

ASSEMBLY-BASED STRUCTURAL VARIATION AND HAPLOTYPES FROM TARGETED SUB-MEGABASE DNA MOLECULES



AGBT 2018
2/13/2018



Stanford
MEDICINE

Ji Research Group
In the Division of Oncology

Contents

- Approach
- Phasing and assembly of
 - *BRCA1* – 0.2 Megabase
 - MHC locus – 4 Megabases
 - 38 structural variants – 0.1 Megabase each



Overview of the approach



Why target Megabase DNA regions?

- Assembling Mb contiguous genomic regions of high sequence complexity
(e.g. structural variations)
- Identifying structural variants present at low allelic fractions
(e.g. genome mixtures such as cancers)



Challenges of sequencing sub-Mb targets

- No efficient enrichment method
- No preservation of intact HMW DNA molecule
- High DNA amounts required for diploid assembly

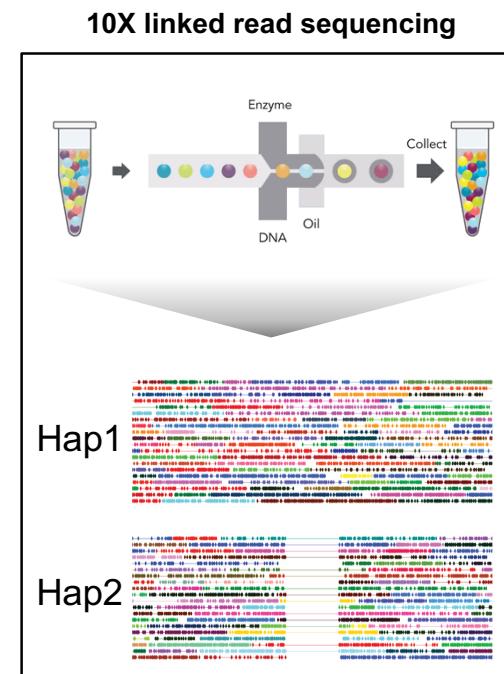
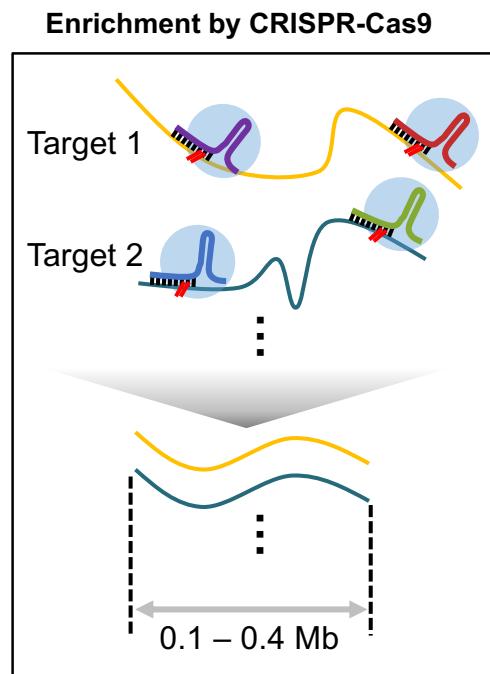


Solutions for sequencing sub-Mb targets

- No efficient enrichment method
 - Target enrichment by *in vitro* CRISPR-Cas9
- No preservation of intact HMW DNA molecule
 - Sage HLS high molecular weight system
- High DNA amount requirement for diploid assembly
 - 10X linked short read sequencing with 1 ng



Steps for targeting Mb region



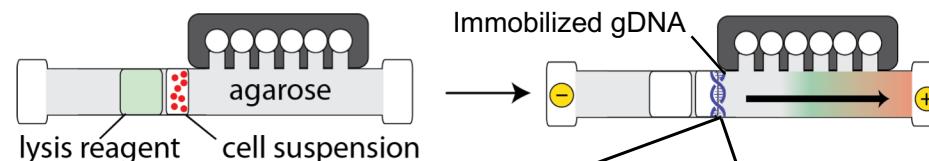
Jiang, W. et al., *Nat. Commun.* (2015)
Shin, G. et al., *Nat. Commun.* (2017)

Zheng, G.X.Y. et al., *Nat. Biotechnol.* (2016)
Bell, J.M. et al., *Nucleic Acids Res.* (2017)
Xia, L.C. et al., *Nucleic Acids Res.* (2017)
Greer, S.U. et al., *Genome Med.* (2017)

