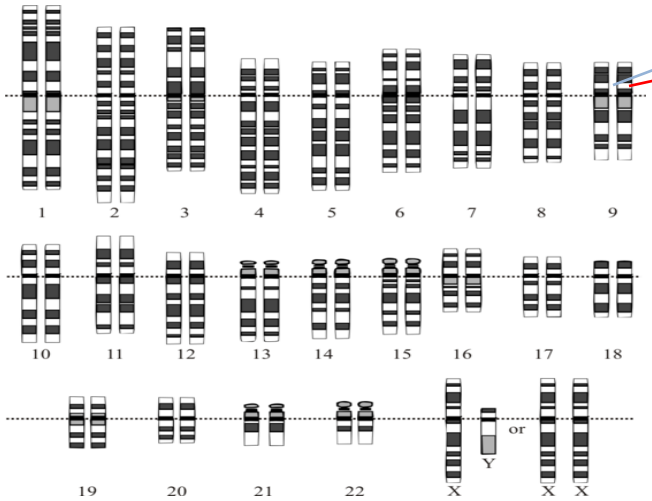
Integrating read-based and population-based phasing for dense and accurate haplotyping of individual genomes

Vikas Bansal
University of California San Diego

ISMB 2019

Humans are diploid



maternal ACGTGCA.....GTACAC.....AGTAAGACTA.....GTCCGTA....

paternal ACGGGCA.....GTTCAC.....AGTAACACTA.....GTCTGTA....

