Structural Variation with PacBio Data

Ali Bashir

Overview

- Very low (<1X) to low (<5X) Coverage Methods
  - Alpha Satellite Detection
    - Alpha-centauri (unpublished)
  - Generalized Breakpoint Approaches
    - MultiGASV
- Medium Coverage (<10X) to High-Coverage
  - Tandem Repeat Detection
    - PacMonSTR
  - Local assembly
  - Hybrid Approaches with other technologies

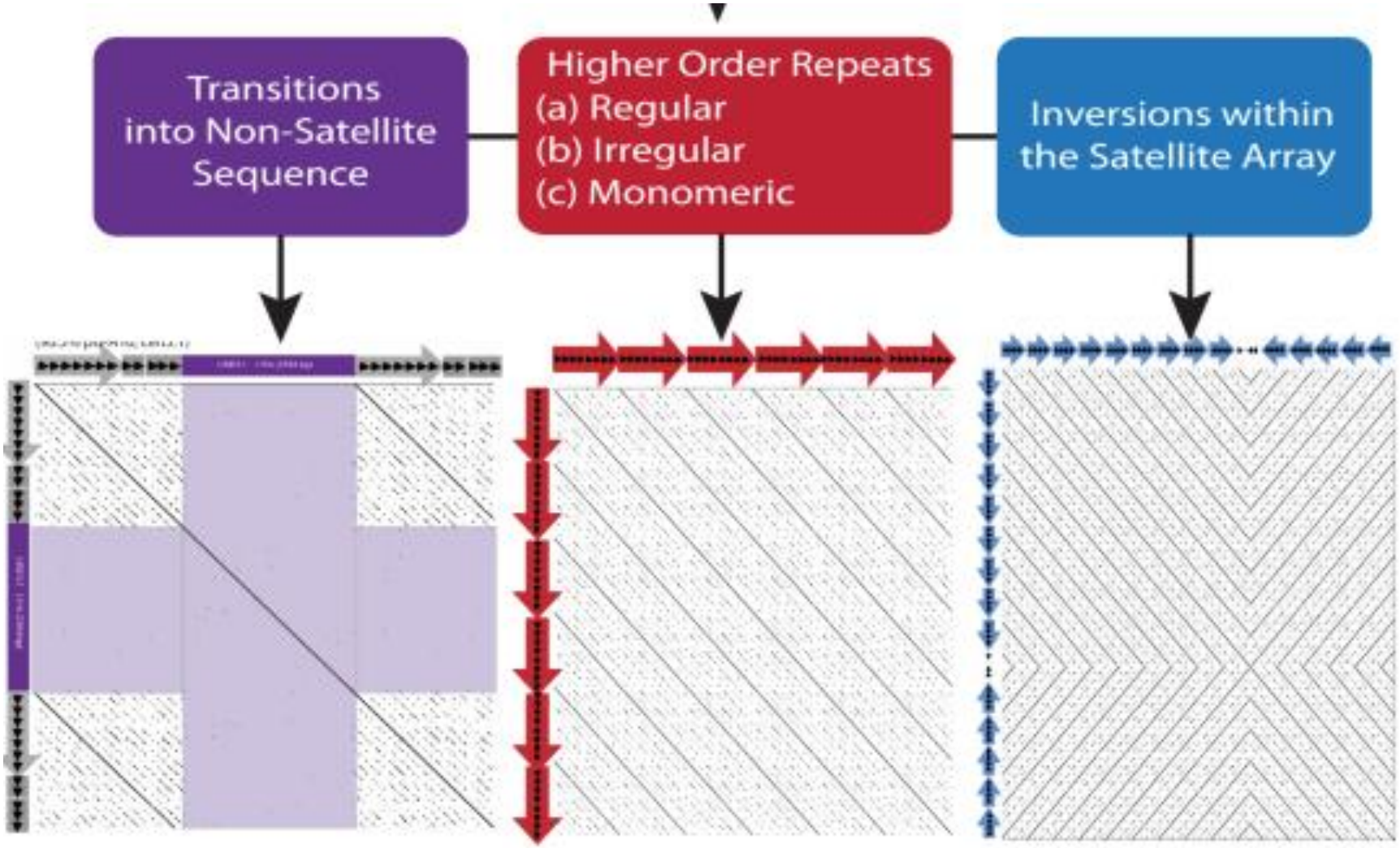
α-Centauri

alpha-satellite and high order repeat structure (in development)

https://github.com/volkansevim/alpha-CENTAURI

Volkan Sevim, Jason Chin, Karen Miga

Automatic Alpha Satellite Identifier and Classifier



MultiBreak-SV

Breakpoint-based SV Detection with Long and short-reads

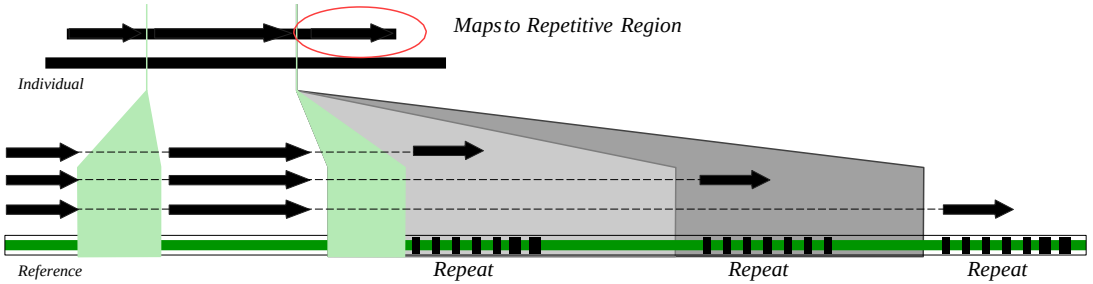
https://github.com/raphael-group/multibreak-sv

Anna Ritz, Suzanne Sindi, Iman Hajirasouliha, Ben Raphael

Ritz et al. *Bioinformatics*. 2014
Ritz et al. *Bioinformatics*. 2010

SV Detection in Human Genomes

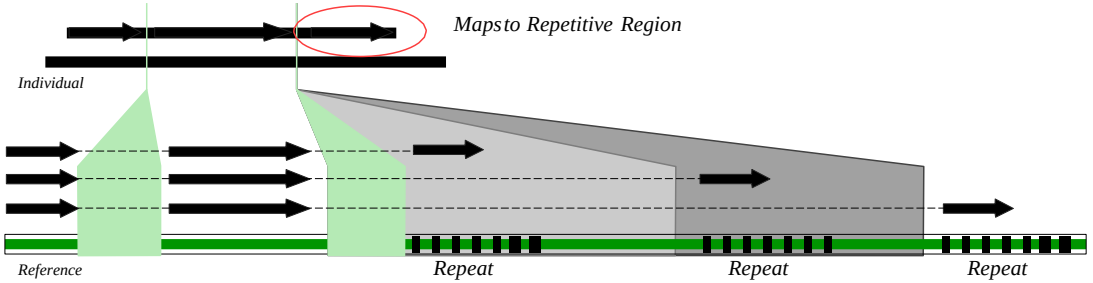
Mammalian genomes contain repetitive regions