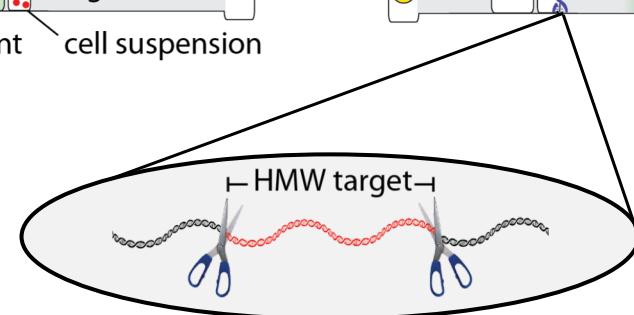


Automated target enrichment with Sage HLS

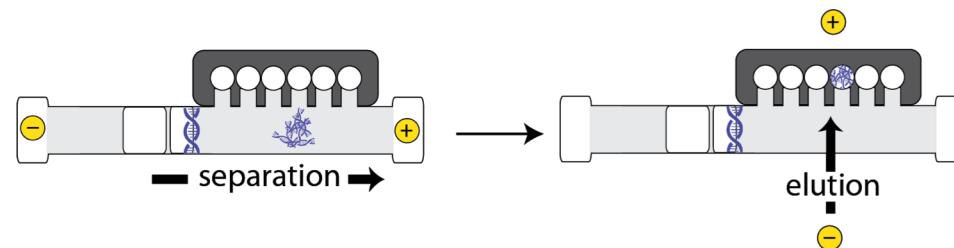
Step 1:
DNA Extraction
from intact cells



