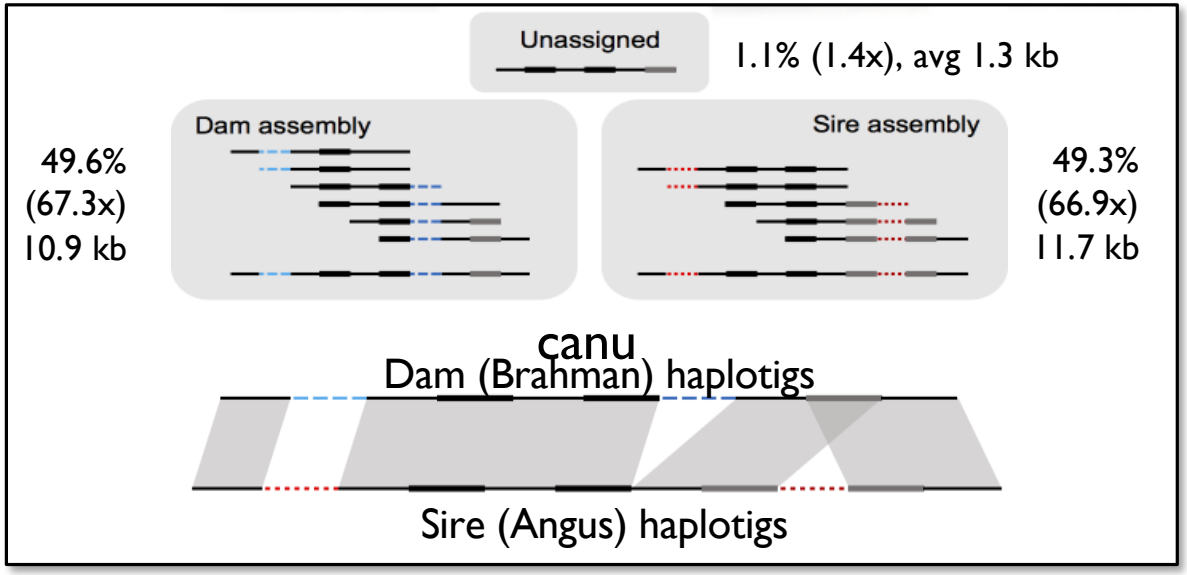
**Requires a second data type & both parents of individual of interest**

**Additional BFX for identifying parent-specific k-mers**

**Additional compute required to run two full assemblies**



Example on F1 cattle hybrid



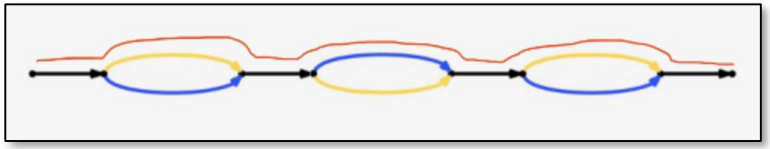

Koren, S. et al. (2018). Complete assembly of parental haplotypes with trio binning. bioRxiv. Available from: https://doi.org/10.1101/271486

# CURRENT APPROACHES TO LONG READ DIPLOID ASSEMBLY

## 2. Haplotype assembly with FALCON-Unzip

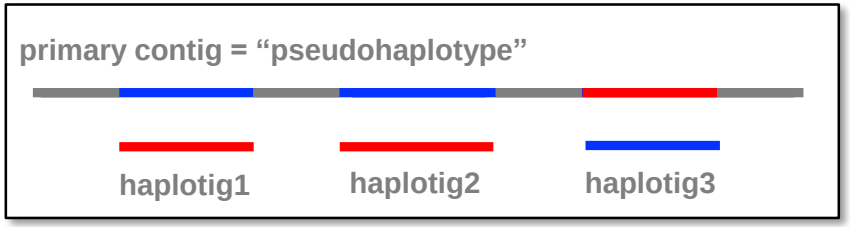
- PacBio data for diploid individual
- Perform FALCON and FALCON-Unzip assemblies
- Outputs primary contigs (longest path through bubbles and collapsed regions) and haplotigs (divergent alleles from bubble regions)

Phase reads in Assembly Graph



Weisenfeld et al. 2017

Emit Psuedohaplotype and Haplotigs




Chin, C.S. et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 13(12), 1050.