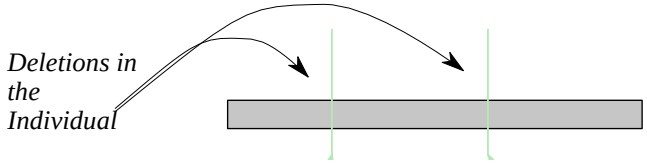
MultiBreakSV expands on conventional paired-read and split read methods

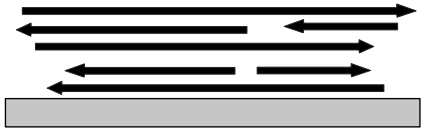
SV Detection in Human Genomes

Mammalian genomes contain repetitive regions

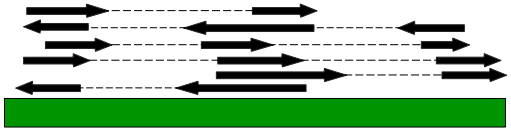


MultiBreak-SV





1. Sequence reads from individual genome using any sequencing platform

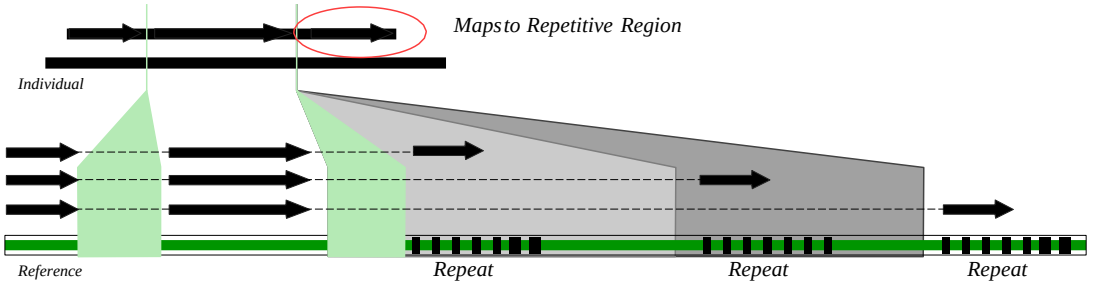


2. Align reads to reference genome; reads may map to different locations

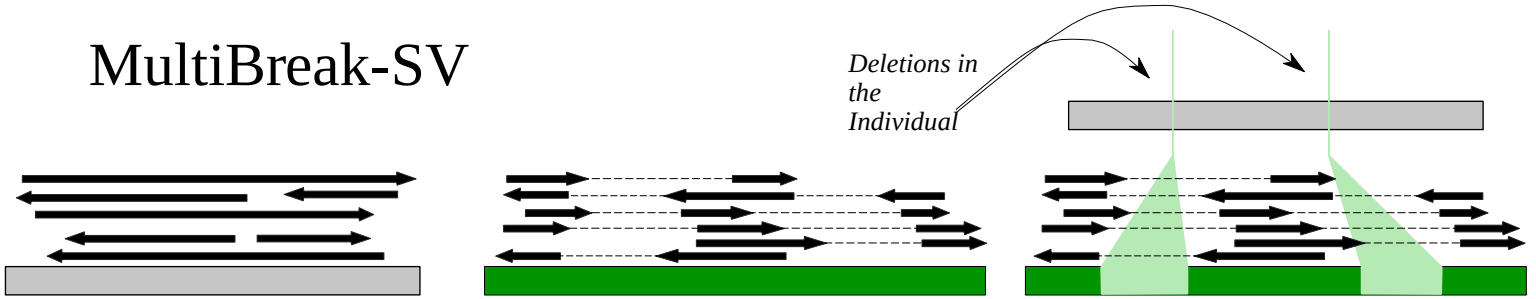
SV algorithms for paired reads [Sindi2012,Quinlan2010,Hormozdiari2009,. . .]
SV algorithm for multi-linked reads [Ritz2010]

SV Detection in Human Genomes

Mammalian genomes contain repetitive regions



MultiBreak-SV



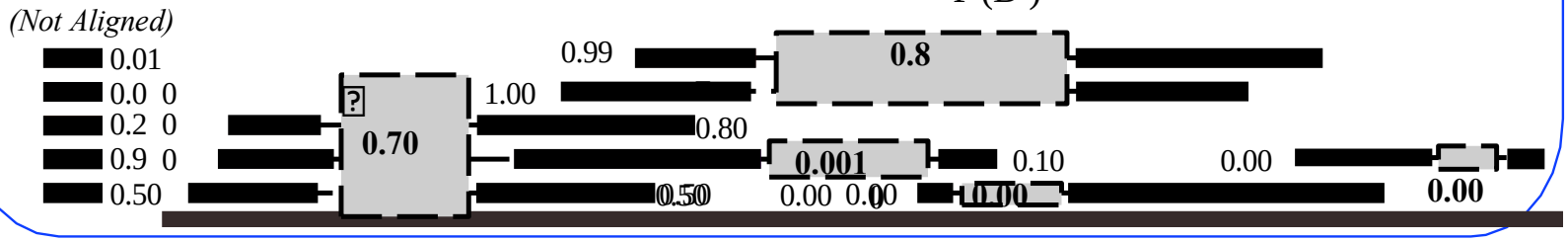
1. Sequence reads from individual genome using any sequencing platform

2. Align reads to reference genome; reads may map to different locations

3. Determine SVs from the alignments and compute novel adjacency probabilities

Mapping M : Select an alignment (or no alignment) for each read
Data D : Read alignments and novel adjacencies