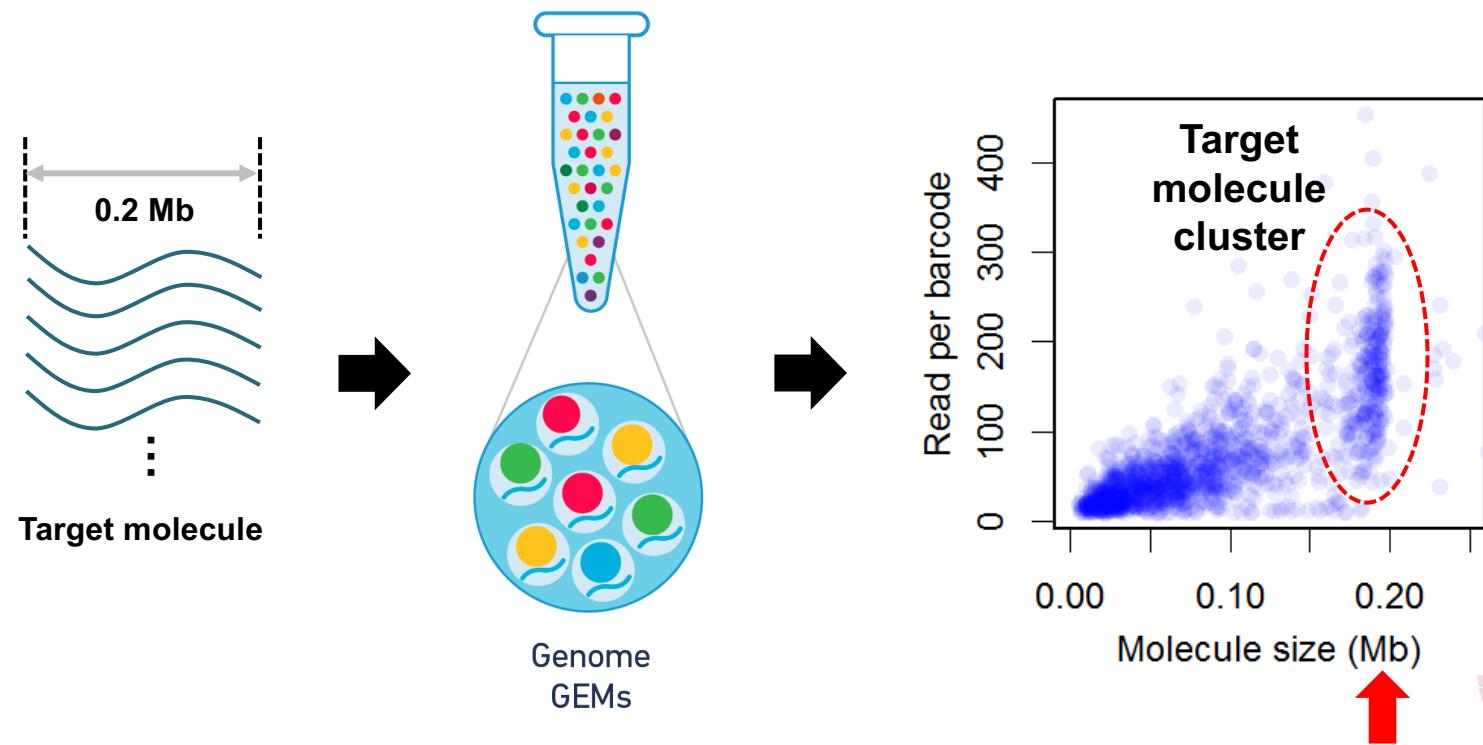
Step 2:
CRISPR-Cas9
digestion



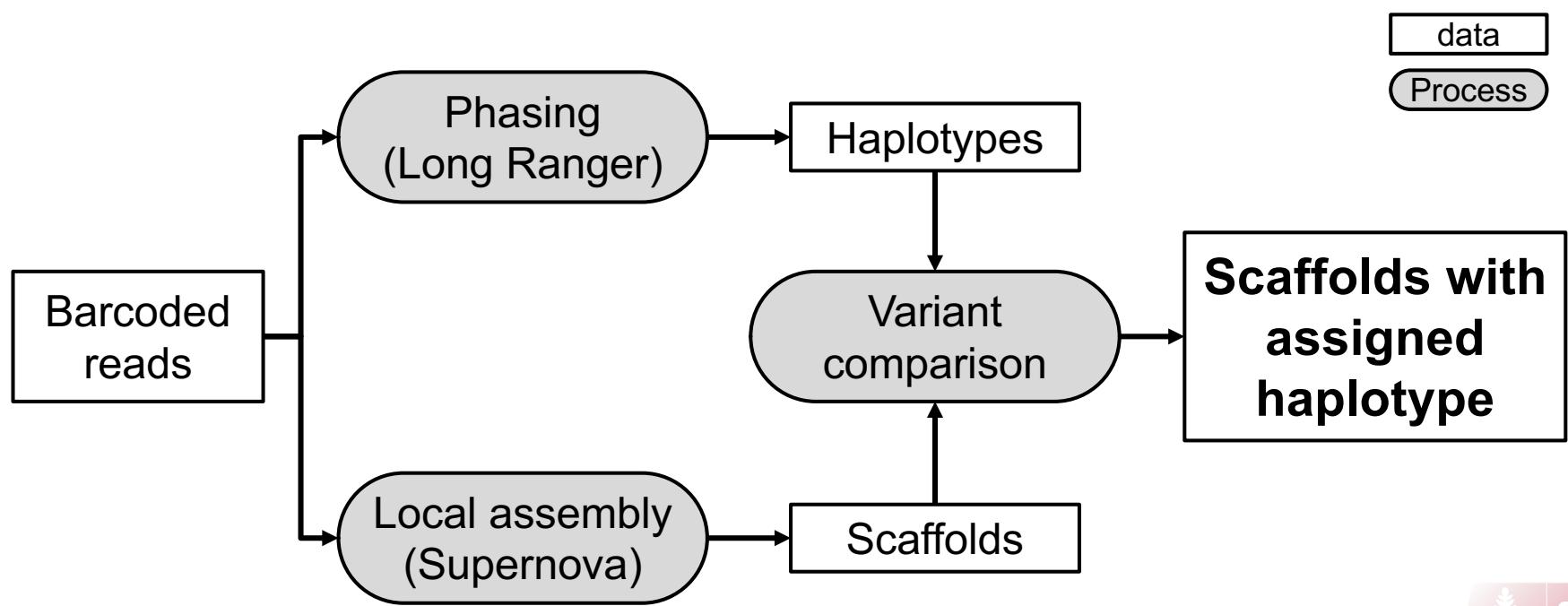
Step 3:
Size Selection



Barcoding HMW target DNA in droplet partitions



Analysis overview



Assembly of *BRCA1* locus

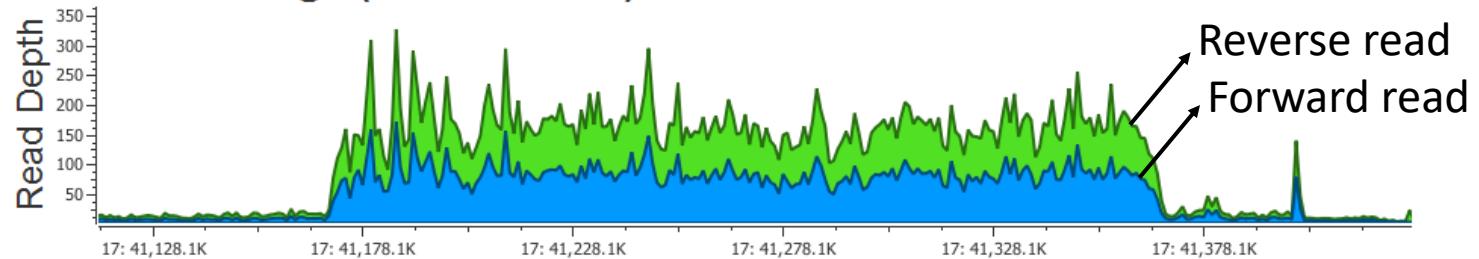


CRISPR-Cas9 enrichment of *BRCA1* molecule

BRCA1 target locus (0.2 Mb)



Binned coverage (bin size: 1kb)

