PACIFIC BIOSCIENCES®

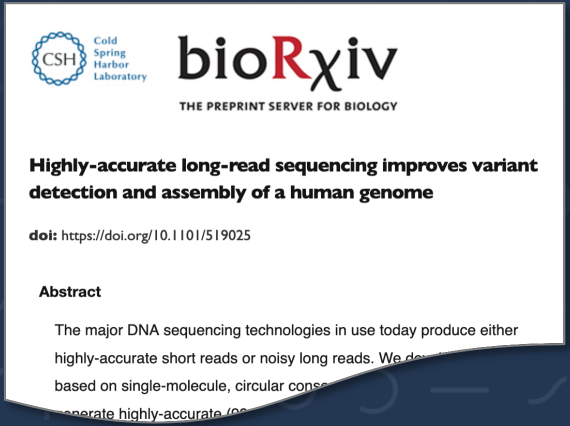
Advantages of HiFi reads for variant discovery and genome assembly

William Rowell, Senior Scientist, Bioinformatics Applications, PacBio

@nothingclever
#SMRTLeiden

AGENDA

- Introduction to HiFi
- Variant Calling
- *De Novo* Assembly
- Coverage Recommendations
- Public HiFi Datasets



CSH Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

Highly-accurate long-read sequencing improves variant detection and assembly of a human genome

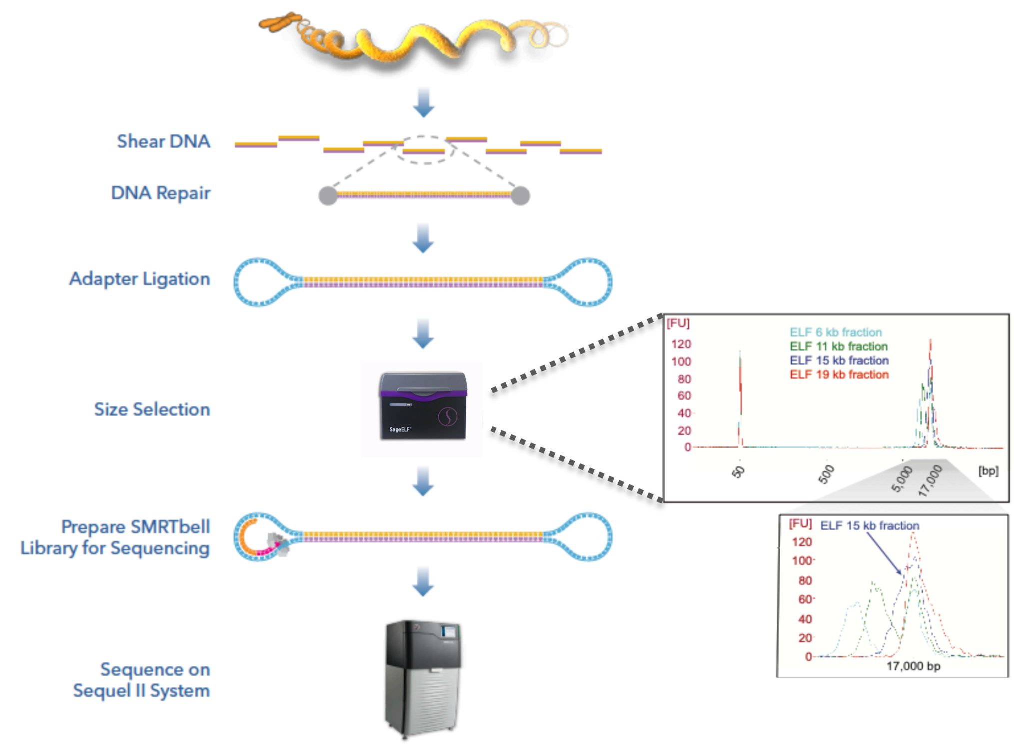
doi: https://doi.org/10.1101/519025

Abstract

The major DNA sequencing technologies in use today produce either highly-accurate short reads or noisy long reads. We de based on single-molecule, circular cons enerate highly-accurate (