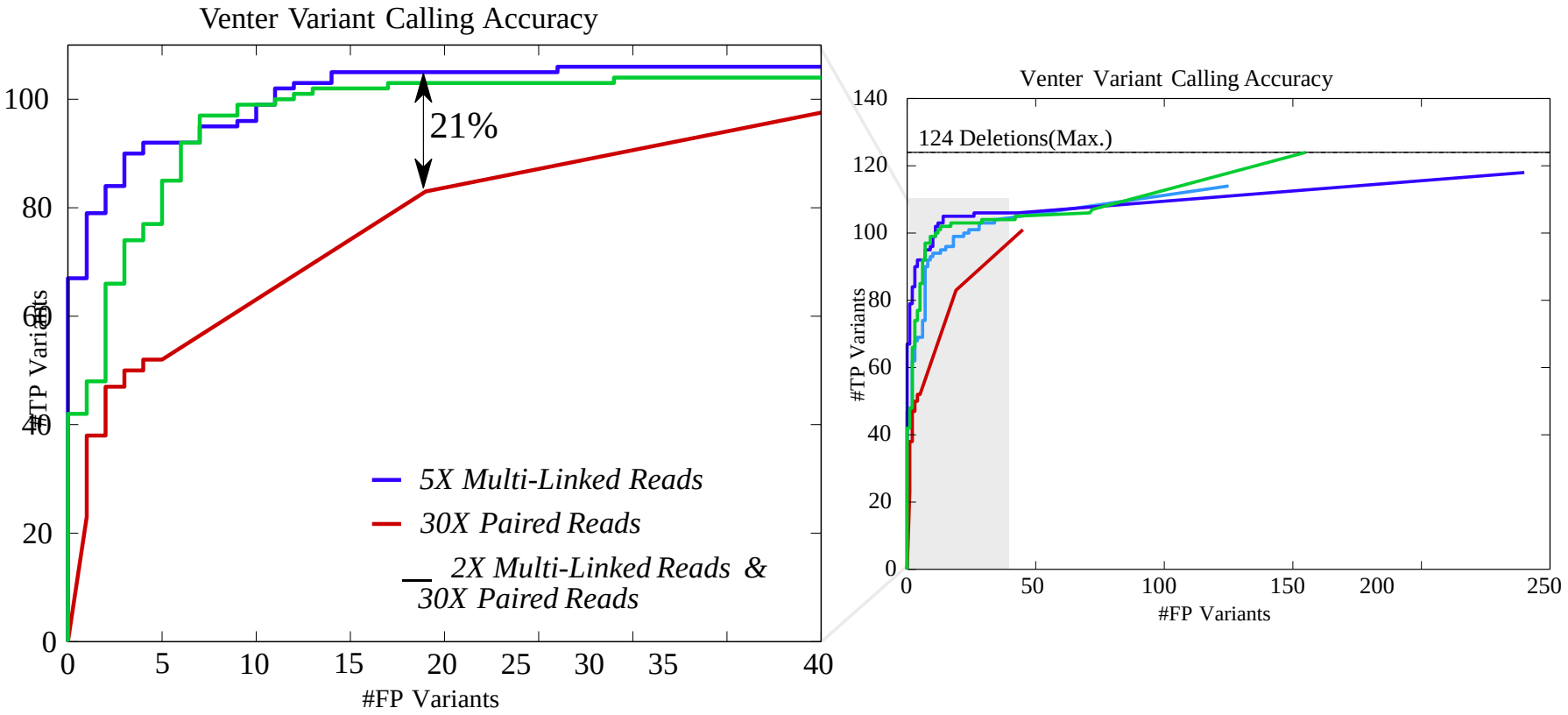
$$P(M \mid D) = \frac{P(\text{alignments} \mid M)\,P(\text{adjacencies} \mid M)\,P(M)}{P(D)}$$

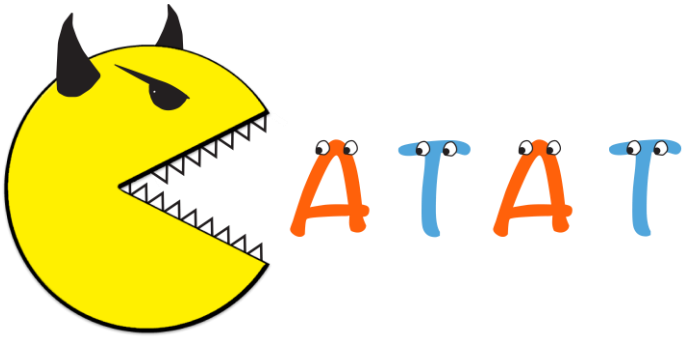


Benchmark on Simulated Chr17

Venter Chromosome 17

- Inserted 17K of Venter's variants into a reference Chr17
- Assess 124 deletions $\geq$ 120bp (112 located in repetitive regions)





