Synthetic long read technologies in genome phasing and beyond

Volodymyr Kuleshov

Stanford University

Batzoglou & Snyder Labs

nature biotechnology

ARTICLES

Whole-genome haplotyping using long reads and statistical methods

Volodymyr Kuleshov[1–4], Dan Xie[3,4], Rui Chen[3,4], Dmitry Pushkarev[2,4], Zhihai Ma[3], Tim Blauwkamp[2], Michael Kertesz[2] & Michael Snyder[3]

The rapid growth of sequencing technologies has greatly contributed to our understanding of human genetics. Yet, despite this growth, mainstream technologies have not been fully able to resolve the diploid nature of the human genome. Here we describe statistically aided, long-read haplotyping (SLRH), a rapid, accurate method that uses a statistical algorithm to take advantage of the partially phased information contained in long genomic fragments analyzed by short-read sequencing. For a human sample, as little as 30 Gbp of additional sequencing data are needed to phase genotypes identified by 50× coverage whole-genome sequencing. Using SLRH, we phase 99% of single-nucleotide variants in three human genomes into long haplotype blocks 0.2–1 Mbp in length. We apply our method to determine allele-specific methylation patterns in a human genome and identify hundreds of differentially methylated regions that were previously unknown. SLRH should facilitate population-scale haplotyping of human genomes.