Introduction to HiFi

HIFI LIBRARY PREP PRODUCES UNIFORM INSERT SIZES


Shear DNA

DNA Repair

Adapter Ligation

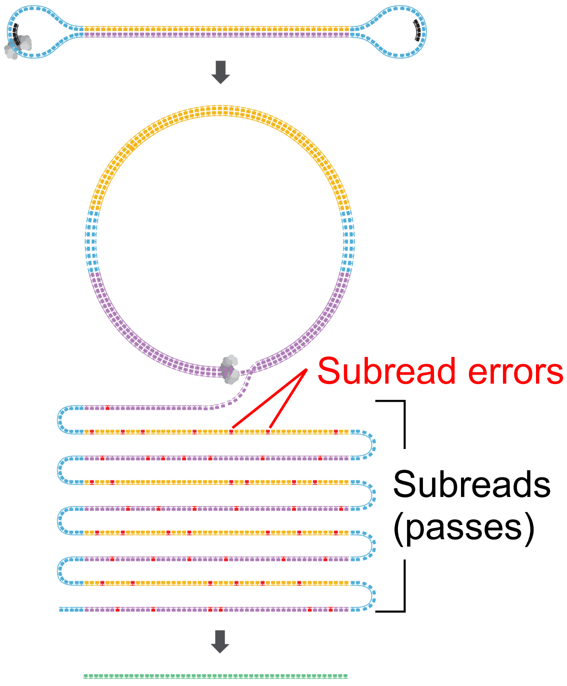
Size Selection

Prepare SMRTbell Library for Sequencing

Sequence on Sequel II System

Wenger, Peluso, et al. (2019). *bioRxiv*. doi:10.1101/519025

PACBIO CIRCULAR CONSENSUS SEQUENCING (CCS)



First round

Rolling circle

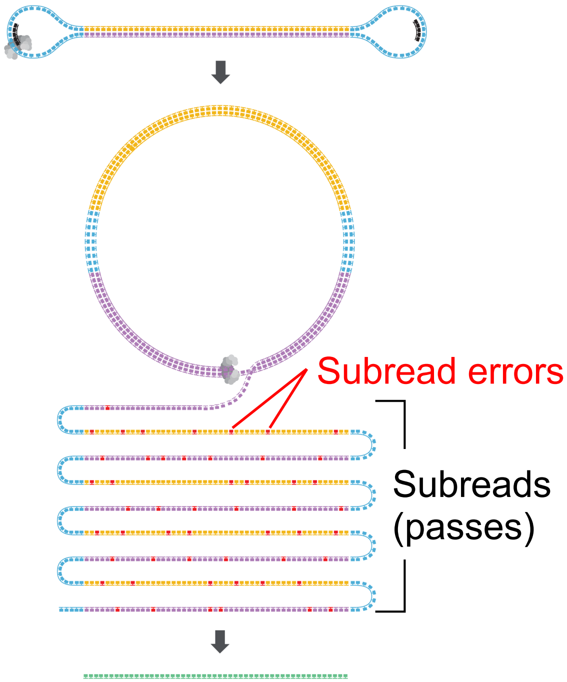
Generate consensus HiFi read

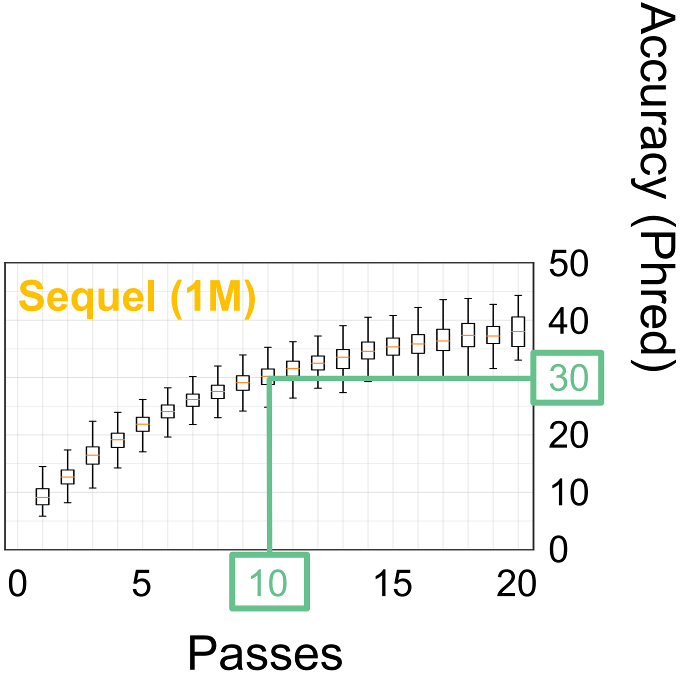
PACBIO CIRCULAR CONSENSUS SEQUENCING (CCS)