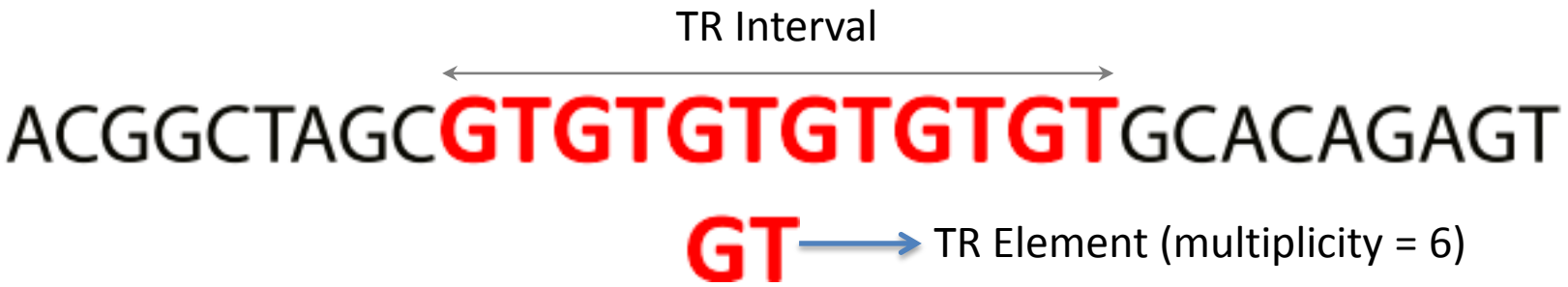
PacMonSTR
Tandem Repeat Resolution

https://github.com/alibashir/pacmonstr

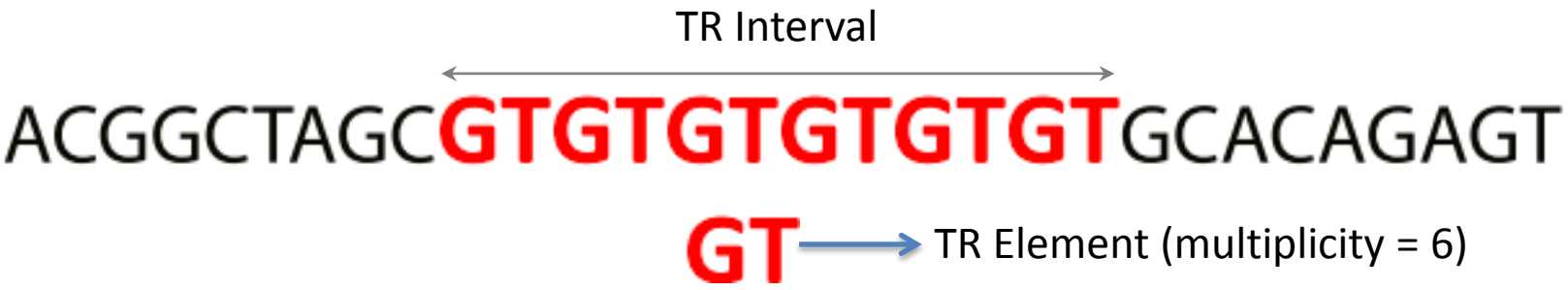
Ajay Ummat, Matthew Pendleton

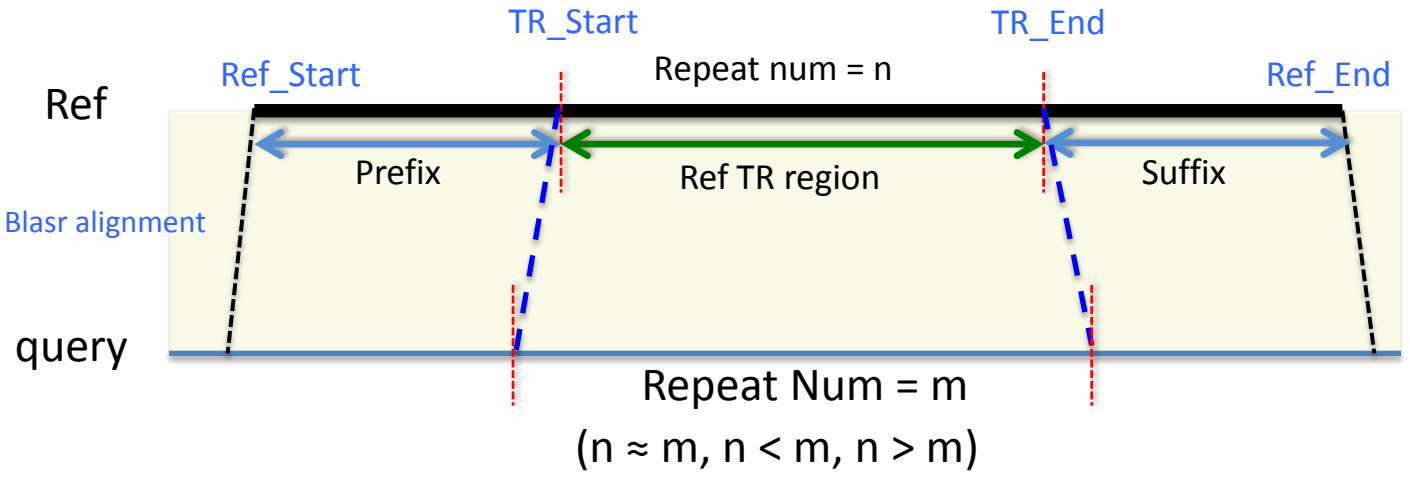
Ummat and Bashir. *Bioinformatics*. 2014

Tandem Repeats Detection Model



Tandem Repeats Detection Model





Any statistically significant deviation between *n* and *m* is an interesting result!

- A number of tools (lobSTR, repeatSeq) have used this ideas to make robust callers for NGS data

Probabilistic estimation of TR via *pairHMM*

Problem

$q = TCTCTCCTTACTCCCTCTTCC...TCCTC$ → *Error prone read*

$tr = TC$ → *Repeat element*

↓

Estimate number of '*tr*' in '*q*'

Observation

Sequence alignment of '*q*' and '*tr*' generates a path or a *pair*wise alignment

$$\frac{q}{tr} \rightarrow \frac{T}{T}\frac{C}{C}\frac{T}{T}\frac{C}{C}\frac{T}{-}\frac{C}{C}\frac{-}{T}\cdots\frac{q_i}{tr_j}\cdots\frac{C}{C}$$