# CURRENT APPROACHES TO LONG READ DIPLOID ASSEMBLY

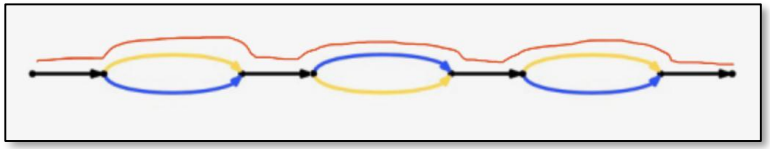
## 2. Haplotype assembly with FALCON-Unzip

- PacBio data for diploid individual
- Perform FALCON and FALCON-Unzip assemblies
- Outputs primary contigs (longest path through bubbles and collapsed regions) and haplotigs (divergent alleles from bubble regions)

**Length of phase block is a function of heterozygosity (biology) and read length (technology)**

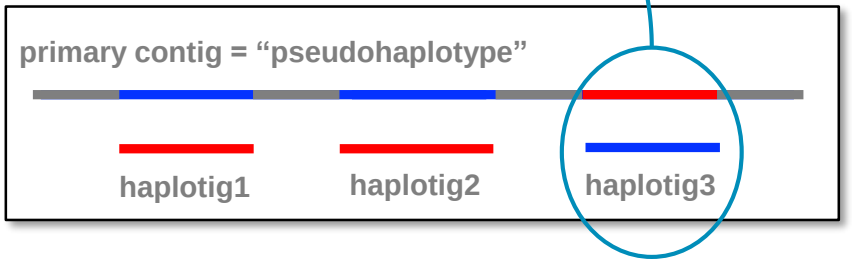
**Phase-switches can occur along primary contig**

Phase reads in Assembly Graph



Weisenfeld et al. 2017

Emit Psuedohaplotype and Haplotigs




Chin, C.S. et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 13(12), 1050.

# *NEW!* FALCON-PHASE FOR PHASED DIPLOID GENOME ASSEMBLIES

**Features:**

- Addresses haplotype/phase switches in FALCON-Unzip assemblies improving base accuracy, gene predictions, scaffolding errors, and eliminating “Franken-haplotypes”
- Increases the length of phase blocks in FALCON-Unzip assemblies
- Utilizes long-range HiC data for phasing & can be run iteratively on contigs and scaffolds
- Easy-to-use, open source software available on Github
- Collaboration with Phase Genomics

PHASE GENOMICS

Kronenberg, Z. et al. (2018). FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. bioRxiv. Available from: https://doi.org/10.1101/327064

