A fully phased accurate assembly of an individual human genome

Tobias Marschall

Heinrich Heine University Düsseldorf, Germany
twitter: @tobiasmarschal

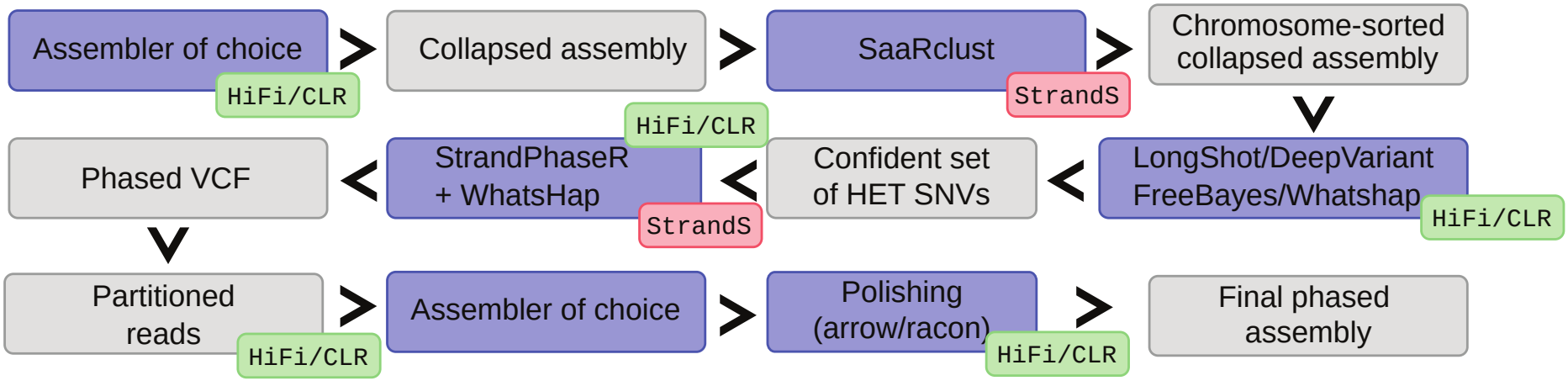
January 15, 2020
SMRT Informatics Developers Conference @ PAG

Preprint/ Acknowledgements

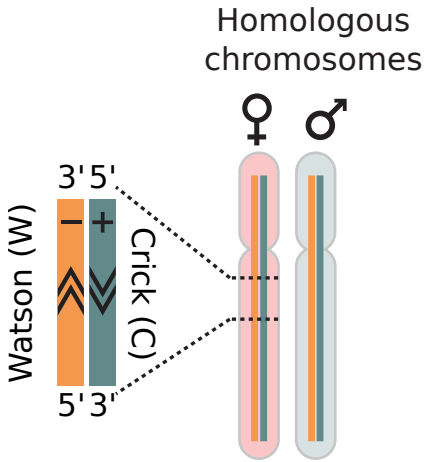
A fully phased accurate assembly of an individual human genome

David Porubsky[1*], Peter Ebert[2*], Peter A. Audano[1], Mitchell R. Vollger[1], William T. Harvey[1], Katherine M. Munson[1], Melanie Sorensen[1], Arvis Sulovari[1], Marina Haukness[3], Maryam Ghareghani[2,4], Human Genome Structural Variation Consortium[5], Peter M. Lansdorp[6,7], Benedict Paten[3], Scott E. Devine[8], Ashley D. Sanders[9], Charles Lee[10], Mark J.P. Chaisson[11], Jan O. Korbel[9], Evan E. Eichler[1,12†], Tobias Marschall[2,4†]

Workflow Overview

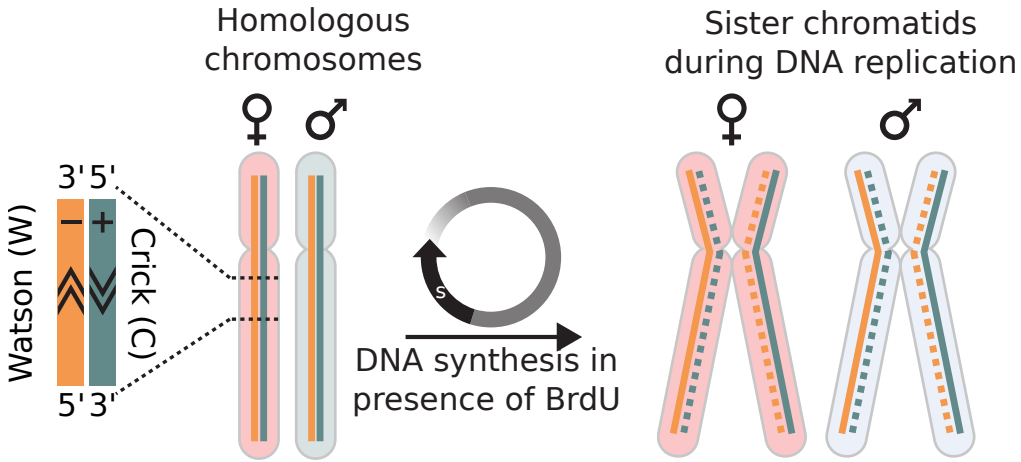


Strand-specific single-cell sequencing (Strand-seq)