Genome phasing

----- [A/T] ------ [C/G] ----- [G/T] ------

----- [A] ------ [G] ----- [G] ------

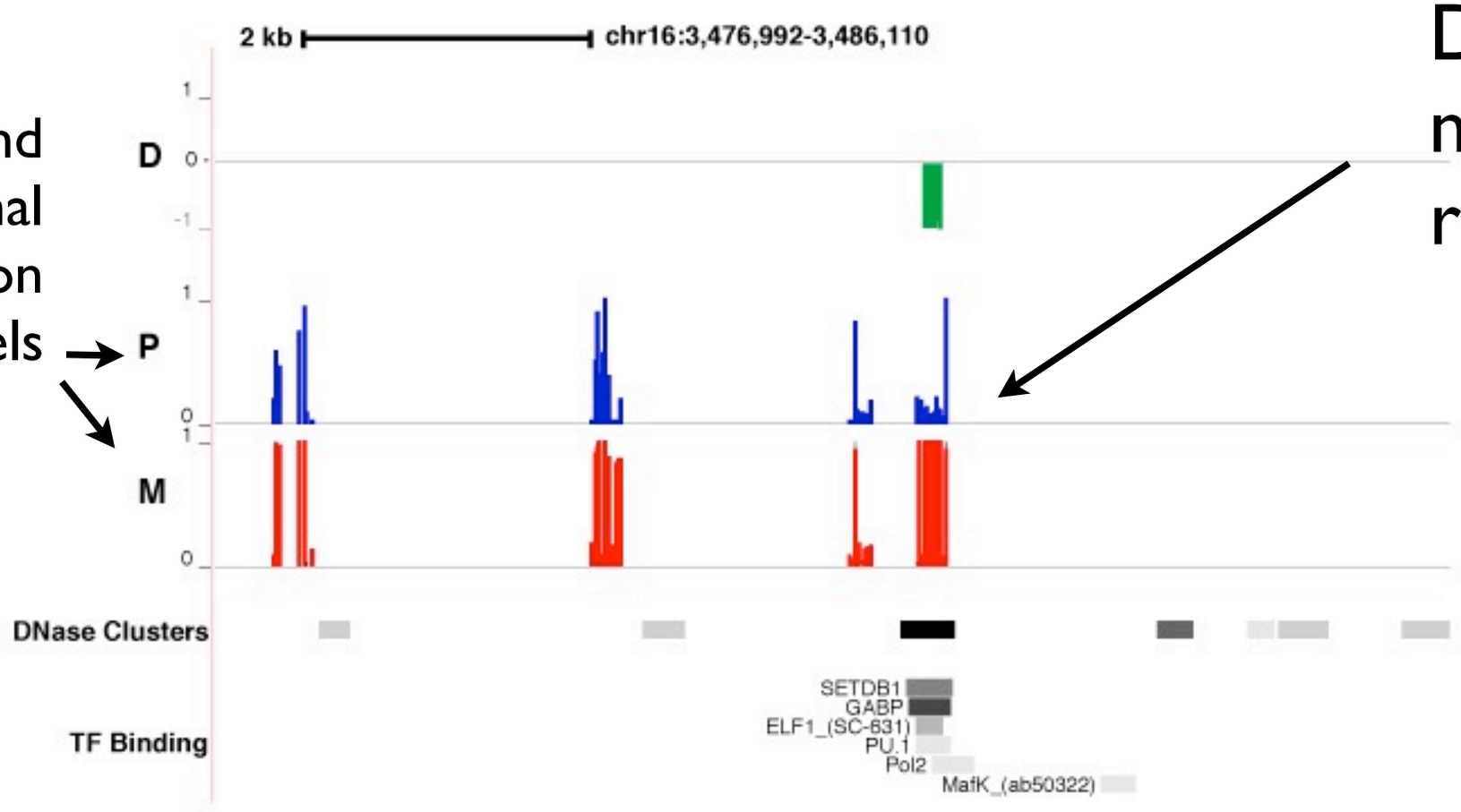
----- [T] ------ [C] ----- [T] ------

Fundamental aspect of human genetics
that is relevant in many applied problems

Scientific application

Allele-specific methylation

Paternal and maternal methylation levels



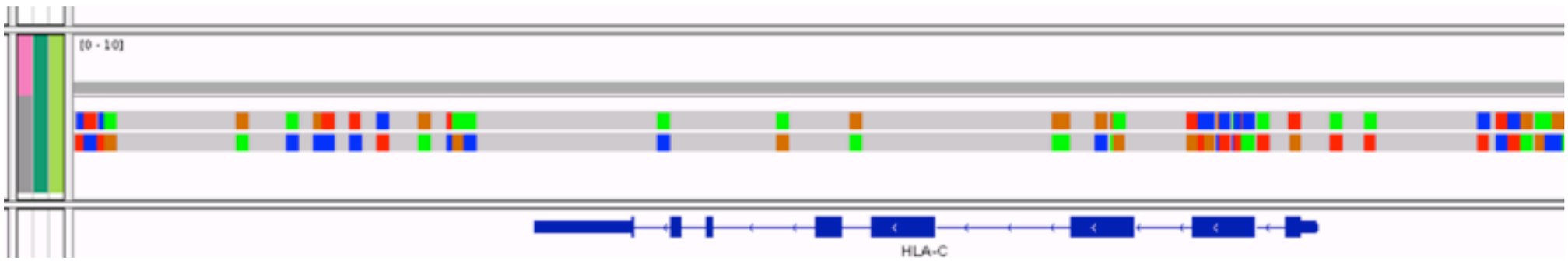
Differentially methylated region

TF factor binding site

Medical application

HLA typing

- Immune response during organ transplantation depends on compatibility between HLA genes



- These genes are highly heterozygous

General principle

unphased genome

----- [A/T] ------ [C/G] ----- [G/T] ------

----- [A] ------ [G] -----

sequence reads

------ [C] ----- [T] ------

----- [T] ------ [C] -----

phased result

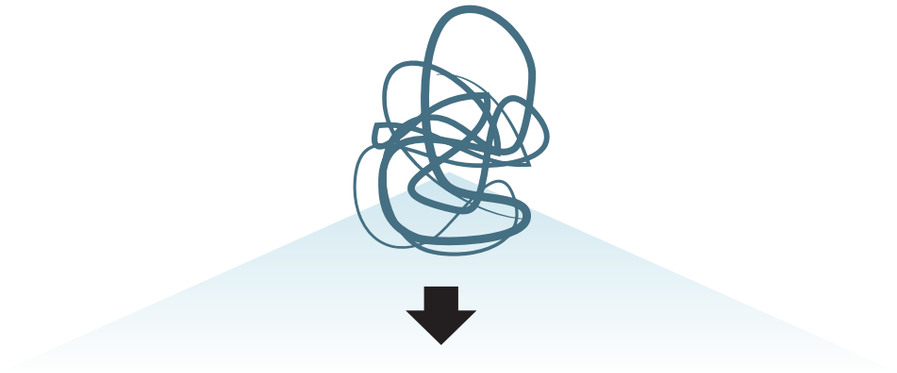
----- [A] ------ [G] ----- [G] ------

----- [T] ------ [C] ----- [T] ------

Long read sequencing

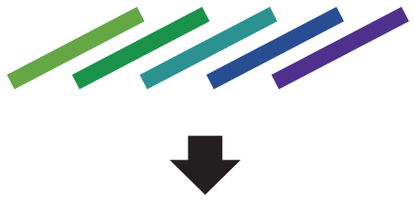
- Phasing is now becoming possible thanks to new synthetic long read technologies
- Examples: Moleculo, Long Fragment Reads (LFR), 10X Genomics
- Produce virtual multi-kb reads on regular sequencers