# *NEW!* FALCON-PHASE INTEGRATES HI-C WITH FALCON-UNZIP

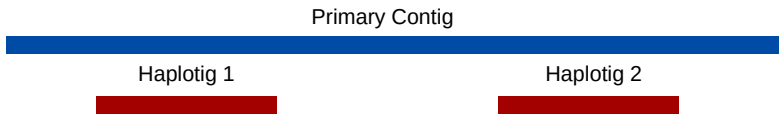
## FALCON-Phase Process

***Input*** a FALCON-Unzip assembly & HiC data

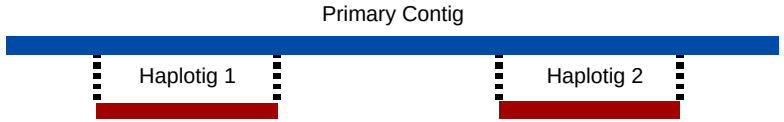
# *NEW!* FALCON-PHASE INTEGRATES HI-C WITH FALCON-UNZIP

## FALCON-Phase Process

***Input*** a FALCON-Unzip assembly & HiC data



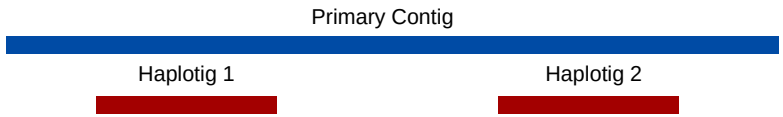

1. Identify placement of haplotigs on primary contigs



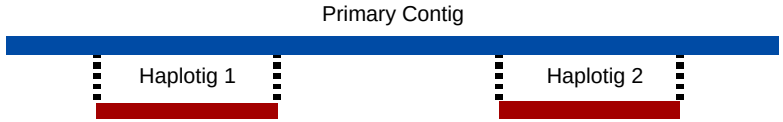
# *NEW!* FALCON-PHASE INTEGRATES HI-C WITH FALCON-UNZIP

## FALCON-Phase Process

***Input*** a FALCON-Unzip assembly & Hi-C data




1. Identify placement of haplotigs on primary contigs

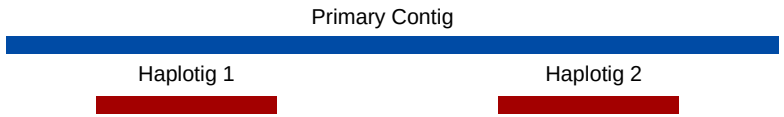



2. Mince primary contigs at edges of haplotig placements to separate homologous haplotigs and collapsed haplotype

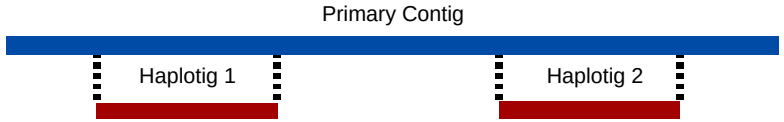
# *NEW!* FALCON-PHASE INTEGRATES HI-C WITH FALCON-UNZIP

## FALCON-Phase Process

***Input*** a FALCON-Unzip assembly & Hi-C data

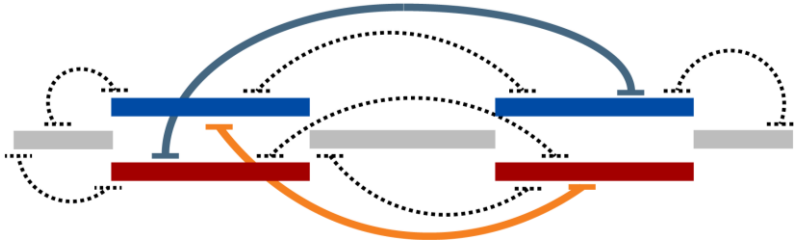



1. Identify placement of haplotigs on primary contigs




2. Mince primary contigs at edges of haplotig placements to separate homologous haplotigs and collapsed haplotype

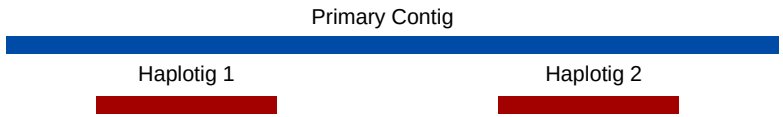

3. Map paired Hi-C reads to minced contigs to identify haplotype switches



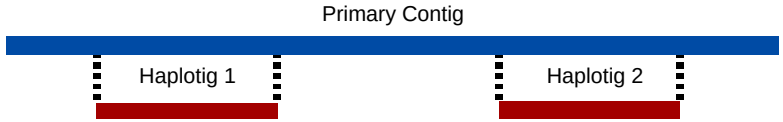
# *NEW!* FALCON-PHASE INTEGRATES HI-C WITH FALCON-UNZIP