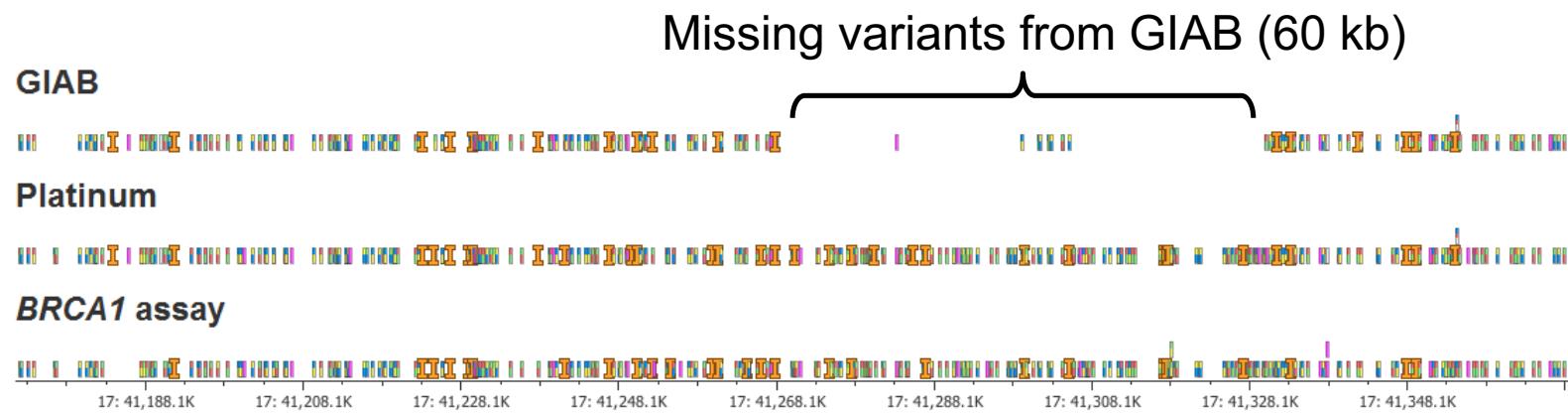


Overall WGS coverage	Target coverage	Fold enrichment
4.4	156	36



BRCA1 locus haplotype is accurate

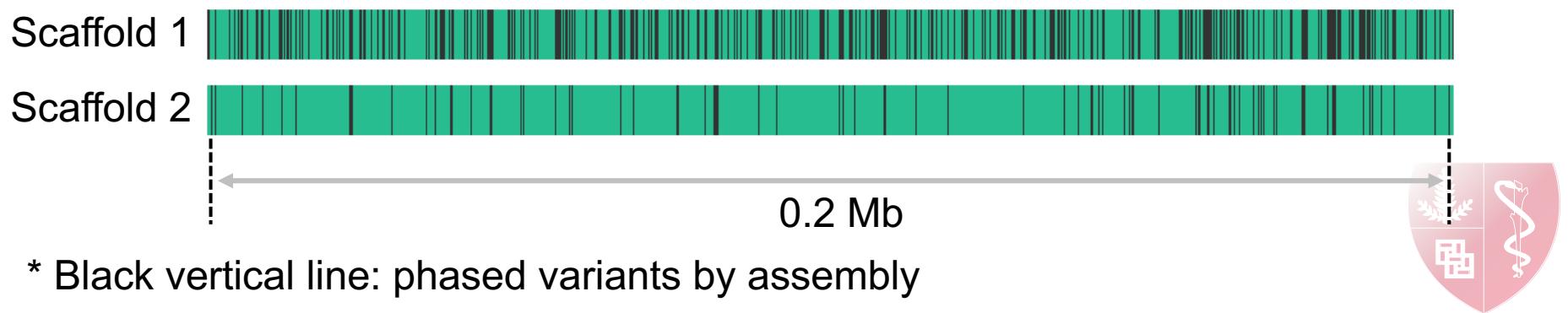
	Fraction concordance (match / common)	
	vs. GIAB	vs. Platinum
<i>BRCA1</i> assay	99.5% (212/213)	99.7% (341/342)