1.



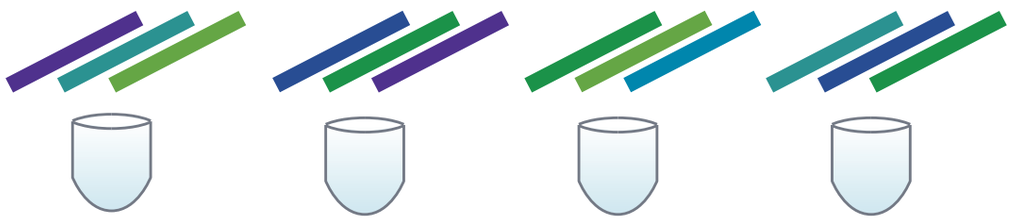
Moleculo starts with quality DNA

2.



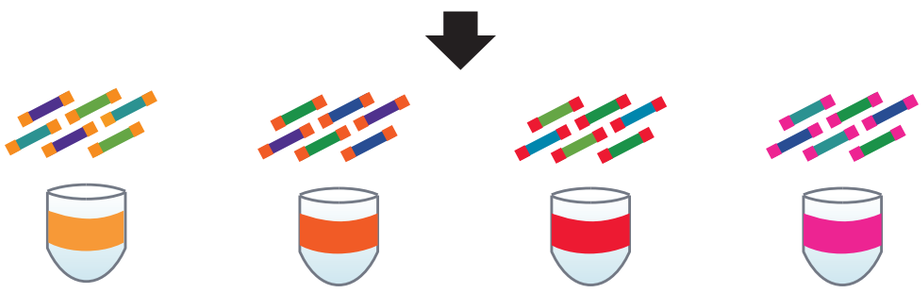
DNA is cut into 10 Kbp fragments

3.



The fragments are placed into wells

4.