## FALCON-Phase Process

***Input*** a FALCON-Unzip assembly & Hi-C data

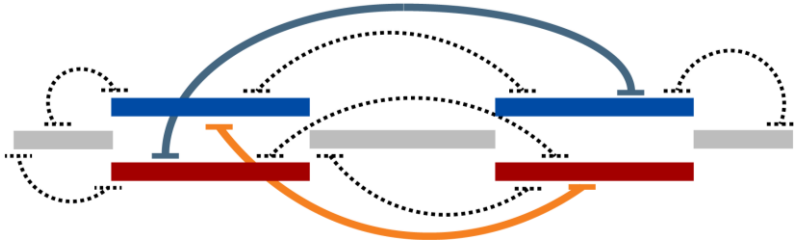



1. Identify placement of haplotigs on primary contigs

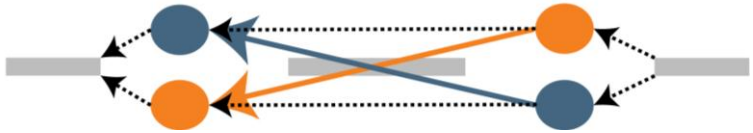



2. Mince primary contigs at edges of haplotig placements to separate homologous haplotigs and collapsed haplotype


3. Map paired Hi-C reads to minced contigs to identify haplotype switches



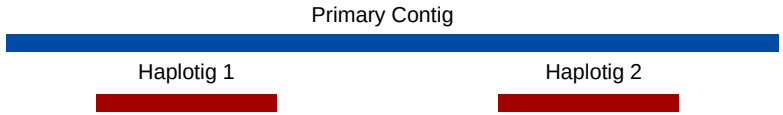
4. Combine same-phase haplotigs and collapsed haplotype contigs into gap-less phased contigs