Local assembly produced two haploid copies

Contig N50 (kb)	142.91
Scaffold N50 (kb)	190.01

Alignment of scaffolds to the reference

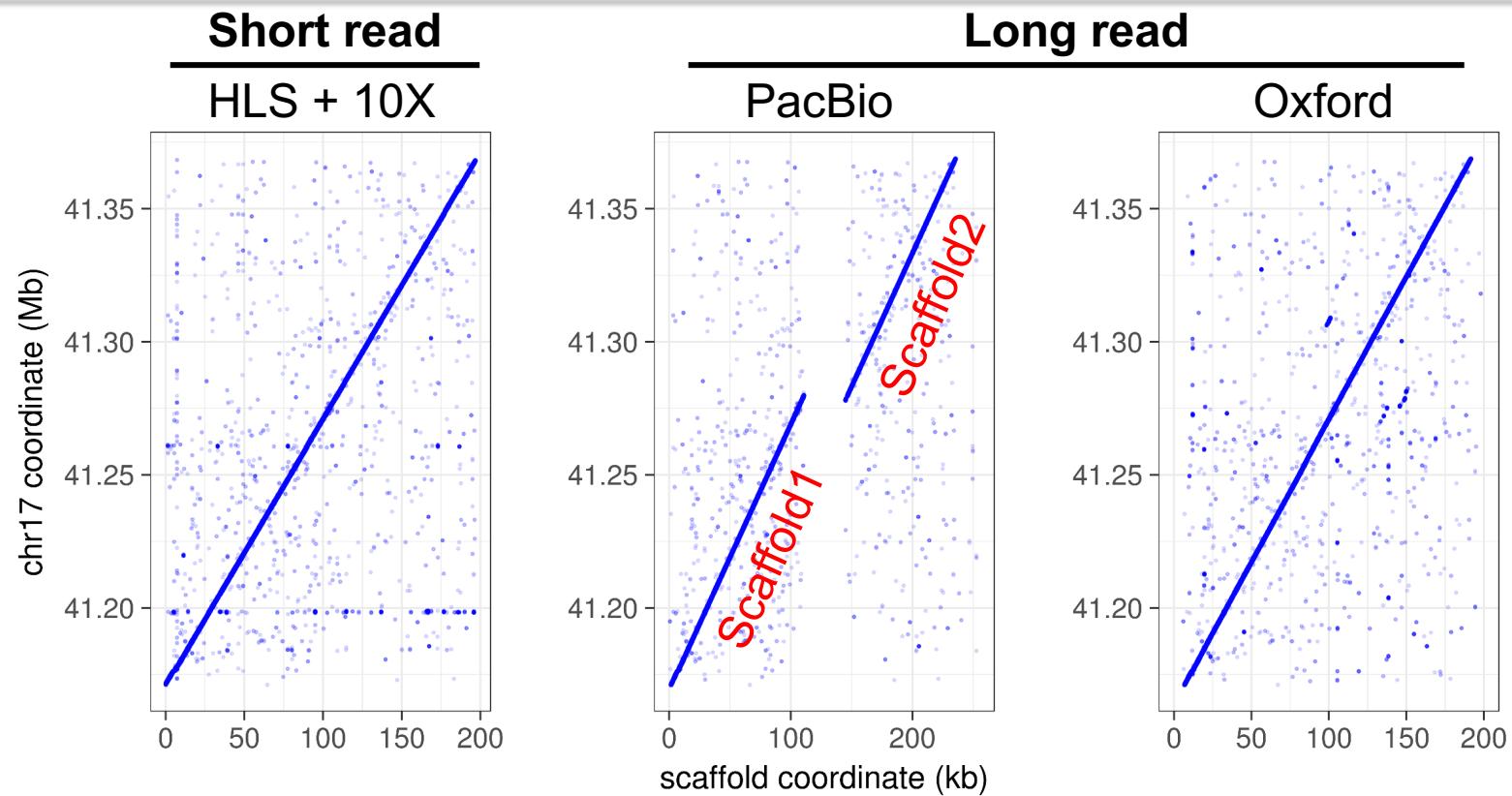


Consistency between assembly- and alignment-based haplotypes

		Fraction concordance (match / common)	
		Alignment Haplotype 1	Alignment Haplotype 2
Assembly Haplotype 1	0%	100%	(0/267) (267/267)
	100%	0%	(267/267) (0/267)