First round

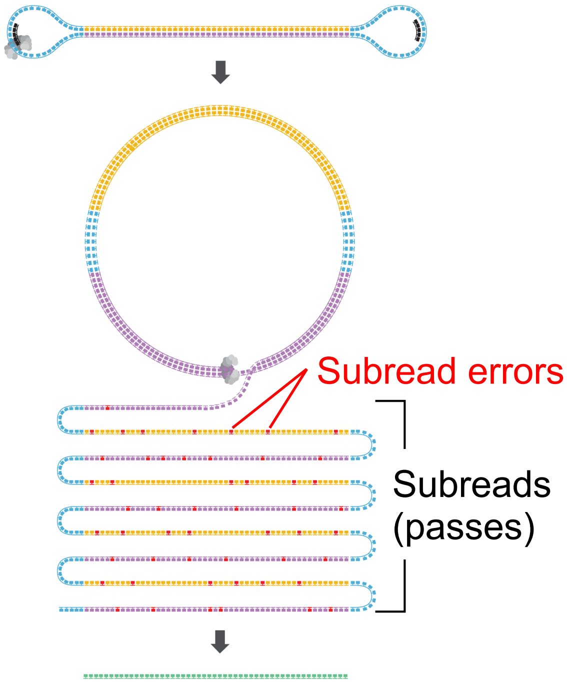
Rolling circle

Generate consensus HiFi read





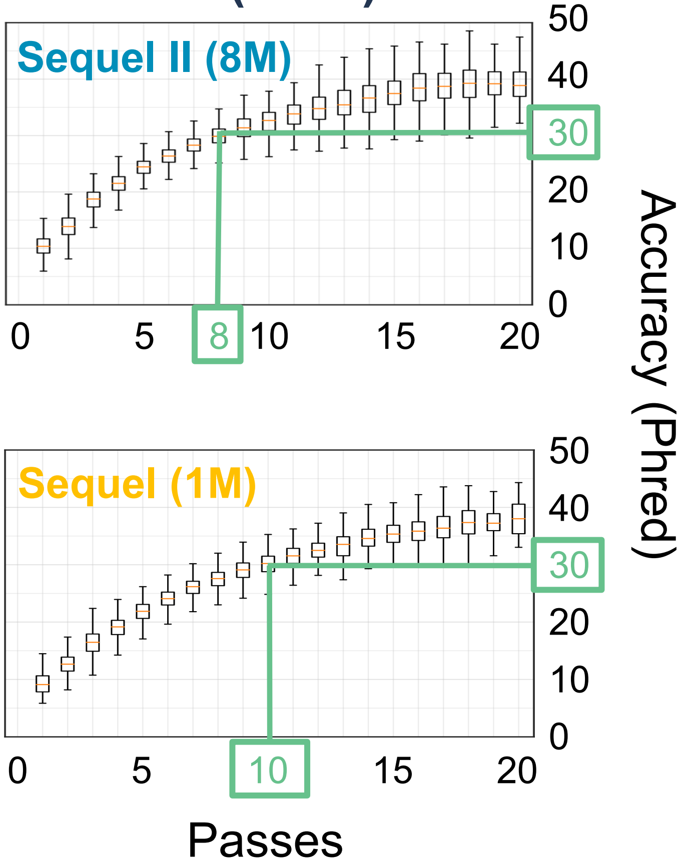
PACBIO CIRCULAR CONSENSUS SEQUENCING (CCS)