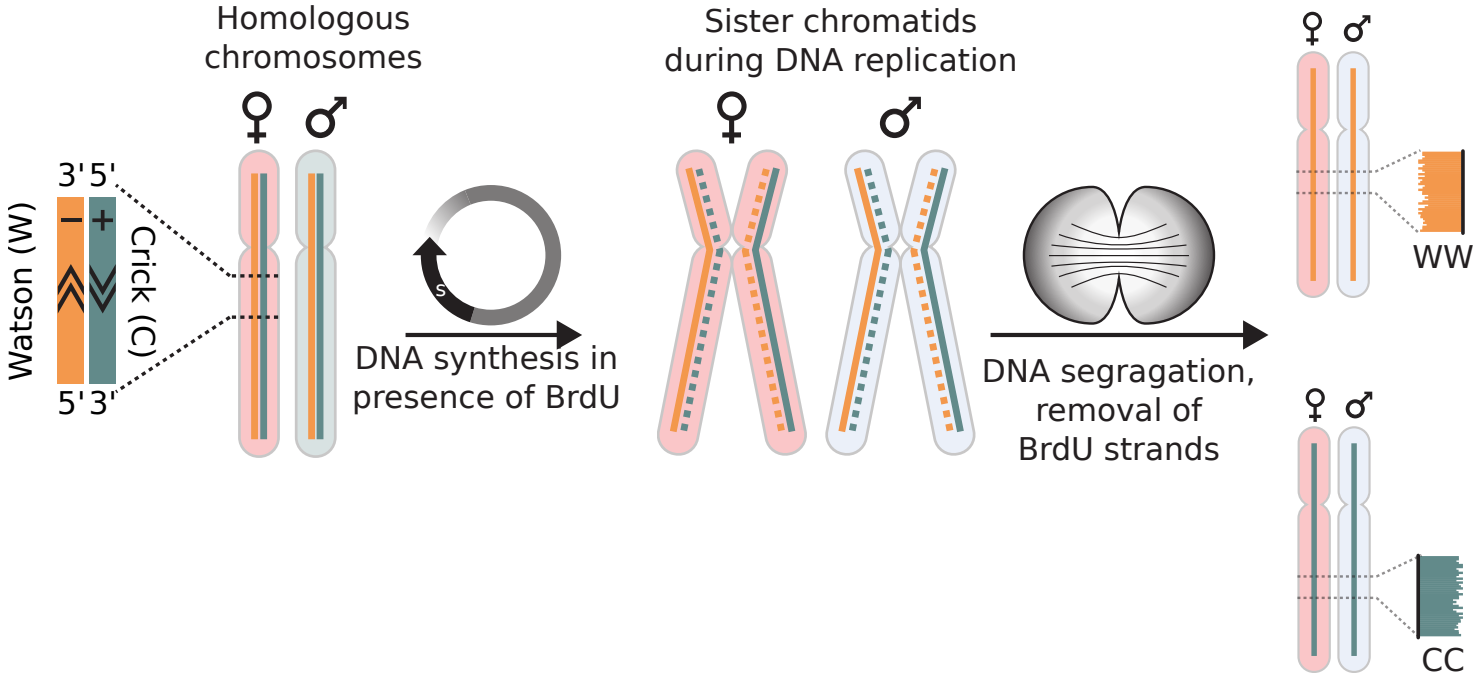
[Figure adapted from D Porubský, A Sanders, et al., Genome Research, 2016]

Strand-specific single-cell sequencing (Strand-seq)