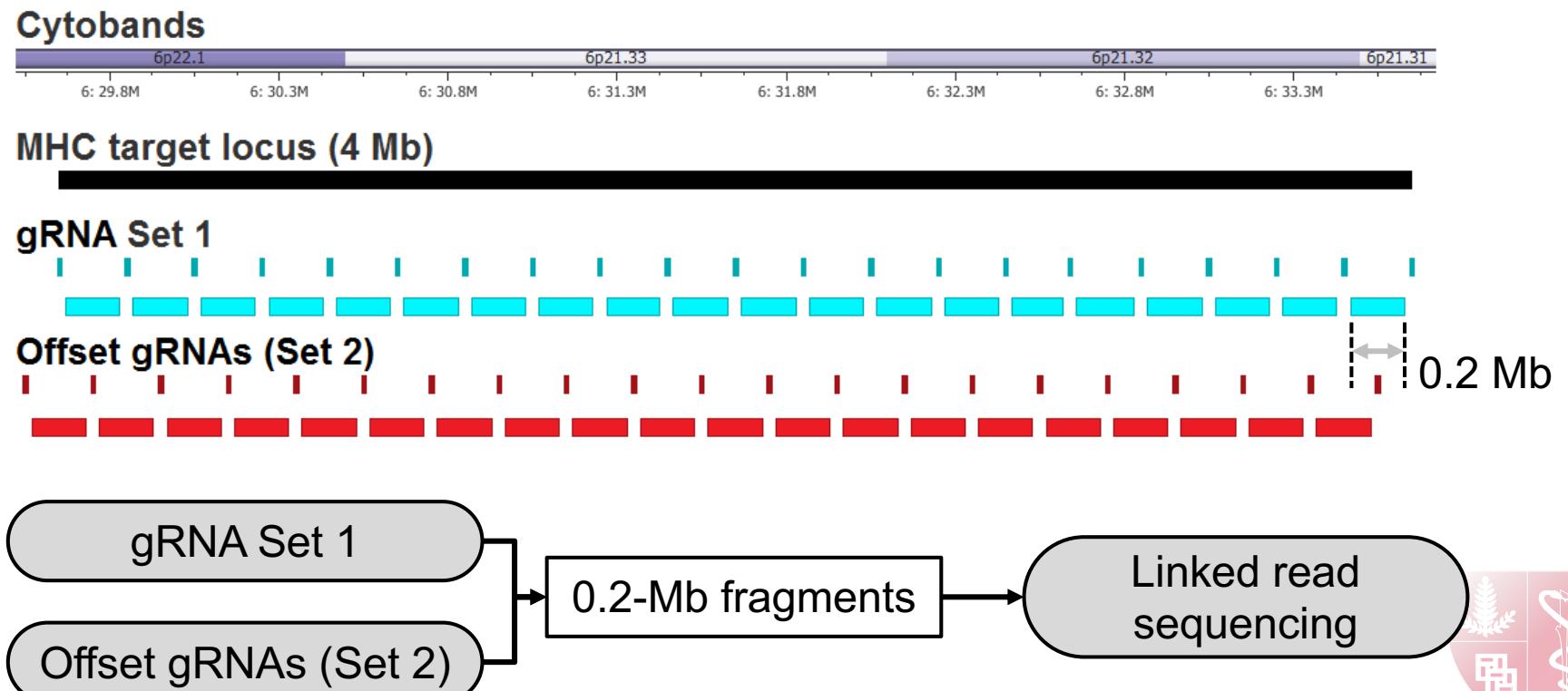
Comparison with other long read assemblies



Assembly of 4-Mb MHC locus



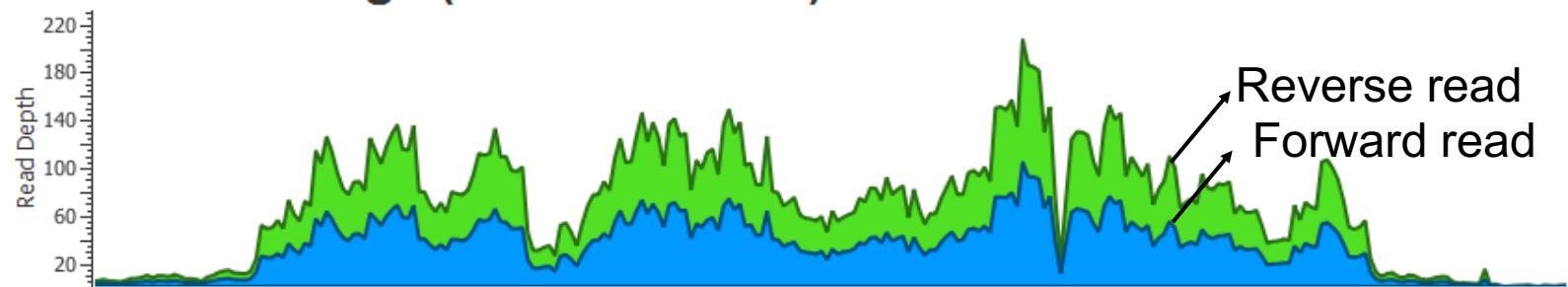
Strategy for targeting 4-Mb MHC region



CRISPR-Cas9 enrichment of MHC molecule

MHC target locus (4 Mb)

Binned coverage (bin size: 20 kb)