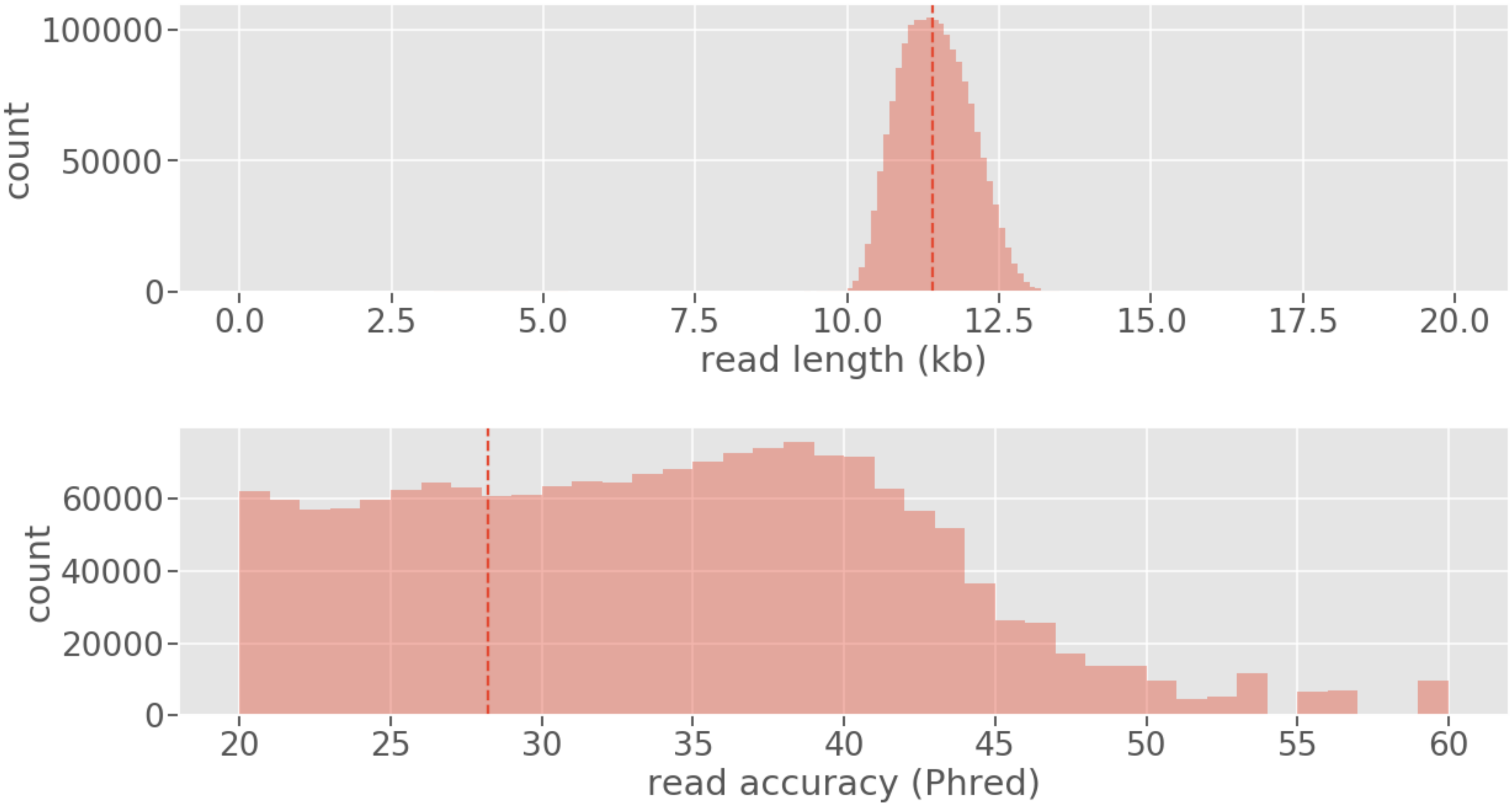
First round

Rolling circle

Generate consensus HiFi read



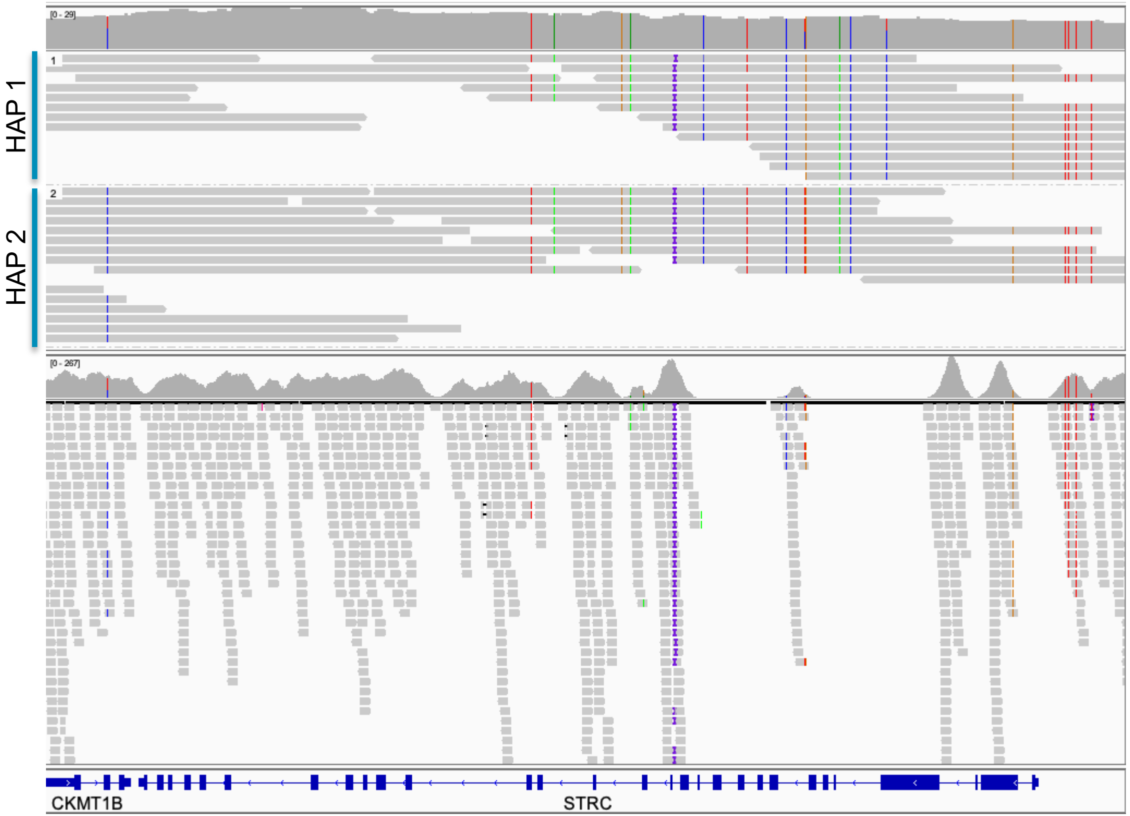
HIFI READS ARE LONG AND ACCURATE



HIFI READS ARE EASILY MAPPED TO REPETITIVE REGIONS

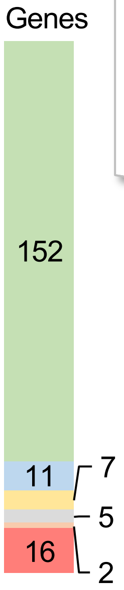
11 kb HiFi

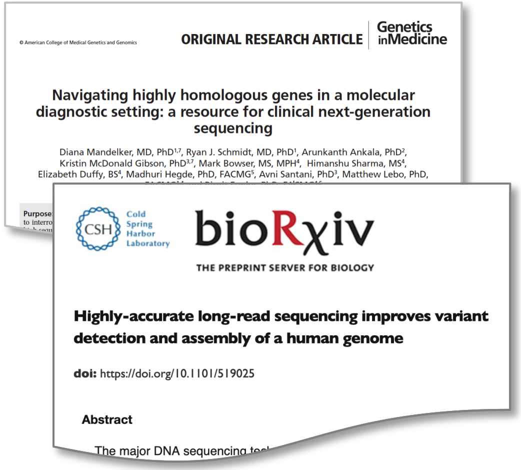
2x250 bp



DETECT MORE VARIANTS IN MEDICALLY-RELEVANT GENES

% problem exons resolved	Genes
100%	*ABCC6, ABCD1, ACAN, ACSM2B, AKR1C2, ALG1, ANKRD11, BCR, CATSPER2, CD177, CEL, CES1, CFH, CFHR1, CFHR3, CFHR4, CGB, CHEK2, CISD2, CLCNKA, CLCNKB, CORO1A, COX10, CRYBB2, CSH1, CYP11B1, CYP11B2, CYP21A2, CYP2A6, CYP2D6, CYP2F1, CYP4A22, DDX11, DHRS4L1, DIS3L2, DND1, DPY19L2, DUOX2, ESRRA, F8, FAM120A, FAM205A, FANCD2, FCGR1A, FCGR2A, FCGR3A, FCGR3B, FLG, FLNC, FOXD4, FOXO3, FUT3, GBA, GFRA2, GON4L, GRM5, GSTM1, GYPA, GYPB, GYPE, HBA1, HBA2, HBG1, HBG2, HP, HS6ST1, IDS, IFT122, IKBKG, IL9R, KIR2DL1, KIR2DL3, KMT2C, KRT17, KRT6A, KRT6B, KRT6C, KRT81, KRT86, LEFTY2, LPA, MST1, MUC5B, MYH6, MYH7, NEB, NLGN4X, NLGN4Y, NOS2, NOTCH2, NXF5, OPN1LW, OR2T5, OR51A2, PCDH11X, PCDHB4, PGAM1, PHC1, PIK3CA, PKD1, PLA2G10, PLEKHM1, PLG, PMS2, PRB1, PRDM9, PROS1, RAB40AL, RALGAPA1, RANBP2, RHCE, RHD, RHPN2, ROCK1, SAA1, SDHA, SDHC, SFTPA1, SFTPA2, SIGLEC14, SLC6A8, SMG1, SPATA31C1, SPTLC1, SRGAP2, SSX7, STAT5B, STK19, STRC, SULT1A1, SUZ12, TBX20, TCEB3C, TLR1, TLR6, TMEM231, TNXB, TRIOBP, TRPA1, TTN, TUBA1A, TUBB2B, UGT1A5, UGT2B15, UGT2B17, UNC93B1, VCY, VWF, WDR72, ZNF419, ZNF592, ZNF674*
[75%, 100%)	*ANAPC1, C4A, C4B, CHRNA7, CR1, DUX4, FCGR2B, HYDIN, OTOA, PDPK1, TMLHE*
[50%, 75%)	*ADAMTSL2, CDY2A, DAZ1, GTF2I, NAIP, OCLN, RPS17*