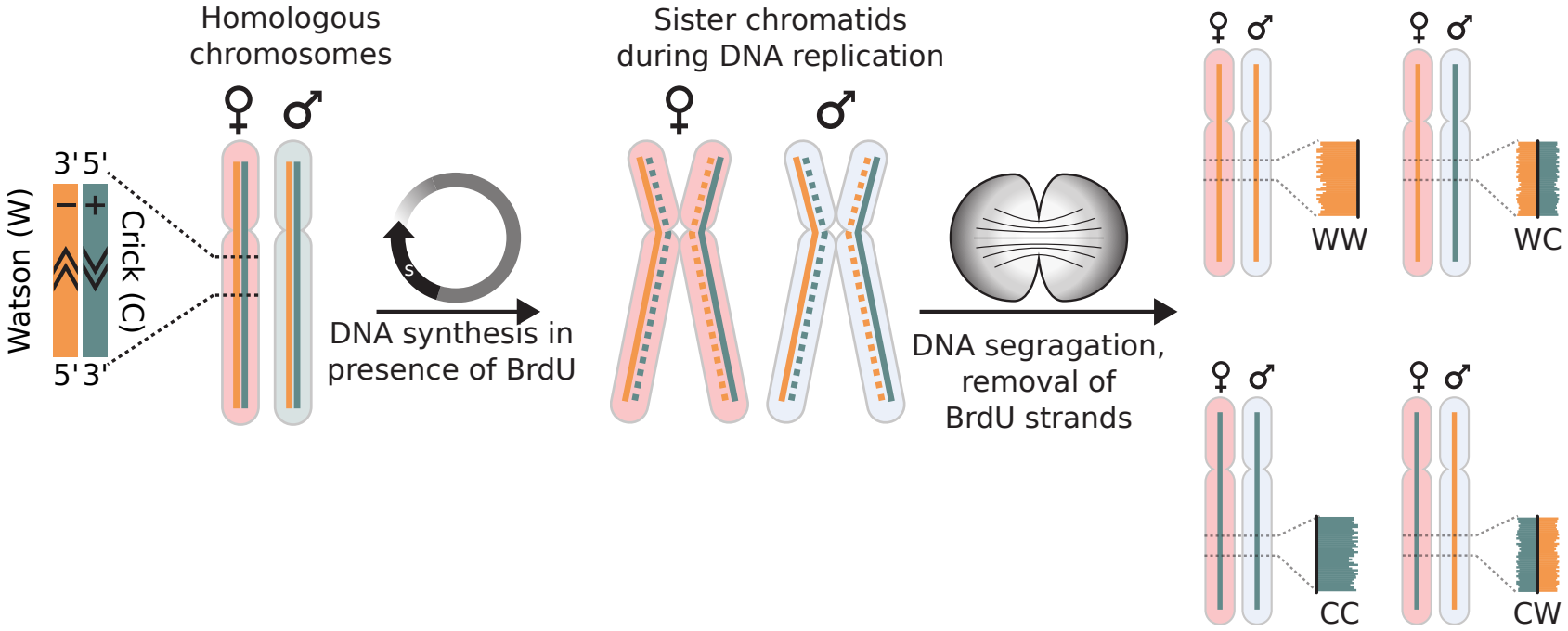
[Figure adapted from D Porubský, A Sanders, et al., Genome Research, 2016]

Strand-specific single-cell sequencing (Strand-seq)



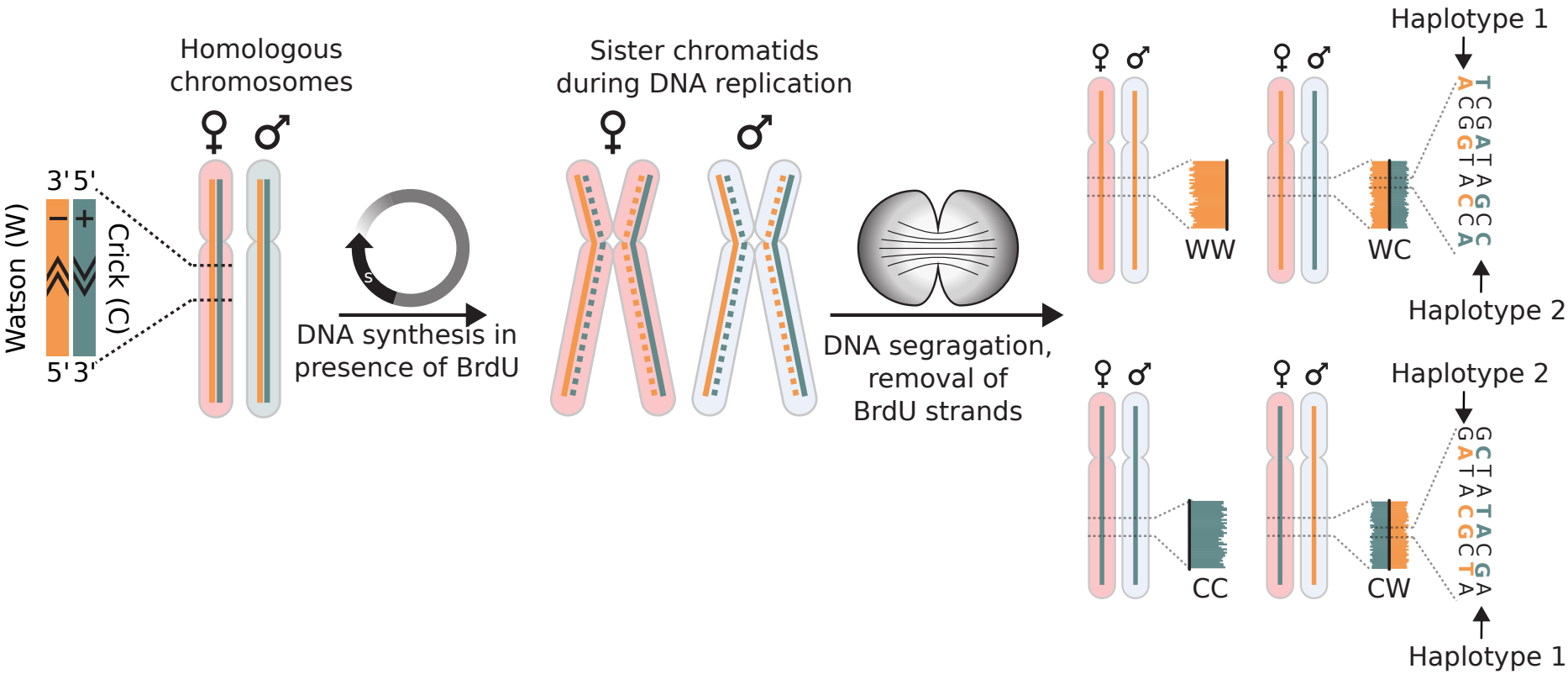
[Figure adapted from D Porubský, A Sanders, et al., Genome Research, 2016]