sequencing

genotypes T/G A/T G/G C/G C/T

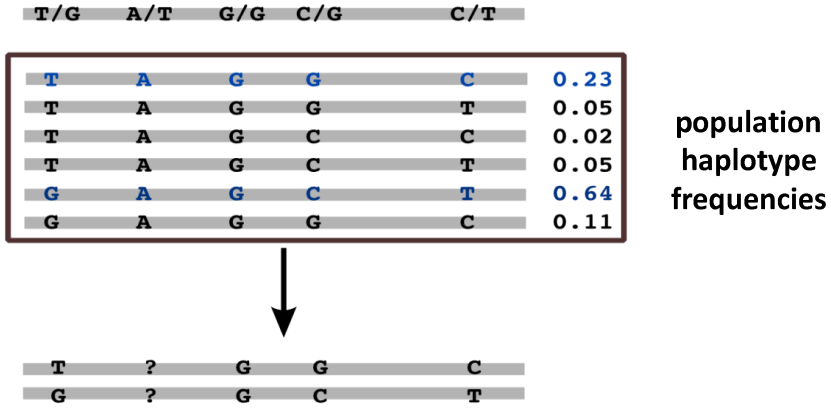
haplotype phasing

missense mutation T/G A G G/C

stop-gain mutation C/T

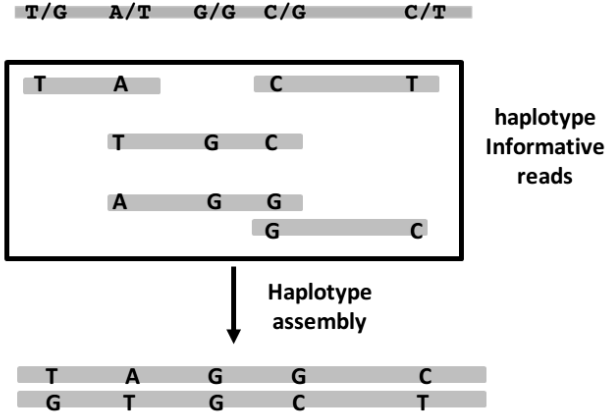
Two approaches for haplotype phasing

1. Population-based (statistical phasing)



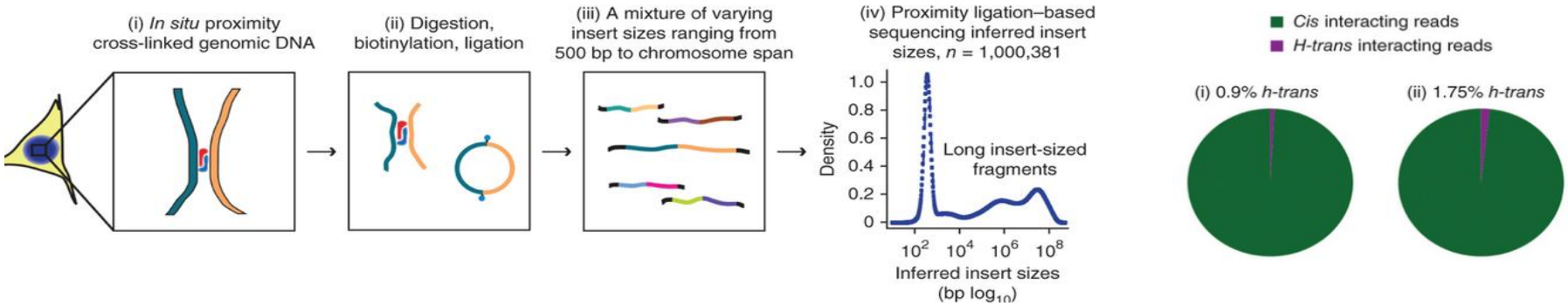
- Uses linkage disequilibrium patterns to infer most likely haplotypes
- Many statistical methods (SHAPEIT, Beagle)
- Limited ability to phase rare variants

2. Read-based (haplotype assembly)



- Requires sequencing with long reads or long-range information
- Equally accurate for common and rare variants

Illumina sequencing using Hi-C enables whole-genome haplotype phasing



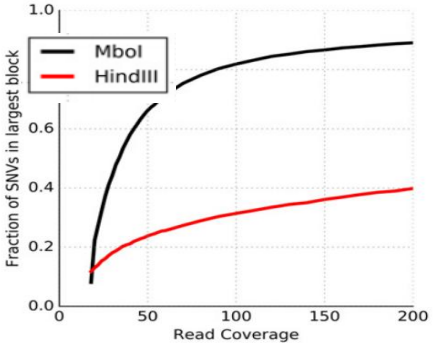
- 98-99% of intra-chromosomal read-pairs are ‘cis’
- haplotypes for human genome (NA12878) assembled from 18x whole-genome Hi-C using HapCUT

1. **Contiguity**: chromosome-spanning haplotype block
2. **Completeness**: 18-22% variants phased per chromosome
3. **Accuracy**: switch error rate of 2-3%

S. Selvaraj, J. Dixon, V. Bansal, B. Ren, Nature Biotech. 2013

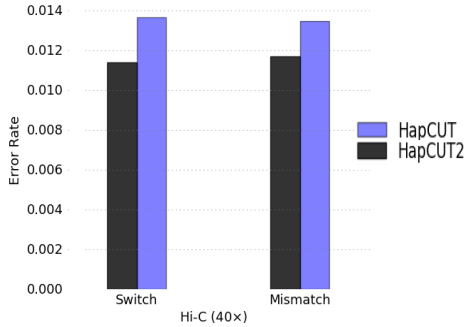
Improving Hi-C based haplotype phasing

- Using the MboI (4 bp cut-site) restriction enzyme for Hi-C library preparation



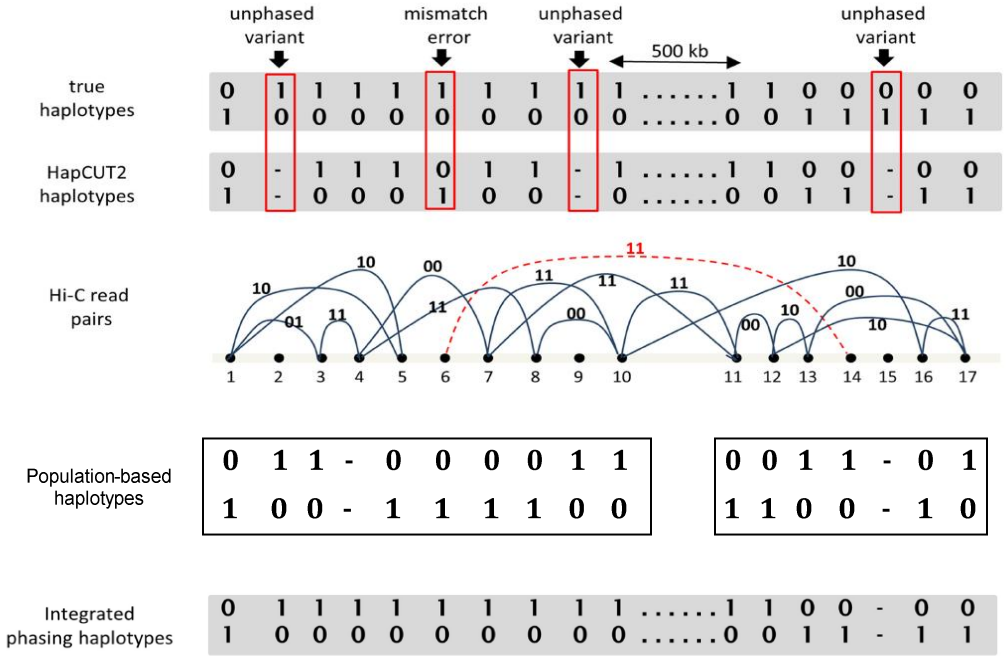
high-depth Hi-C data (Rao et al. 2014)