Number of '*tr*' in '*q*' is related to such a pairwise alignment

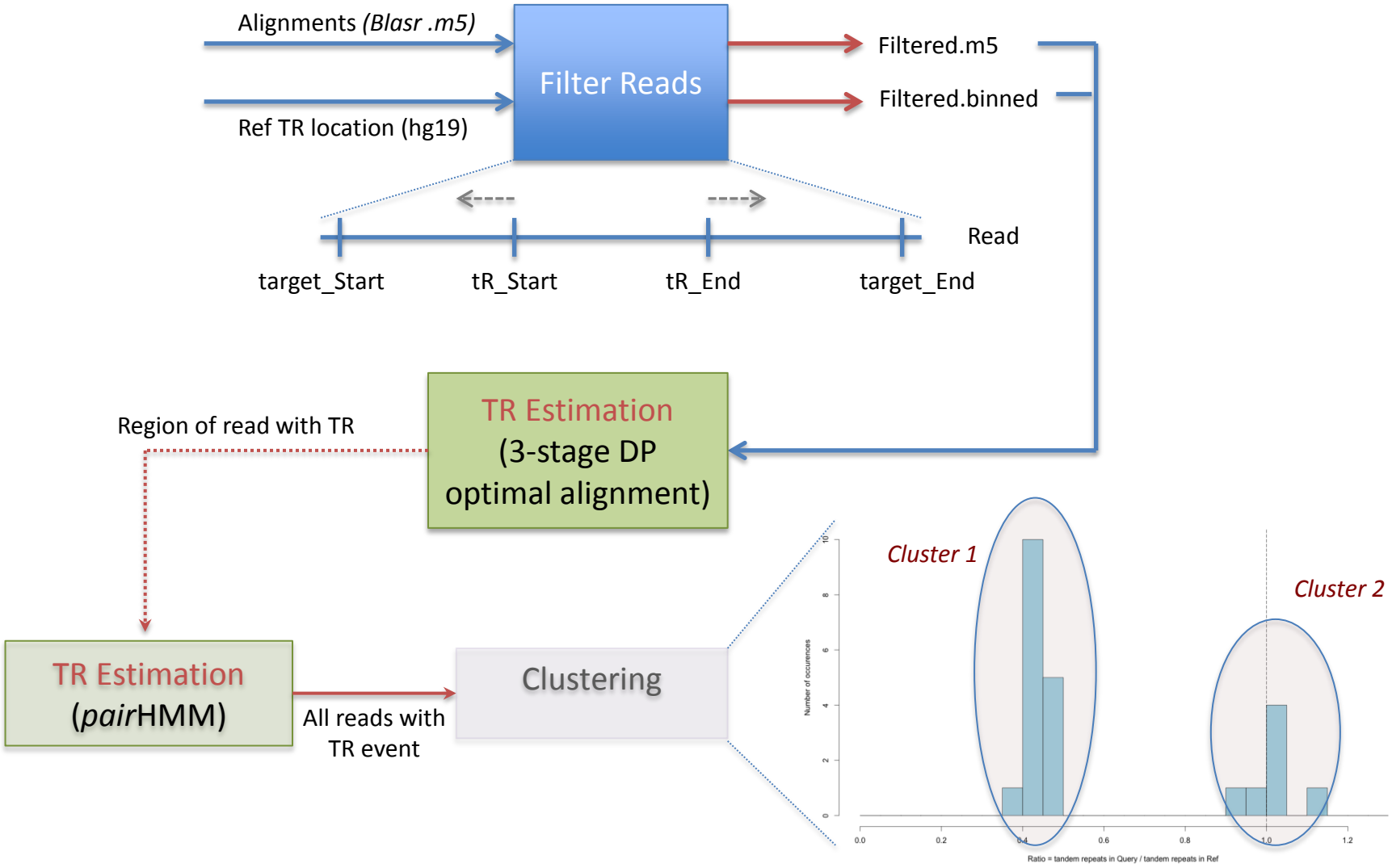
A possible solution

*pair*HMM can assign probabilities to any pairwise alignment

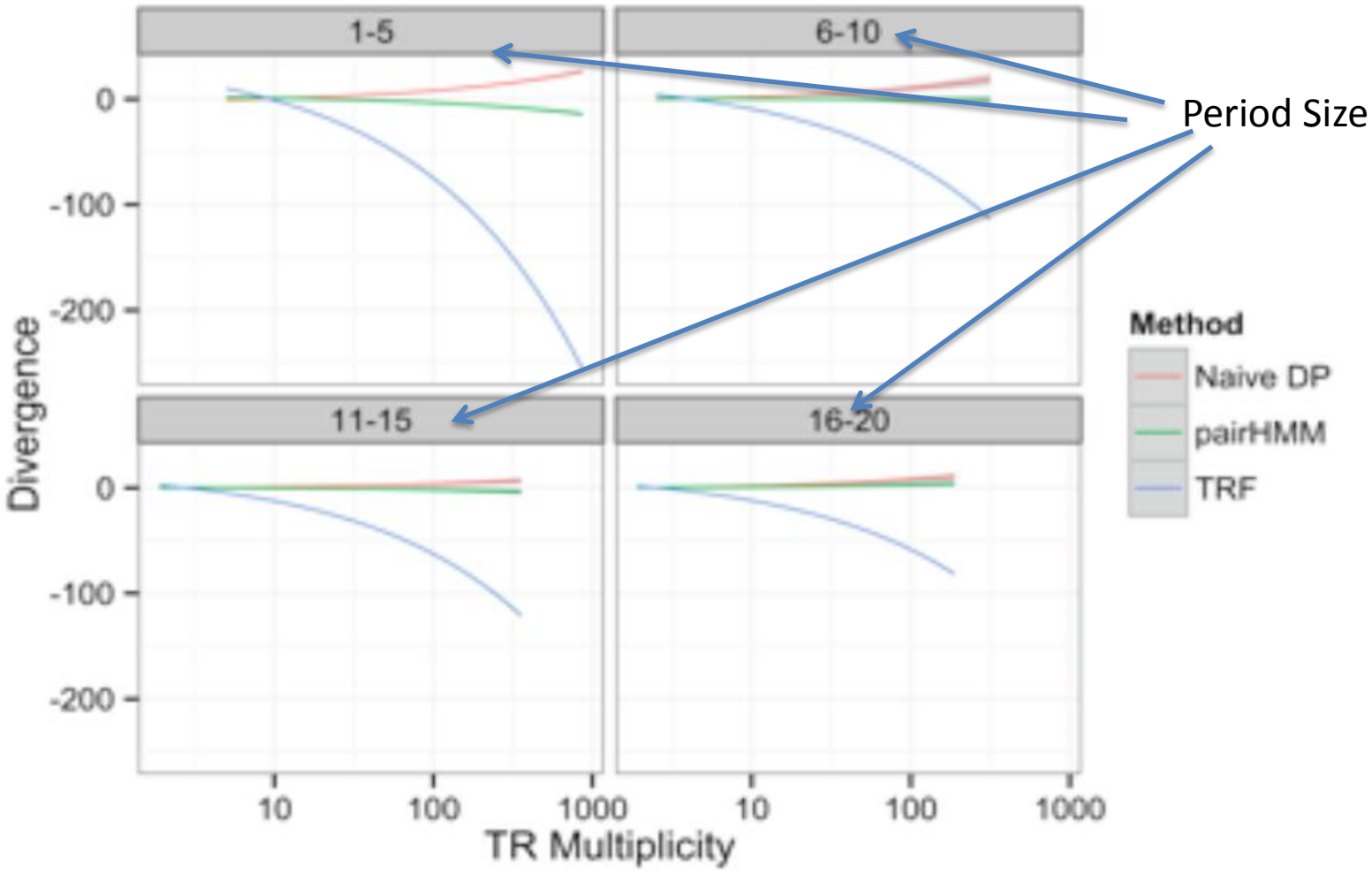
Provide a probabilistic way to estimate the number of '*tr*' in '*q*'