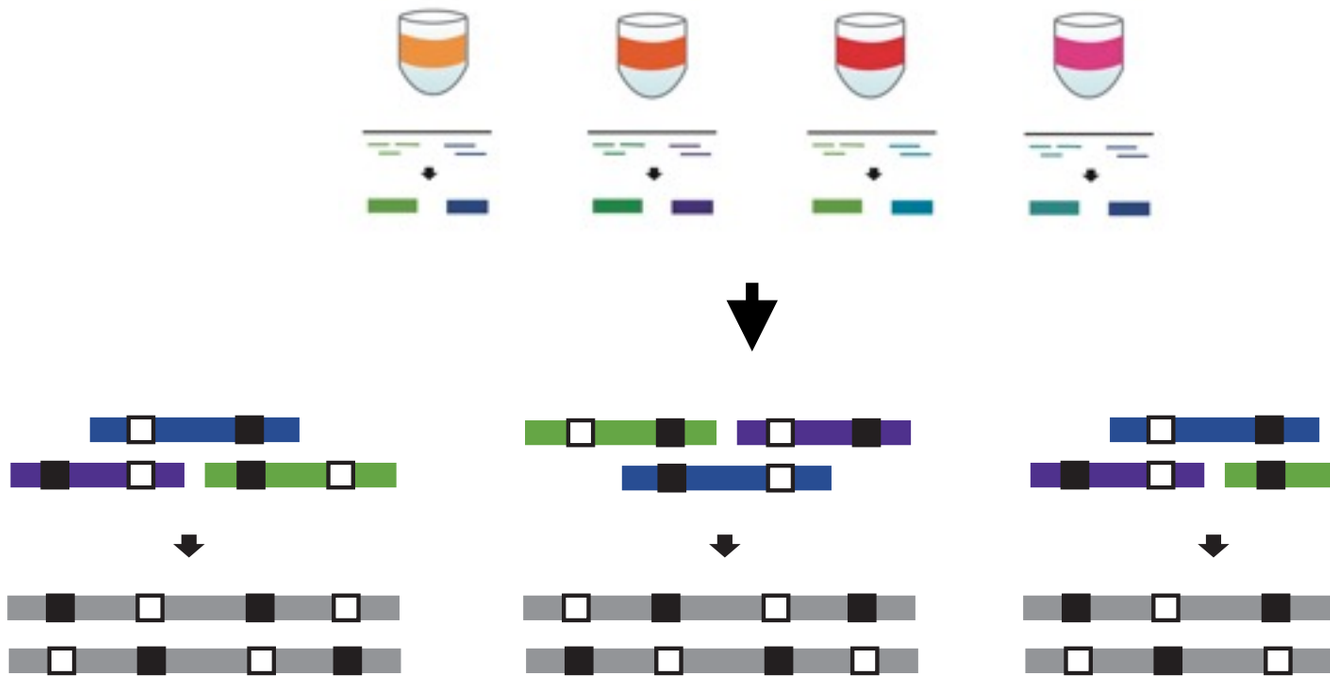


Wells are assigned a unique barcode

5.

The contents of each well are sequenced with short reads and reconstructed on a computer

Locally phased blocks



- Phasing as inference in a probabilistic model (ECCB14)
- 11% more accurate than RefHap
- Produces useful confidence scores

Shortcomings

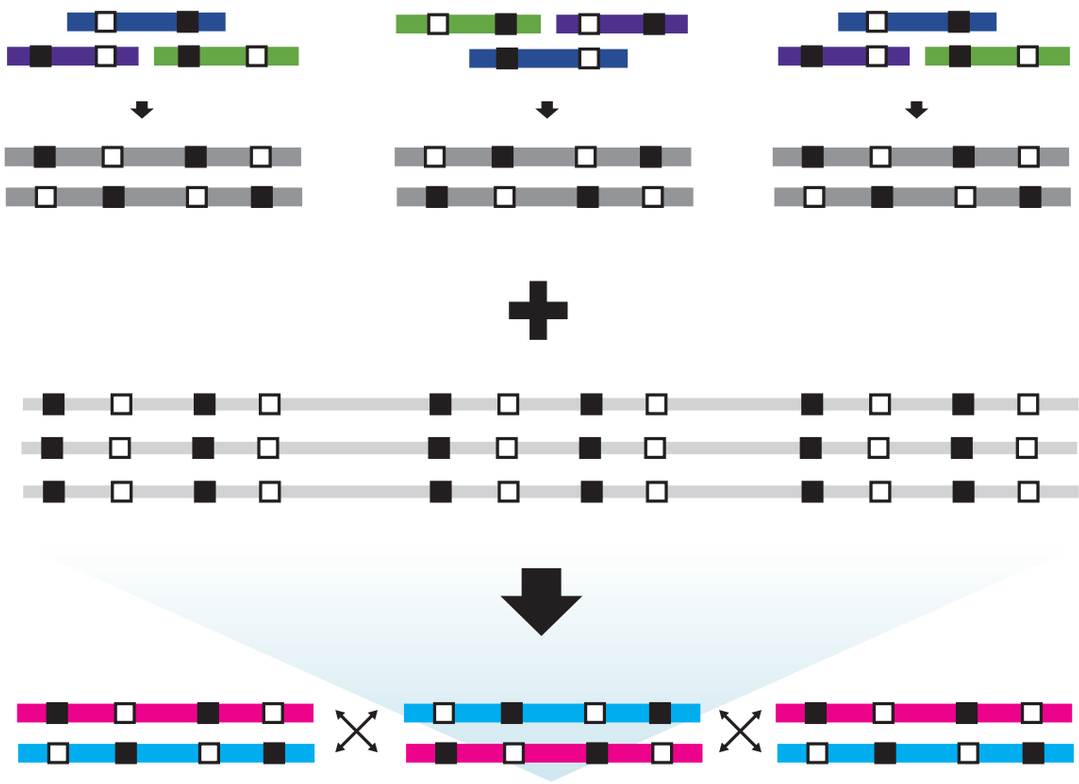
- Reads too short relative to other methods
- 10% of variants unphased due to sequencing bias