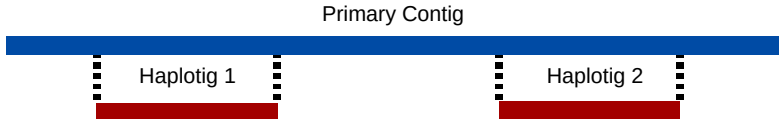
# *NEW!* FALCON-PHASE INTEGRATES HI-C WITH FALCON-UNZIP

## FALCON-Phase Process

***Input*** a FALCON-Unzip assembly & Hi-C data

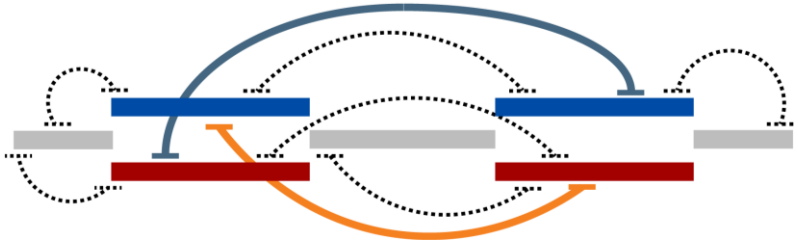



1. Identify placement of haplotigs on primary contigs

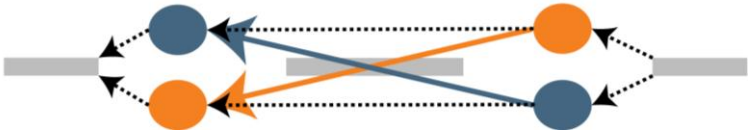



2. Mince primary contigs at edges of haplotig placements to separate homologous haplotigs and collapsed haplotype

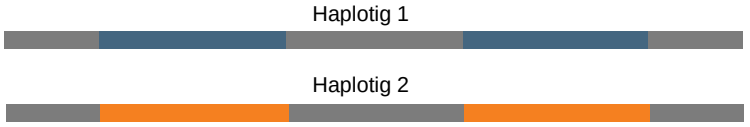

3. Map paired Hi-C reads to minced contigs to identify haplotype switches



4. Combine same-phase haplotigs and collapsed haplotype contigs into gap-less phased contigs

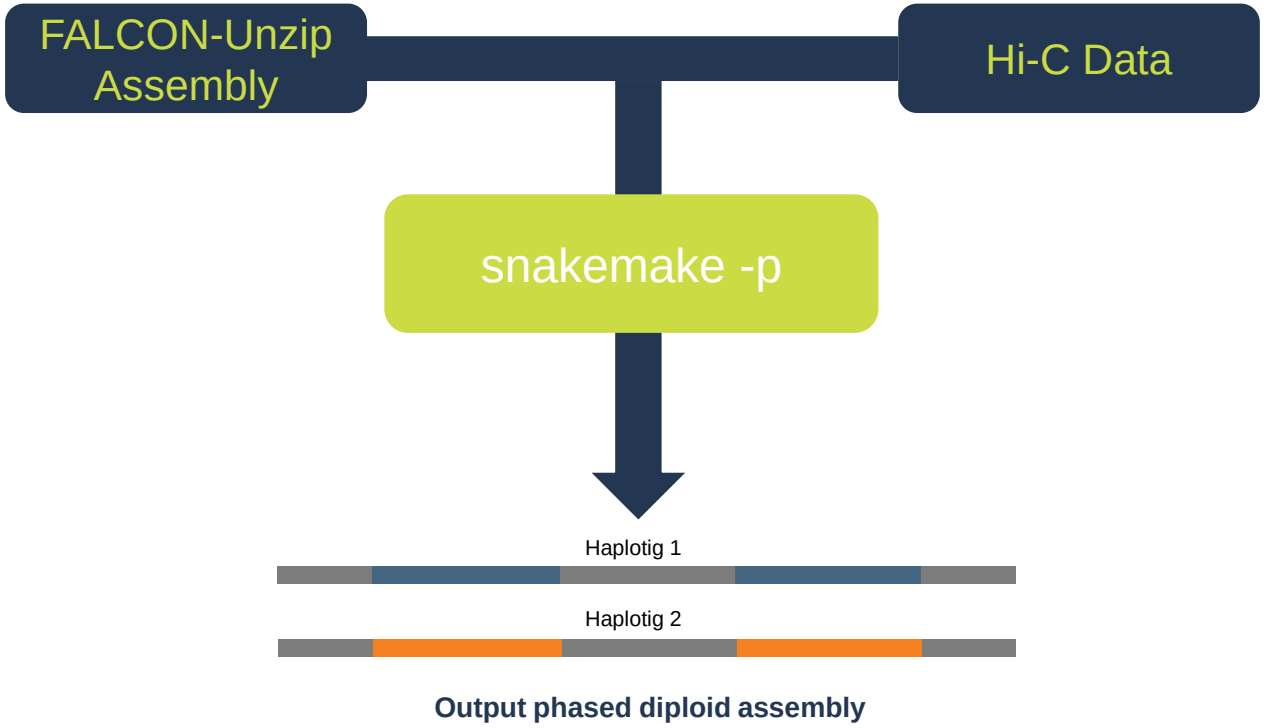


***Output*** a phased diploid assembly



# *NEW!* FALCON-PHASE INTEGRATES HI-C WITH FALCON-UNZIP

## FALCON-Phase Workflow




**Output phased diploid assembly**

While it is a command line tool, it can be run on a laptop with no compute cluster and uses standard bioinformatics dependencies, making it as easy-to-use as it is powerful for phased genome assembly.

Kronenberg, Z. et al. (2018). FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. bioRxiv. Available from: https://doi.org/10.1101/327064

# FALCON-PHASE FOR PHASED DIPLOID GENOME ASSEMBLIES

## Conclusions:

- Only method to combine the megabase size contigs of PacBio assemblies and long range HiC information to give phased diploid assemblies
- Easy-to-use single command tool that can be run iteratively on contigs and scaffolds
- Does ***not*** require compute cluster resources
- Can be run retroactively on FALCON-Unzip assemblies with HiC data from any source

For more information on FALCON-Phase please see the the GitHub page and the preprint on bioRxiv.

PACBIO®

www.pacb.com