Overall WGS coverage	Target coverage	Fold enrichment
3.2	80	25

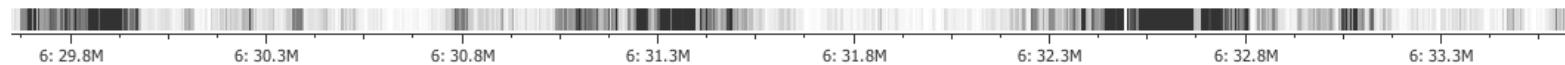


One phase block for the 4 Mb locus

10X WGS phase blocks



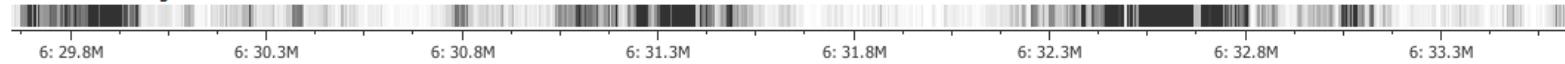
10X WGS variants



MHC assay phase block



MHC assay variants



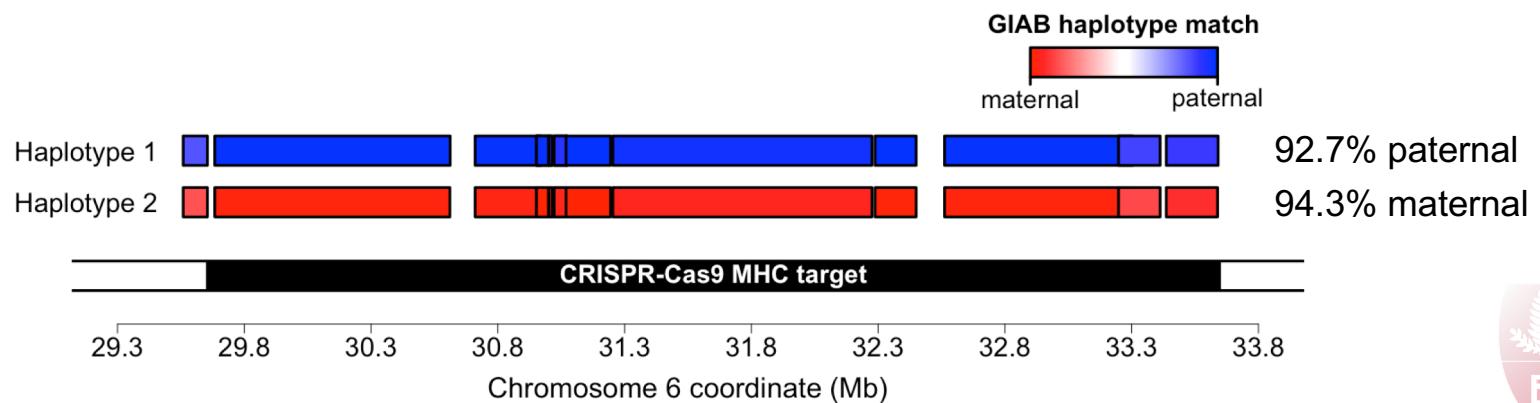
99.2% match to the Platinum genome

Based on alignment



Haplotype-assigned assembly of MHC

Scaffold N50 (kb)	882.4
Assembly size (Mb)	3.86
Number scaffold >10kb	30



HLA genotypes determined by assembly

Gene	Haplotype 1 (Closest alleles)	Haplotype 2 (Closest alleles)	Relative distance (Mb)
HLA-A	A*01:01:01:01	A*11:01:01:01	0
HLA-C	C*07:01:01:01	C*01:02:01:01	1.3
HLA-B	B*08:01:07 + 9 alleles	B*56:01:01:01 + 2 alleles	1.4
HLA-DQA1	DQA1*05:01:01:02	DQA1*01:01:01:01	2.7
HLA-DQB1	DQB1*02:01:01	DQB1*05:01:01:03	2.7
HLA-DPB1	DPB1*04:01:01:01	DPB1*14:01:01:01	2.7

- ✓ Total of 30 HLA genes were genotyped and phased by assembly
- ✓ All exons + introns were used