	Moleculo	*LFR*
N50	60 Kbp	**600 Kbp**
% phased	90%	**95%**

Idea:

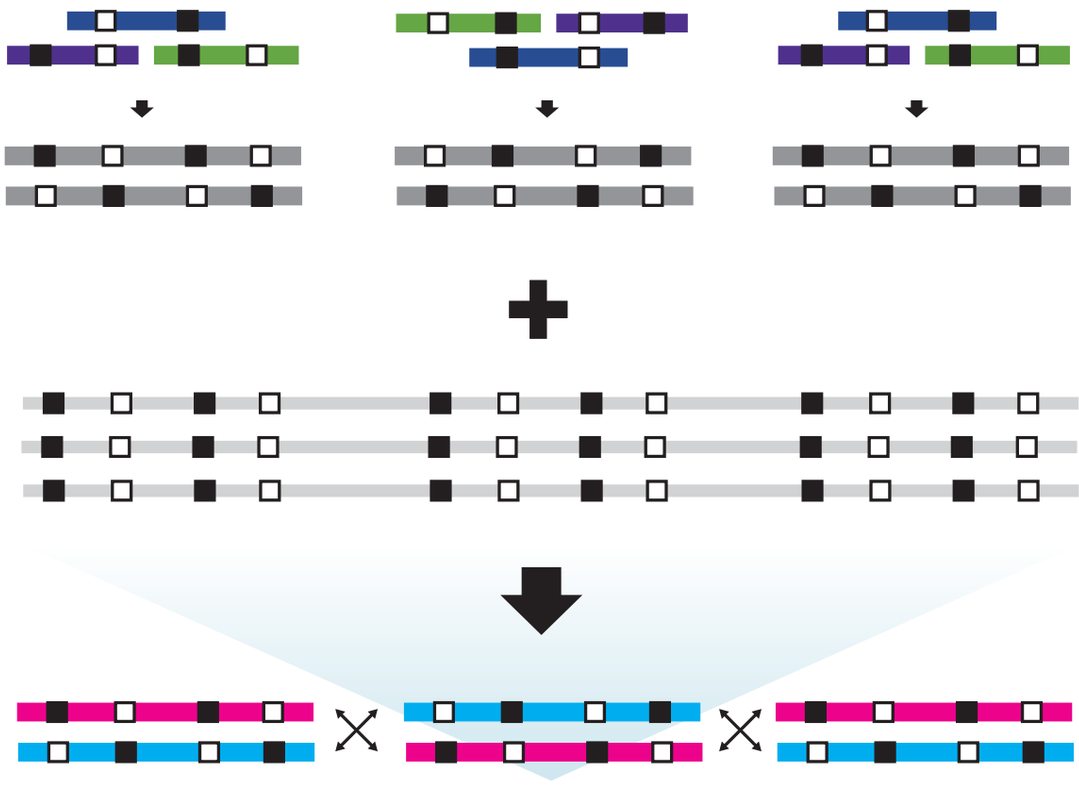
Use statistical phasing!

Prism Statistical Phaser



- Extends earlier methods to handle pre-phased blocks
- Prior information from blocks significantly improves accuracy
- Works best where molecular phasing fails
- Produces useful confidence scores

Prism Statistical Phaser



- Augments the HMM model of Li and Stephens (used in Impute2, Shape-IT, etc) with additional variables
- Determines scores using probabilistic inference in the model

Experiments

500 Kbp N50

< 1 error/Mbp

	Haplotype block N50 length (bp)	Phasing rate over SNVs	Switches per Mbp
NA12878 (two libraries)	563,801	99.00%	0.47
NA12891 (two libraries)	647,599	99.25%	0.68
NA12892 (two libraries)	531,804	98.84%	0.75

99% of SNVs phased

Comparison

	Kuleshov et al.	Kaper et al.	Peters et al.	Kitzman et al.
Frag. size	8 Kbp	10-20 Kbp	64-82 Kbp	37 Kbp
Sequencing	30-60 Gbp	203-409 Gbp	238-496 Gbp	110 Gbp
% phased	99%	97%	92-97%	94%
N50 (Kbp)	450-560	358	530-600	386
Long sw. acc.	99.91%	99.4%	2.2% of blocks	n/a
Short sw. acc.	99.90%	99.7%	99.99%	99.7%