Strand-specific single-cell sequencing (Strand-seq)



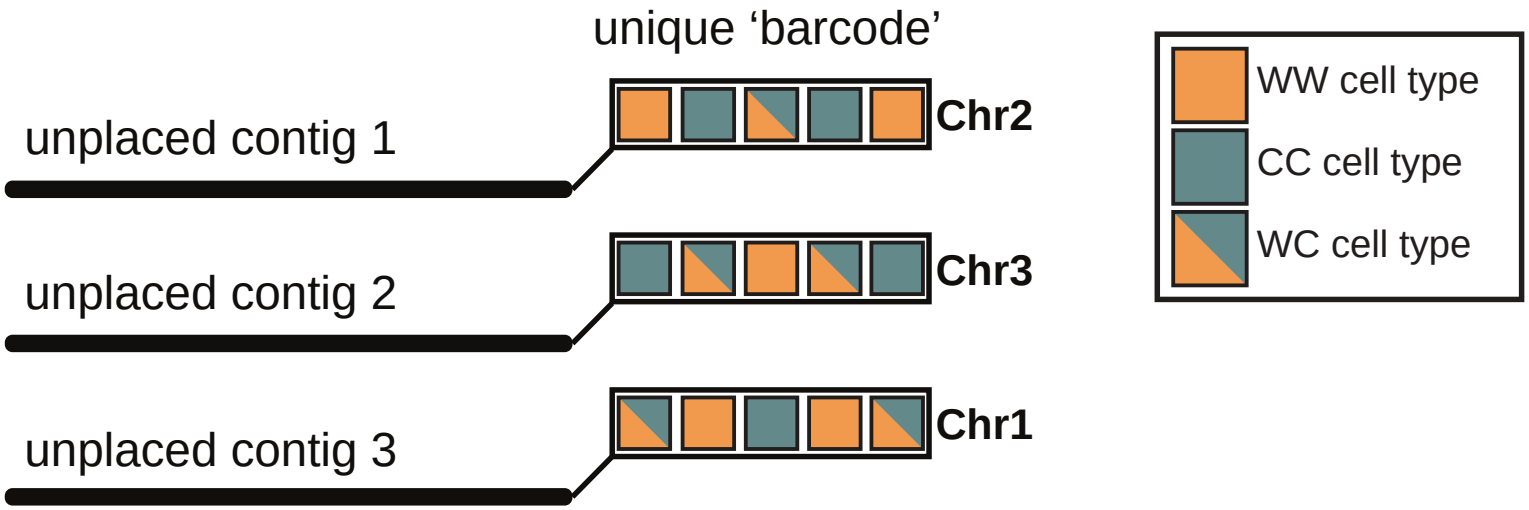
[Figure adapted from D Porubský, A Sanders, et al., Genome Research, 2016]

Strand-specific single-cell sequencing (Strand-seq)



[Figure adapted from D Porubský, A Sanders, et al., Genome Research, 2016]