- Modeling trans-errors in Hi-C using a likelihood-based model for haplotyping (HapCUT2)

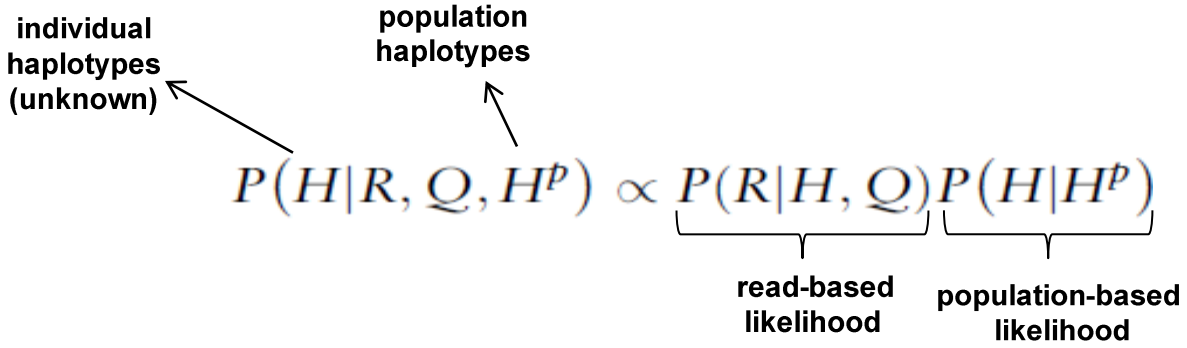


Edge, Bafna and Bansal, Genome Research 2017

Population-based phase information can improve Hi-C phasing

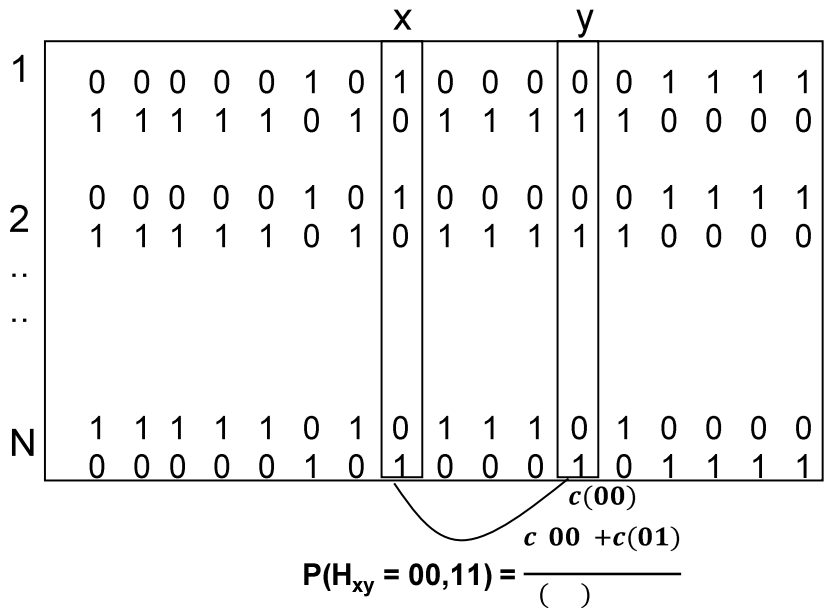


Joint likelihood model for phasing



- SHAPEIT2 HMM cannot model long-range Hi-C information (Delaneau et al. 2013)
- HapCUT2 can optimize read-based likelihood for Hi-C reads
- **Approach:** approximate population-based likelihood using second-order probability distributions (estimated using SHAPEIT2) and optimize using HapCUT2