This shows how clever algorithms can greatly improve sequencing technology

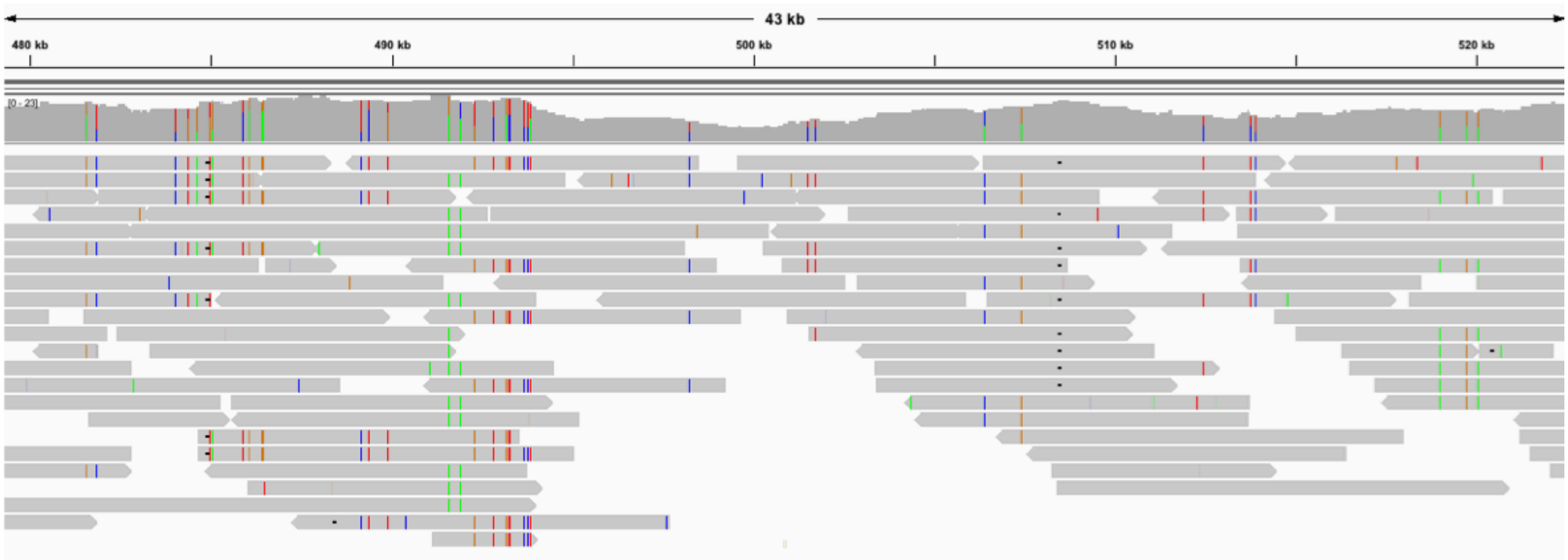
Metagenomics

- We used Moleculo to assemble the human gut microbiome, which led to:
  - Very long contigs
  - High resolution analysis of strains
- Enabled by new software package called Nanoscope