Durbin et al., 1998
Rabiner, 1989

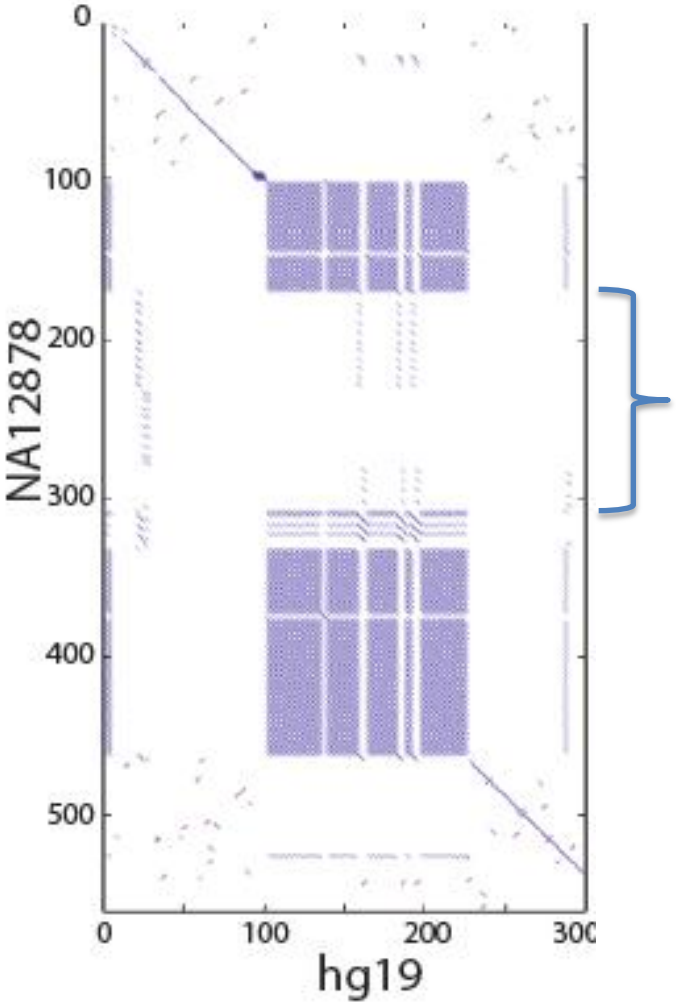
TR estimation pipeline



Comparing Methods

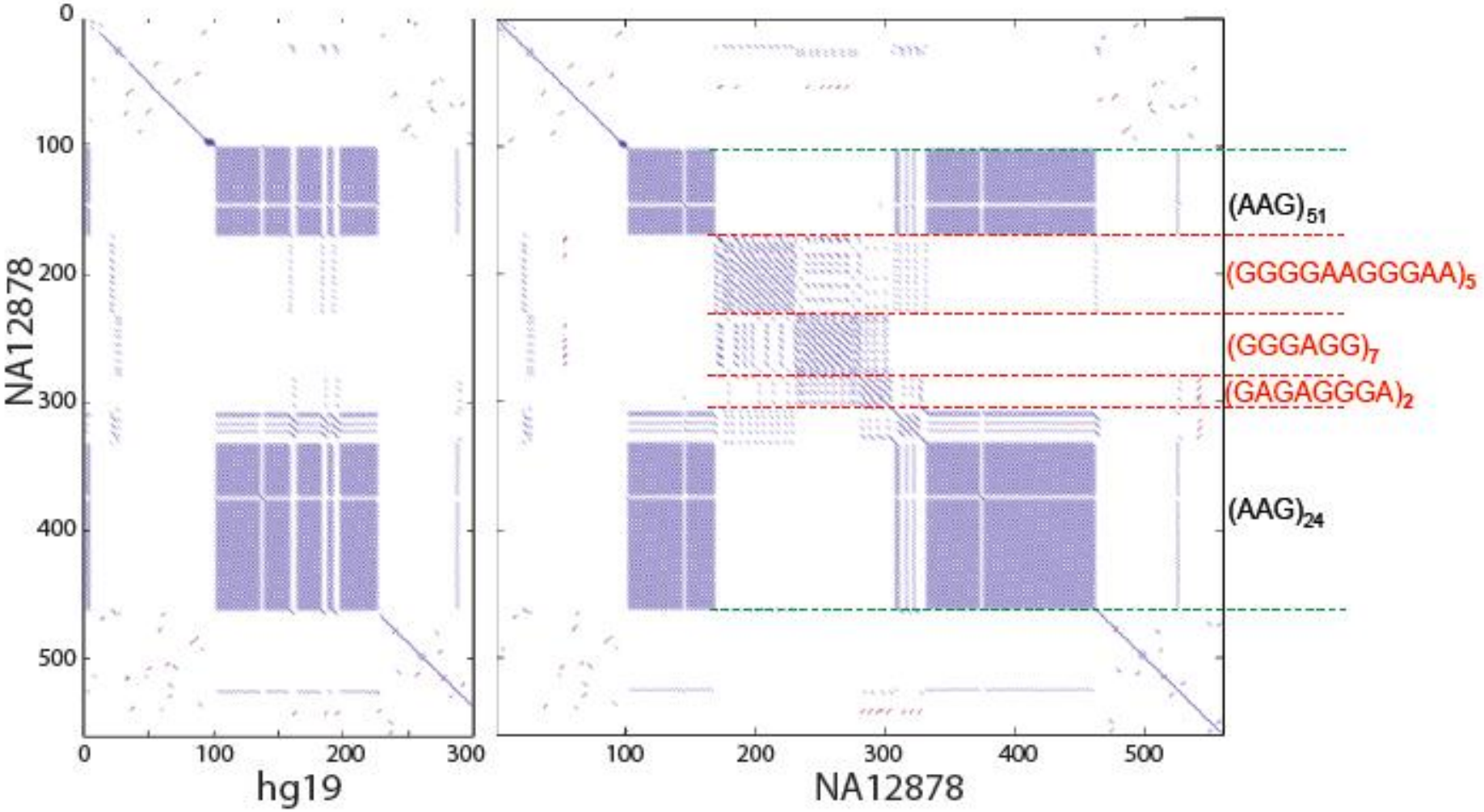


Sometimes multiple distinct TR states are identified



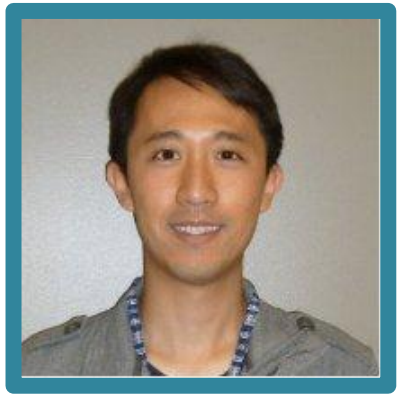
This interval clearly appears to be different (and expanded) in our sequence relative to the reference