Approximating population-based likelihood using second-order probability distributions

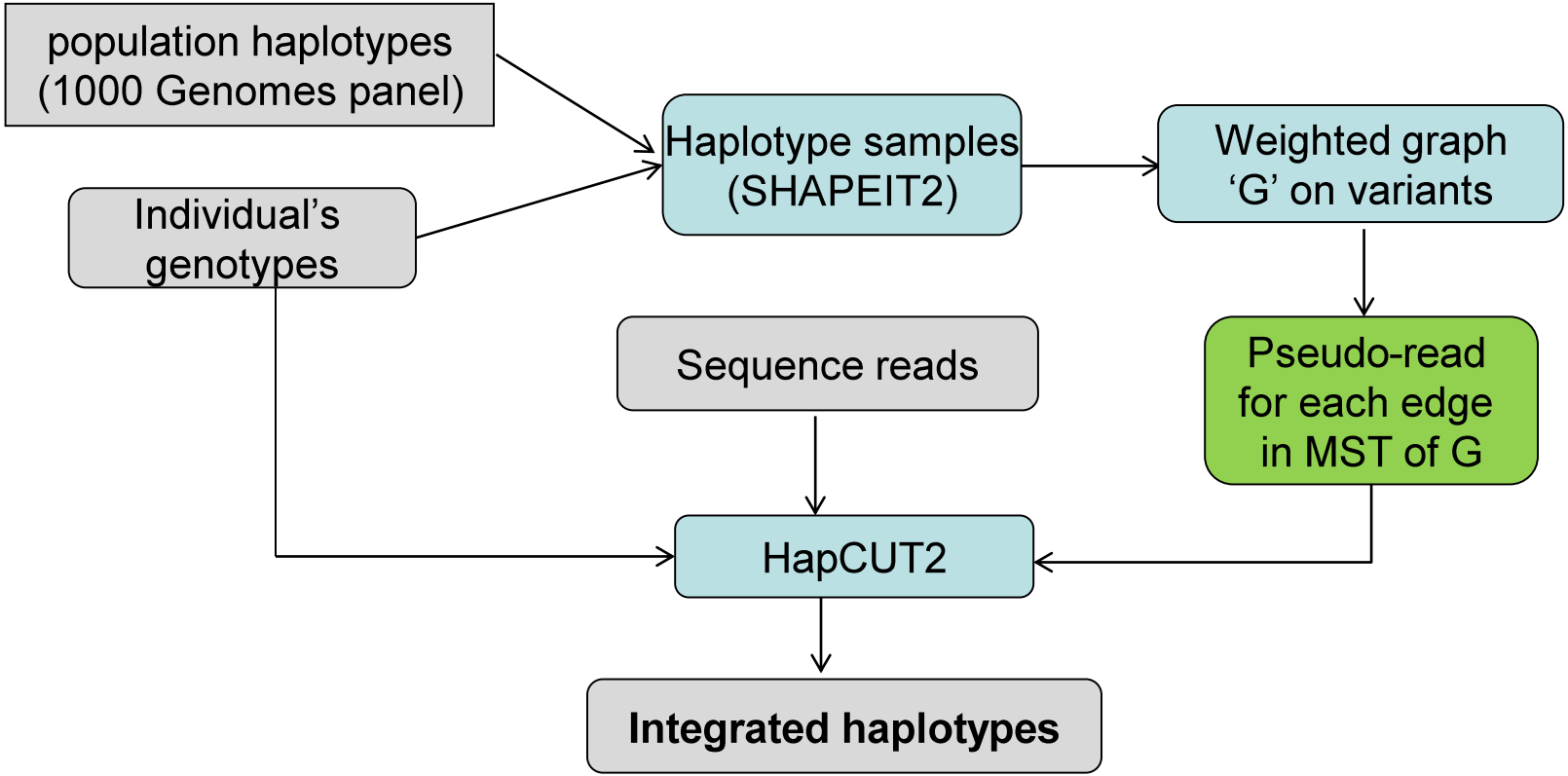


$$P(H|H^p) \approx P(H_{x_1}) \prod_{i=1}^{n-1} P(H_{x_{i+1}}|H_{x_i}),$$

- $x_1, x_2, \ldots, x_n$ is a permutation of the variants
- Select permutation that has minimum KL distance to full distribution
- Chow-Liu (1968): maximum spanning tree of graph with edge weights equal to mutual information gives optimal permutation