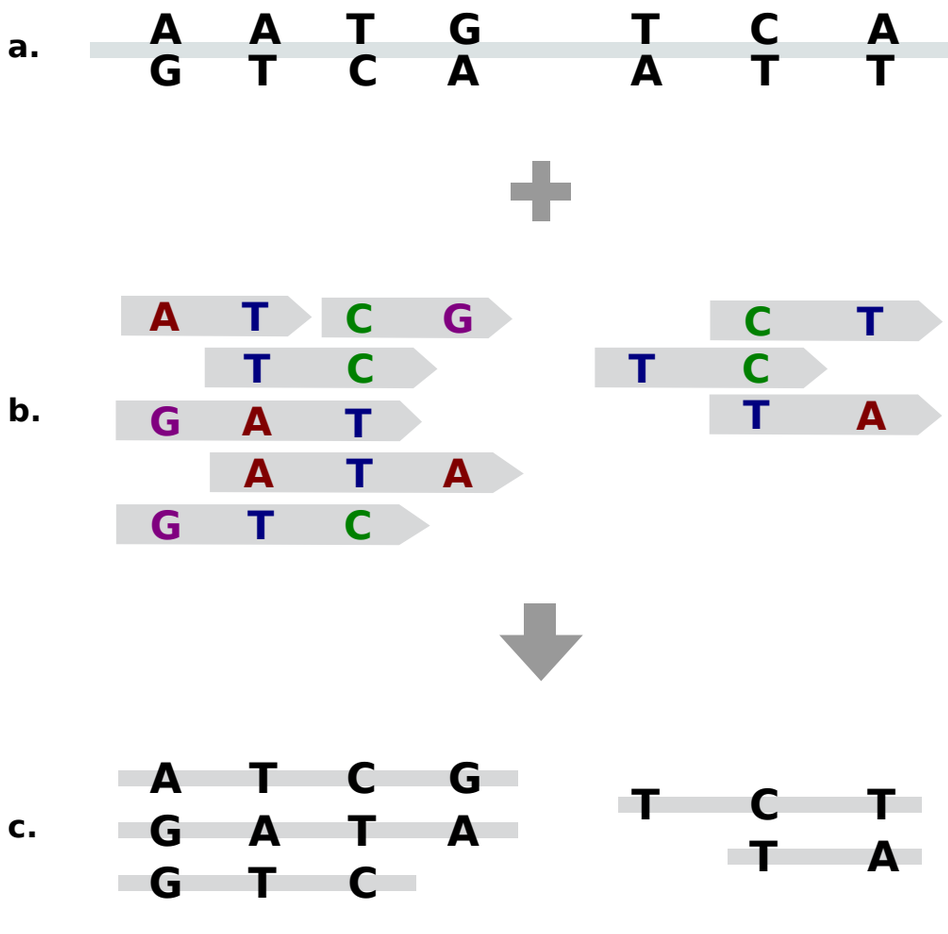
Assembly results

- 650 Mbp of sequence as 50 Kbp (N50) contigs (7x longer than with Illumina)
- Several megabase-long contigs, including a recently discovered species

Sub-strain identification



Sub-strain identification

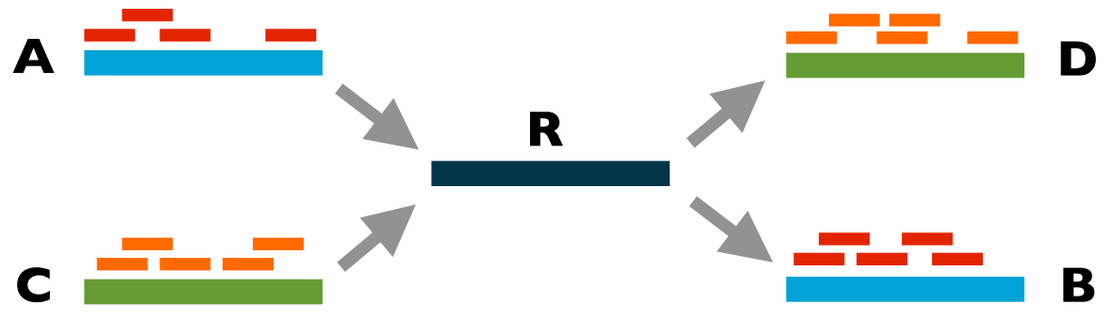


- Lens phasing algorithm reconstructs bacterial haplotypes.
- Over 200K variants
- Haplotype N50 length of 22 Kb
- Several long haplotypes of over 120 Kbp

De-novo Assembly

Two regions with repeat R covered by long reads

Repeat structure in the assembly graph



Resolving the repeat using raw short reads

Conclusion

- Synthetic long reads are a promising sequencing technology that can make progress on important genomics problems
- This technology requires developing novel computational methods, which opens a new research direction

Acknowledgements

- **Snyder Lab:** Mike Snyder, Dan Xie, Chao Jiang, Wenyu Zhou

- **Batzoglou Lab:** Serafim Batzoglou, Alex Bishara, Yuling Liu

- **Moleculo Team:** Dmitry Pushkarev, Michael Kertesz, Tim Blauwkamp

- **Funding Agencies**
  - NIH Training Grant
  - NSERC Canada Graduate Fellowship
  - ISCB for travel support

Thank you!