[25%, 50%)	*DAZ2, DAZ3, KIR3DL1, OPN1MW, PPIP5K1*
(0%, 25%)	*NCF1, RBMY1A1*
0%	*BPY2, CCL3L1, CCL4L1, CDY1, CFC1, CFC1B, GTF2IRD2, HSFY1, MRC1, OR4F5, PRY, PRY2, SMN1, SMN2, TSPY1, XKRY*




© American College of Medical Genetics and Genomics

ORIGINAL RESEARCH ARTICLE | Genetics inMedicine

Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing

Diana Mandelker, MD, PhD[1,7], Ryan J. Schmidt, MD, PhD[1], Arunkanth Ankala, PhD[2], Kristin McDonald Gibson, PhD[3,7], Mark Bowser, MS, MPH[4], Himanshu Sharma, MS[4], Elizabeth Duffy, BS[4], Madhuri Hegde, PhD, FACMG[5], Avni Santani, PhD[3], Matthew Lebo, PhD,

Purpose

to interro

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

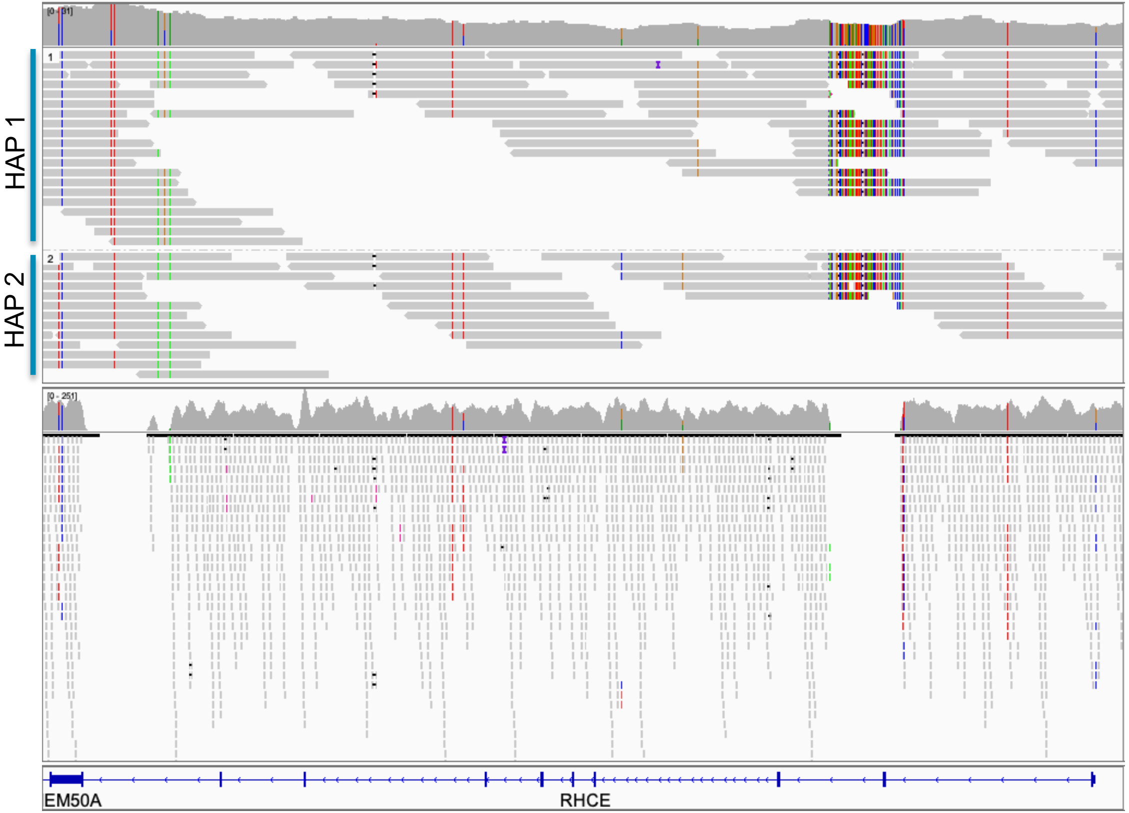
Highly-accurate long-read sequencing improves variant detection and assembly of a human genome