Strand-seq can cluster contigs by chromosome



Bioinformatics, 34, 2018, i115–i123
doi: 10.1093/bioinformatics/bty290
ISMB 2018

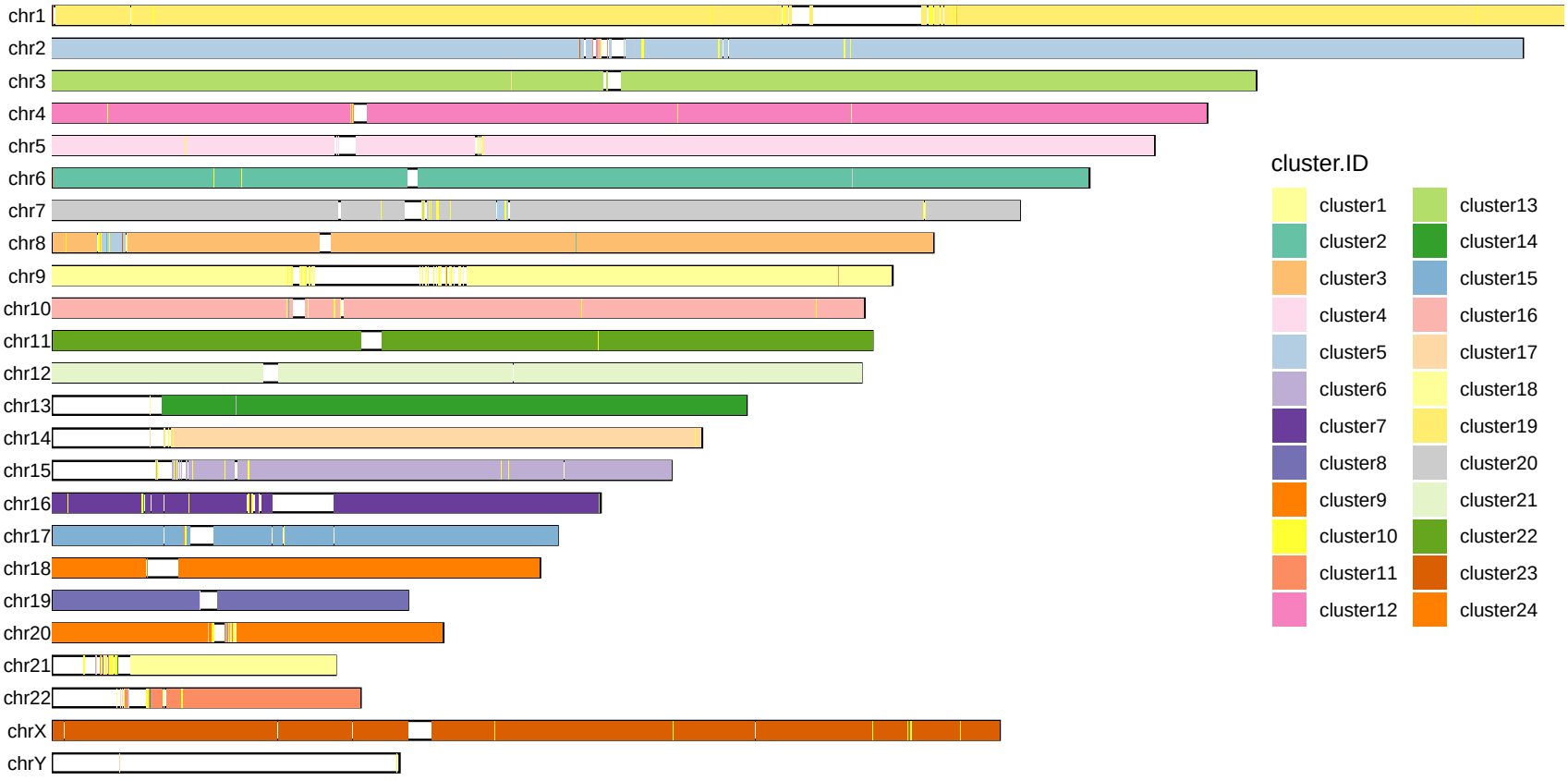
OXFORD

Strand-seq enables reliable separation of long reads by chromosome via expectation maximization

Maryam Ghareghani[1,2,3,†], David Porubskỳ[1,2,†], Ashley D. Sanders[4], Sascha Meiers[4], Evan E. Eichler[5,6], Jan O. Korbel[4] and Tobias Marschall[1,2,*]