Sometimes multiple distinct TR states are identified

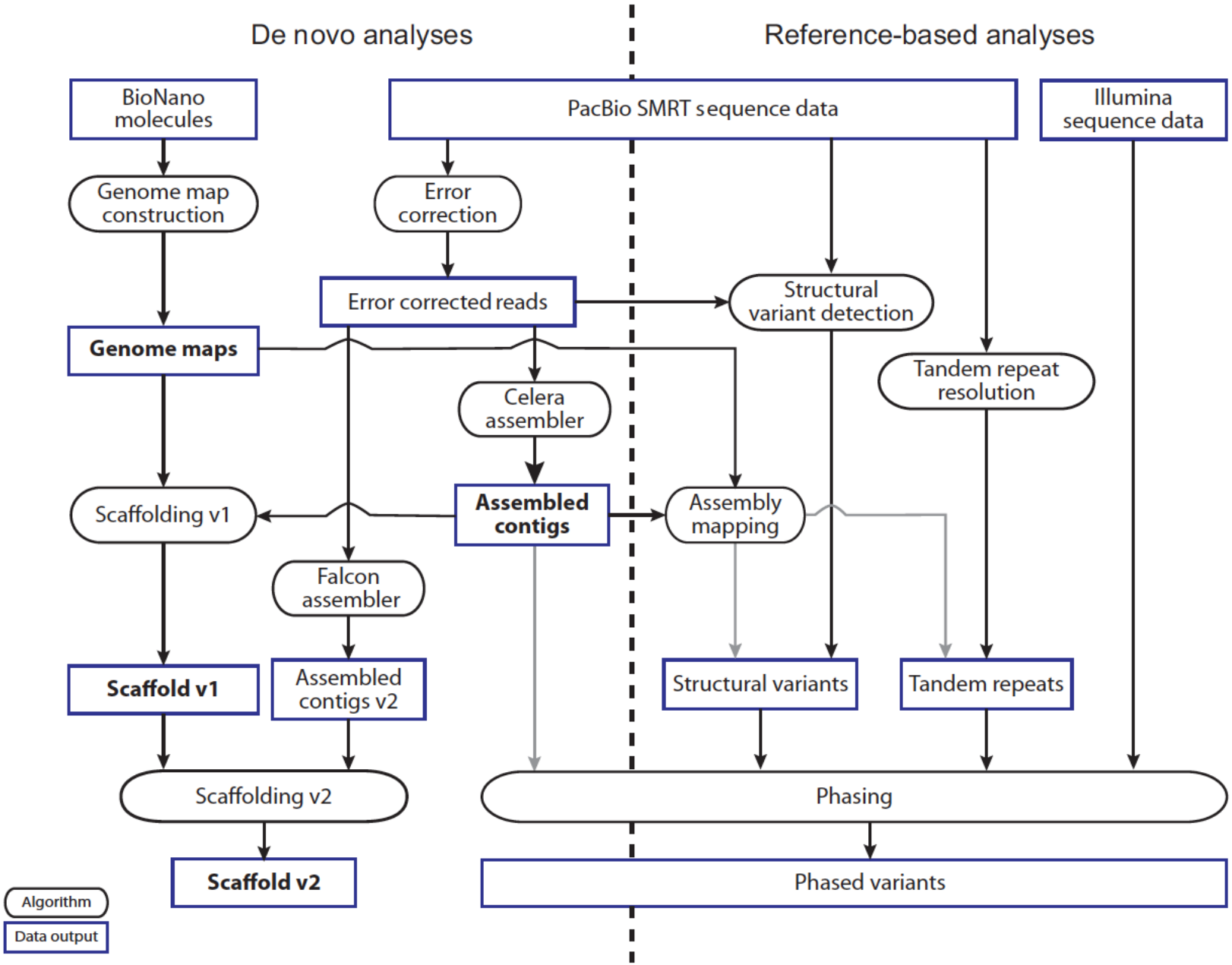


Assembly Based SV calling and integrating PacBio with other technologies

Many people, but on the software side:
Andy Pang, Heng Dai, Matthew Pendleton

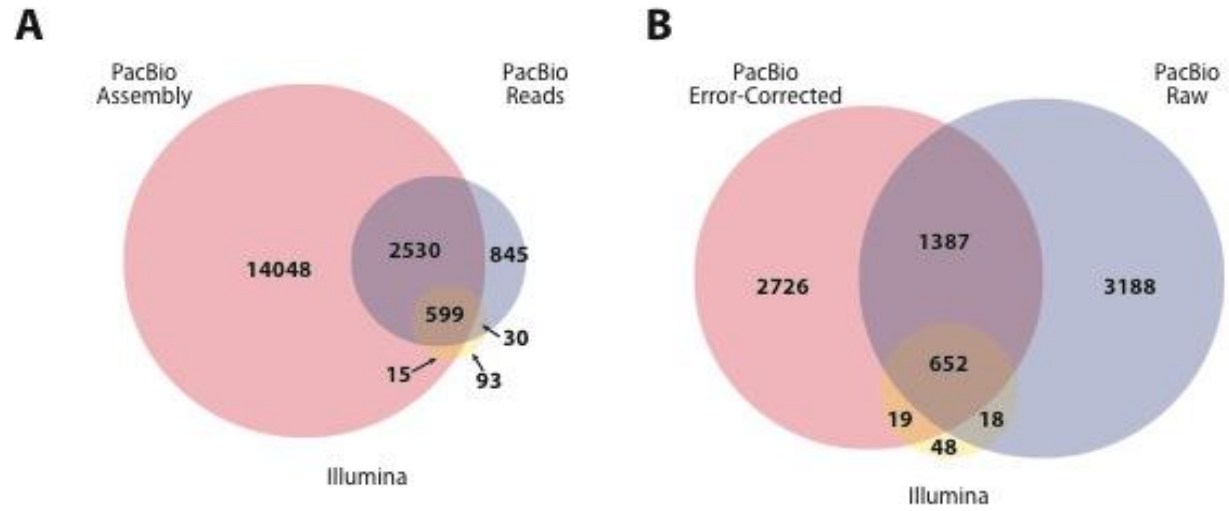


Pendleton et al. *Nature Methods*. 2015

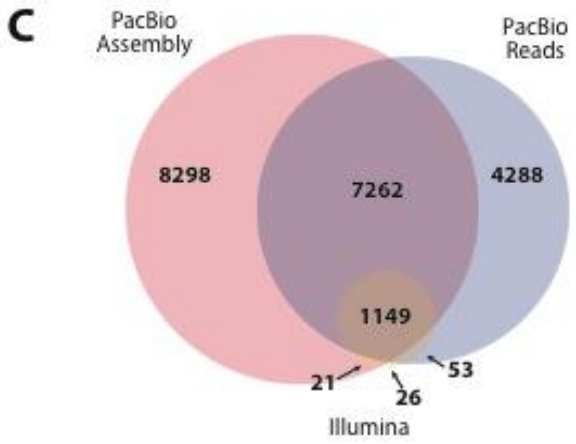

De novo analyses
Reference-based analyses
BioNano molecules
PacBio SMRT sequence data
Illumina sequence data
Genome map construction
Error correction
Error corrected reads
Structural variant detection
Genome maps
Tandem repeat resolution
Celera assembler
Scaffolding v1
Assembled contigs
Assembly mapping
Falcon assembler
Scaffold v1
Assembled contigs v2
Structural variants
Tandem repeats
Scaffolding v2
Phasing
Scaffold v2
Phased variants
Algorithm
Data output