Each term can be encoded as a pseudo-read 'r' s.t. $P(H_x|H_y) = P(r|H_{xy}, q_x, q_y)$

Integrated phasing method



Results: Hi-C data

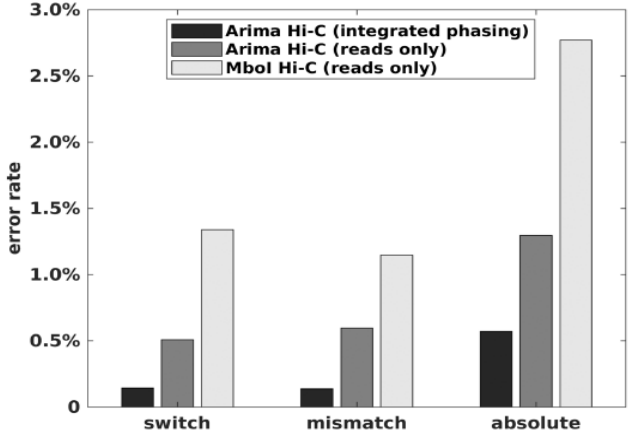
- data for NA19240 from 1KG project (30x coverage)
- Aligned reads to reference genome using BWA-mem
- Phasing accuracy measured using 1KG trio haplotypes

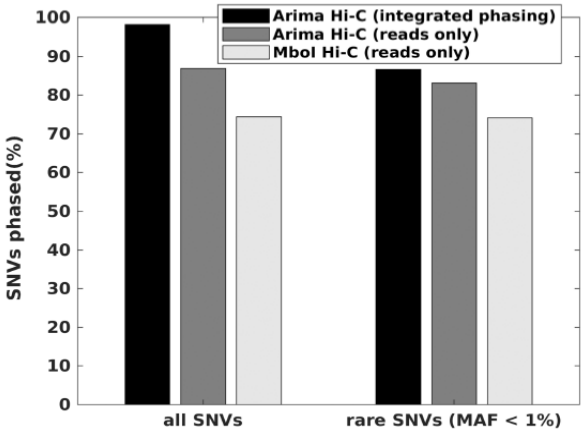
Method	SNVs phased (%)	Absolute error rate (%)	Switch error rate (%)	Mismatch rate (%)	Run time
Reads only	51.30	0.49	0.20	0.365	02:43
Integrated phasing	97.32	0.31	0.034	0.266	08:57
SHAPEIT2	98.67	42.1	0.27	0.76	04:57

Results: Hi-C data

- Data for NA12878 generated using two different Hi-C protocols
  - MboI RE
  - multi-enzyme chemistry (Arima Genomics)

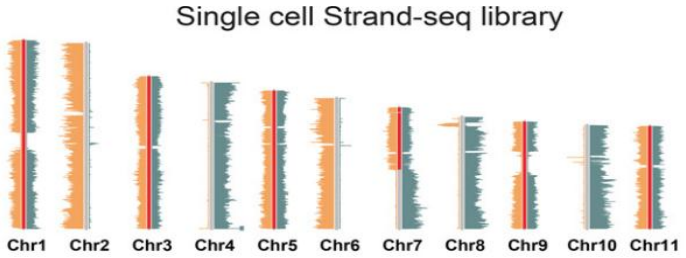
- Aligned reads to reference genome using BWA-mem
- Phasing accuracy measured using Platinum Genomes haplotypes





Results: Strand-seq data

- Single-cell sequencing method that provides sparse haplotype information across entire chromosomes
- Strand-seq enables phasing of 70-80% of variants for human genomes (Porubsky et al. 2016)



(Image from Porubsky et al. 2016)

NA12878 data for 133 cells

Method	SNVs phased (%)	Switch error rate (%)	Mismatch error rate (%)
Reads only	71.38	0.091	0.268
Integrated phasing	94.56	0.0364	0.134

Conclusions

- Novel likelihood based method that can integrate phase information from sequence reads and population haplotype panel
- Significantly improves completeness and accuracy of phasing using two different sparse sequencing methods
- Multi-enzyme Hi-C sequencing (30-40x WGS) enables highly accurate and dense whole-genome haplotyping

Acknowledgements

Siddarth Selvaraj (Arima Genomics)

Peter Edge (UCSD)

Funding:

Department of Pediatrics, UCSD

NHGRI (NIH)