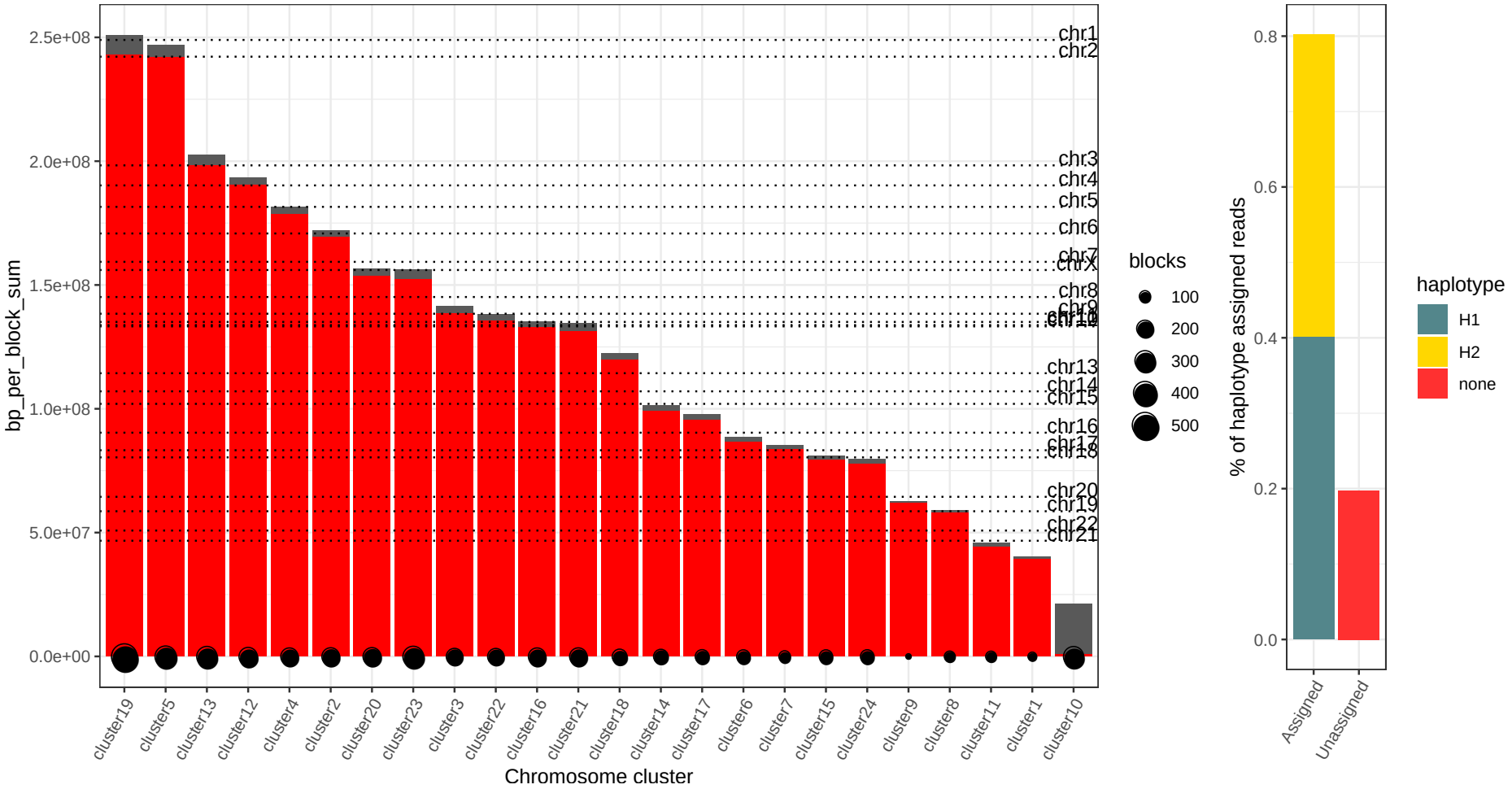
Clustering “collapsed” Canu assembly

HiFi reads of HG00733

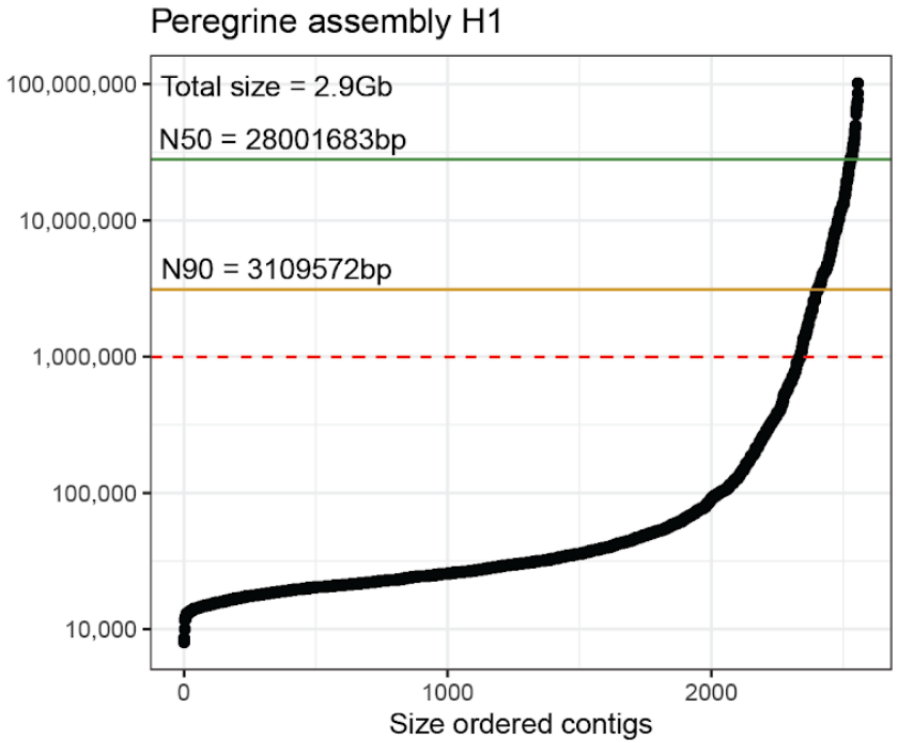


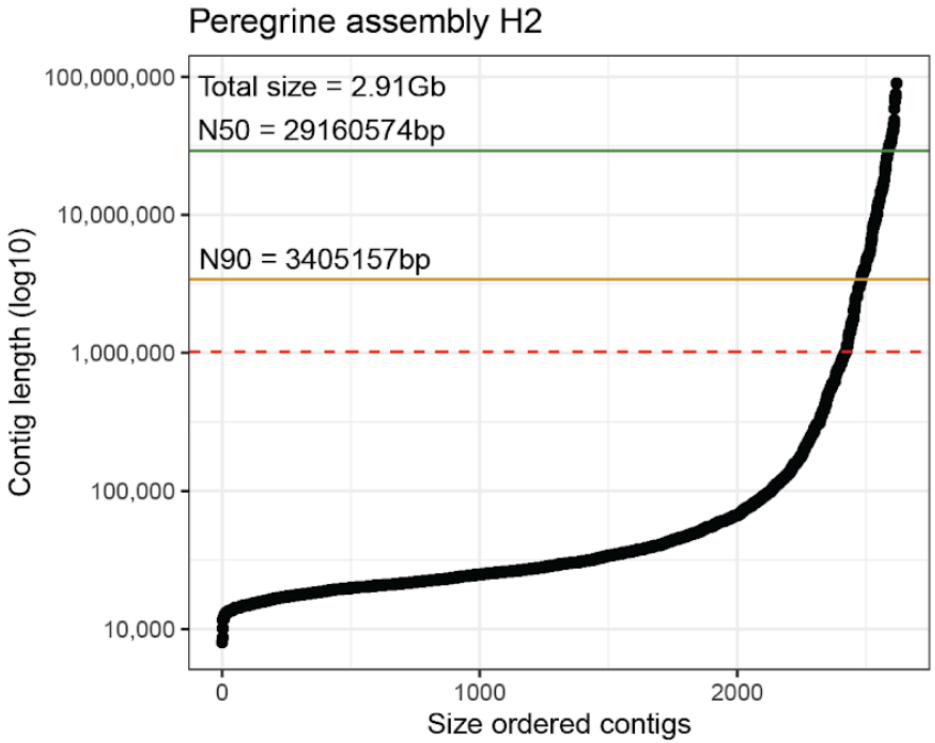
99.6% of contig sequence maps back to correct chromosome