Orthogonal Methods provide many new calls

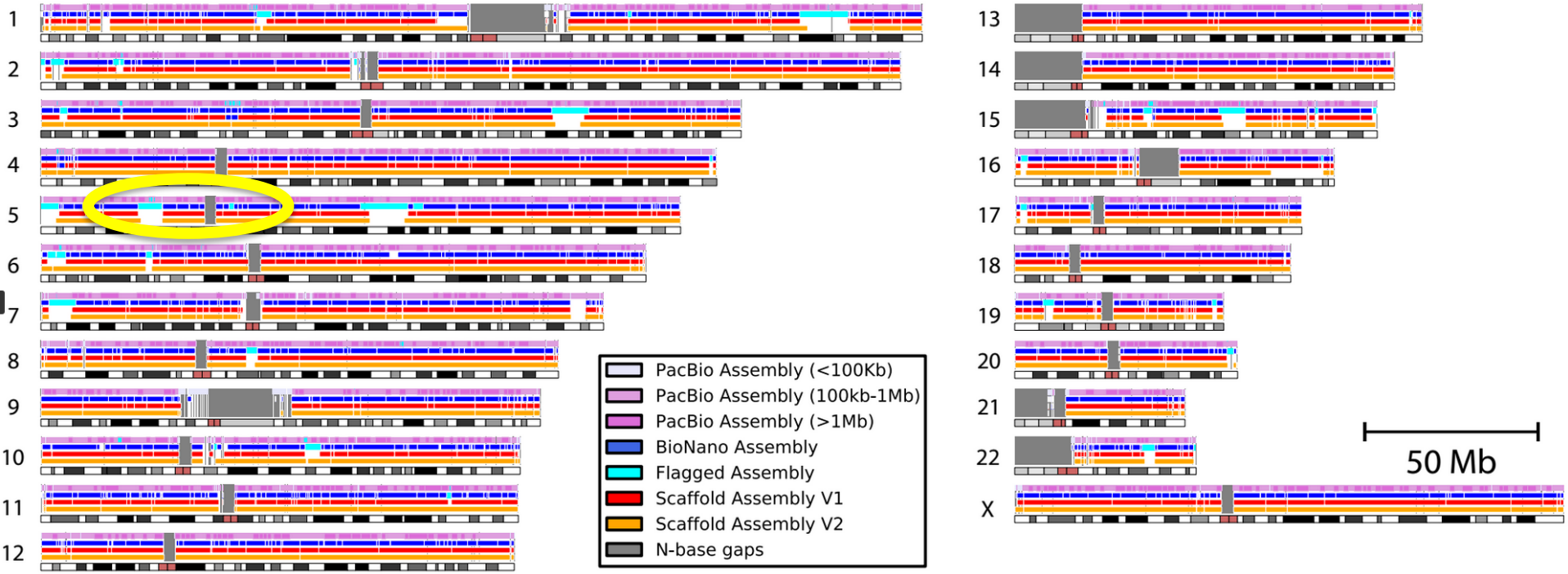
Deletions

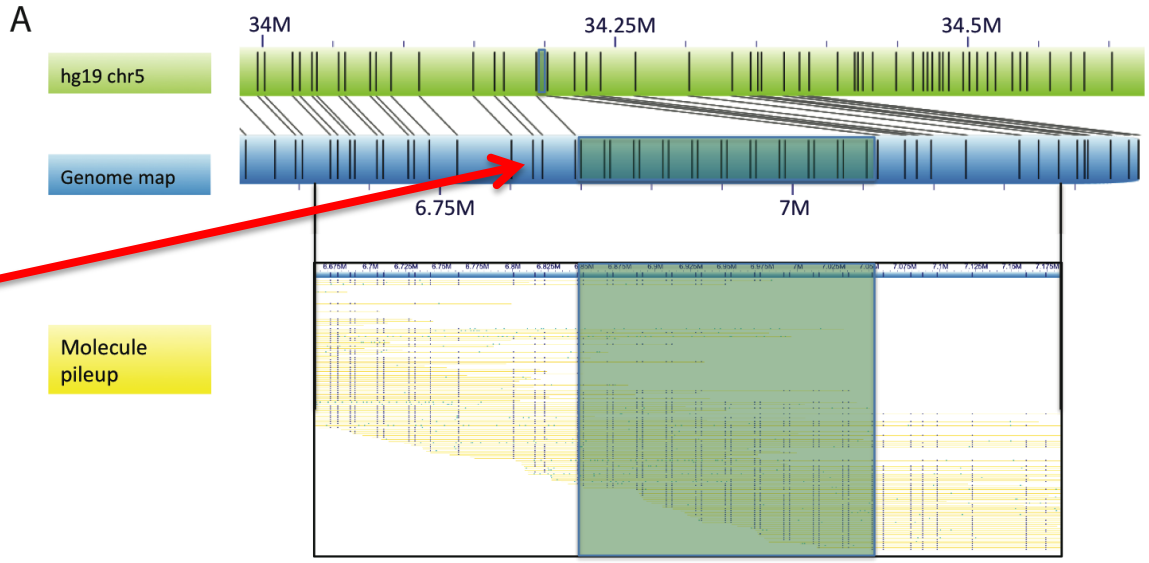


Mobile Element Insertions



Hybrid Approach Improves Assembly and reveals large SVs





Even large (>100kb)
TRs can be resolved!

Acknowledgements

- Mount Sinai
  - **Eric Schadt**
  - **Matt Pendleton**
  - **Ajay Ummat**
  - Oscar Franzen
  - Gintaras Deikus
  - Ariella Cohain
  - **Robert Sebra**
- BioNano Genomics
  - Han Cao
  - Alex Hastie
  - Heng Dai
  - **Andy Pang**
  - Will Stedman
  - Thomas Anantharaman
- PacBio
  - **Jason Chin**
  - Ellen Paxinos
  - Yan Guo
  - Jonas Korlach
  - Volkan Sevim
- Weill Cornell
  - Christopher Mason
  - Scott Blanchard
  - Russel Durret
  - Roger Altman
- CSHL
  - Richard McCombie
- Rockefeller
  - Bob Darnell
- UCSF
  - Pui Kwok
- EMBL
  - Jan Korbel
  - Markus H.-Y. Fritz
  - Tobias Rausch
- UCSC
  - Karen Miga
- 1000 Genomes SV Working Group