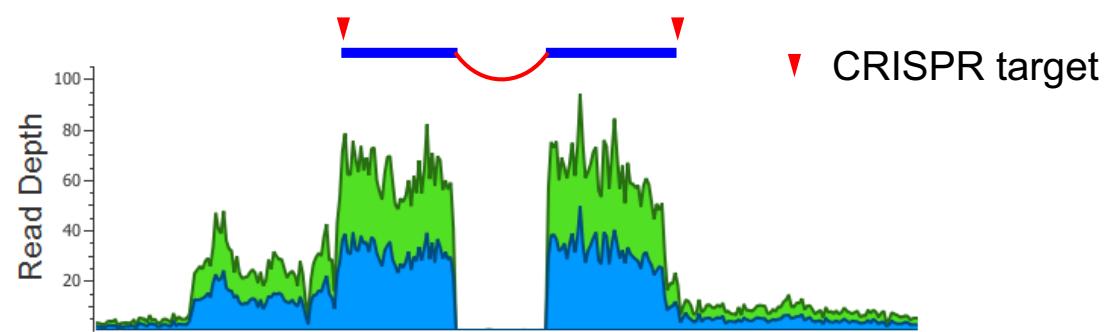


Assembly of 38 candidate SV regions

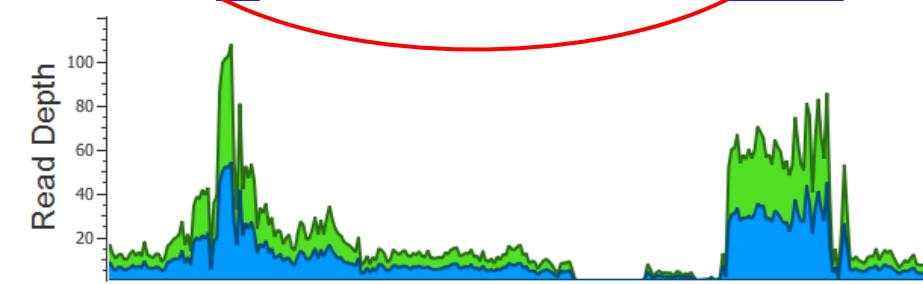


Capture different SV classes

29 Deletions

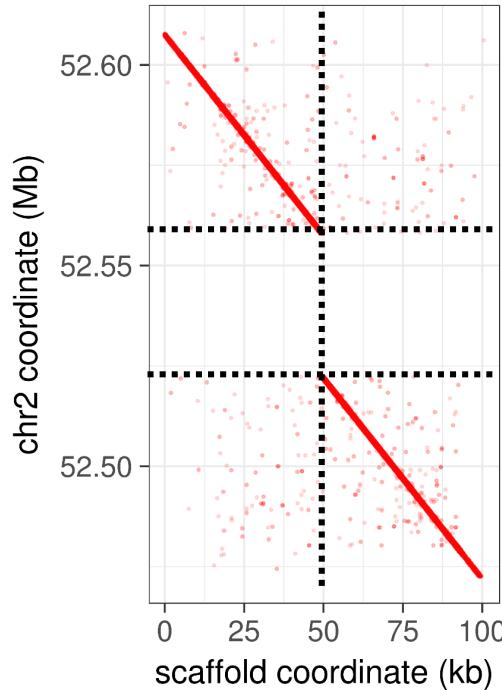


9 Inversion

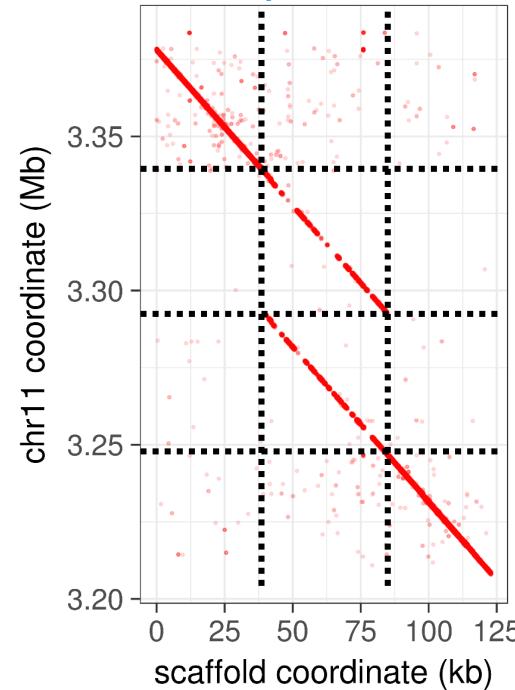


Assembly examples

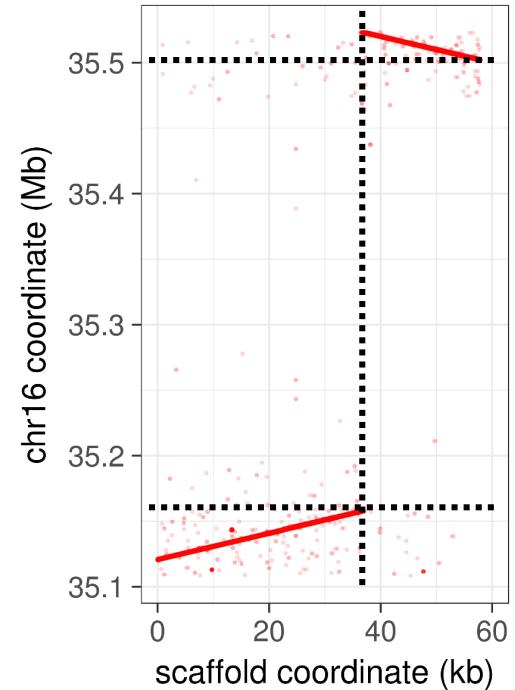
Deletion



Deletion in tandem duplication



Inversion



Conclusion & Highlights

- *In vitro* CRISPR-Cas9 enrichment of targeted Mb molecules
- Complete phasing of target molecule by short read-based assembly
- Assembly of extremely heterogeneous region (MHC)
- Assembly of structural variation



Funding

Funding – Ji Research Group

- NIH/NHGRI R01HG006137
- NIH/NHGRI P01HG00205 to Ji Lab

Funding – Sage Science

- NHGRI Award Number R44HG008720



Contributors

- **Ji Research Group
Stanford University**
 - Stephanie Greer
 - HoJoon Lee
 - Anuja Sathe
 - Billy Lau
 - Charlie Xia
 - Matt Kubit
 - Sue Grimes
 - John Bell
- **Sage Science, Inc**
 - Chris Boles
 - Jun Zhou
 - Todd Barbera

Hanlee Ji, Principle Investigator
hanleeji@stanford.edu