Phasing of each cluster



Resulting phased assembly





N50 after fixing assembly errors:
25,893,285 bp

N50 after fixing assembly errors:
28,927,707 bp

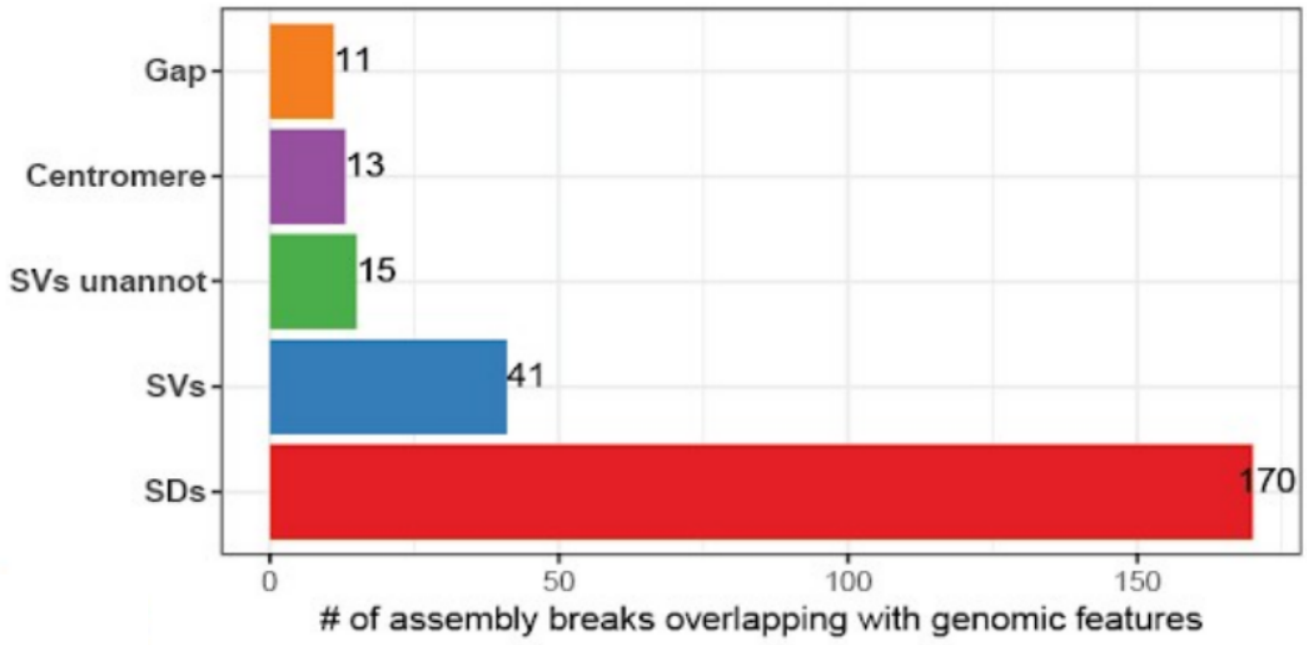
QV 40.8 as estimated from 77 BACs

Phasing validation using parents

Switch error rate 0.4%

Universal Assembly Breaks

Assembly breaks shared by Peregrine and Canu



- 150/250 shared with trio-binned ONT assembly using Shasta
- 64/250 shared with T2T assembly of CHM13

Summary

- Partitioning reads without reference and parent information
- Haplotype-resolved assemblies N50 >25Mbp, QV >40
- Also works with CLR (not shown here)
- Universal assembly breaks shared across sequencing platforms and assembly tools

Funding:

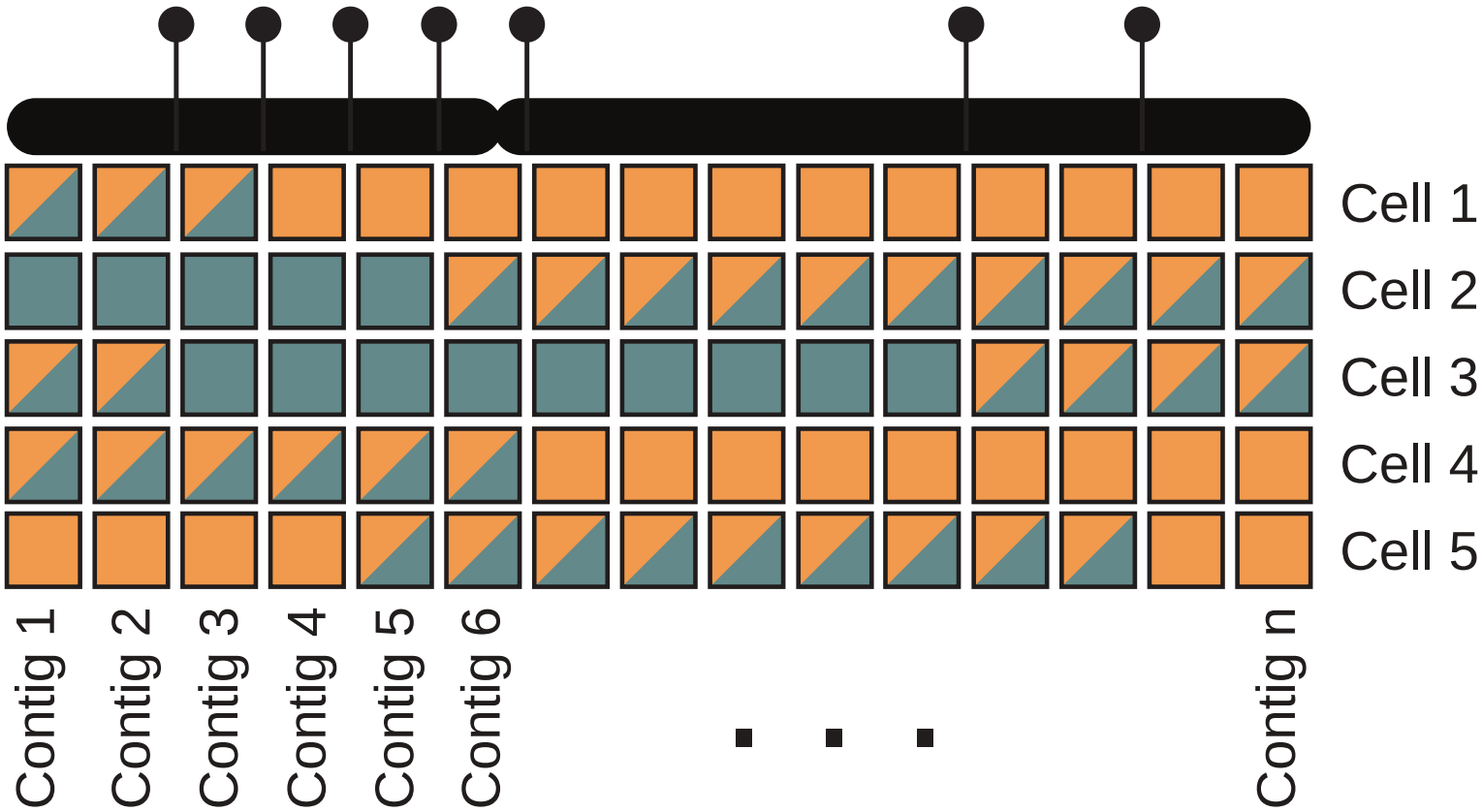
NIH National Institutes of Health

DFG Deutsche Forschungsgemeinschaft

Bundesministerium für Bildung und Forschung

Appendix

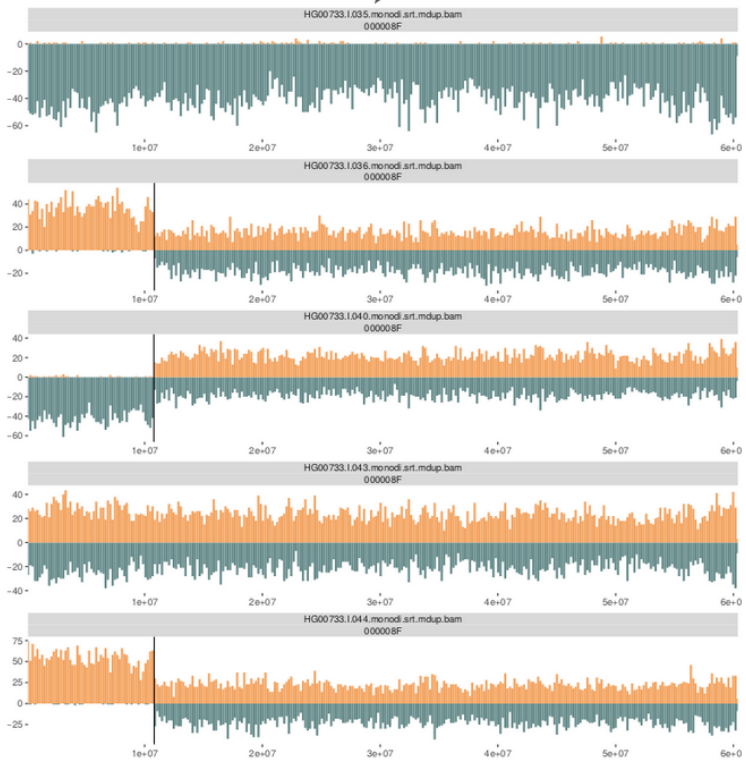
Strand-seq can cluster contigs by chromosome



Using Strand-seq to detect assembly errors

ctg	start	end	direction	cluster.ID
000008F	1	10772584	dir	cluster2
000008F	10772585	60481328	dir	cluster1
000006F	1	25346285	dir	cluster15
000006F	25346286	66162004	revcomp	cluster15

Chimerism



Misorientation