doi: https://doi.org/10.1101/519025

Abstract

The major DNA sequencing tec

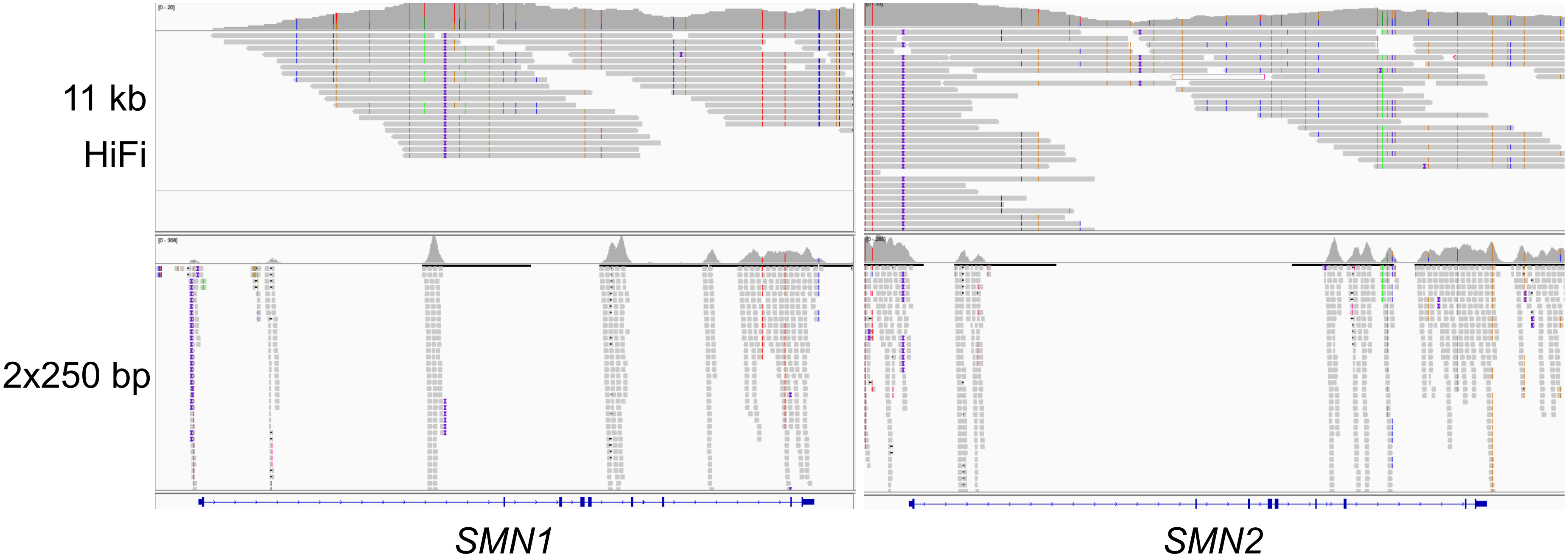
IMPROVED MAPPING IN REFERENCE-DIVERGENT REGIONS



11 kb HiFi

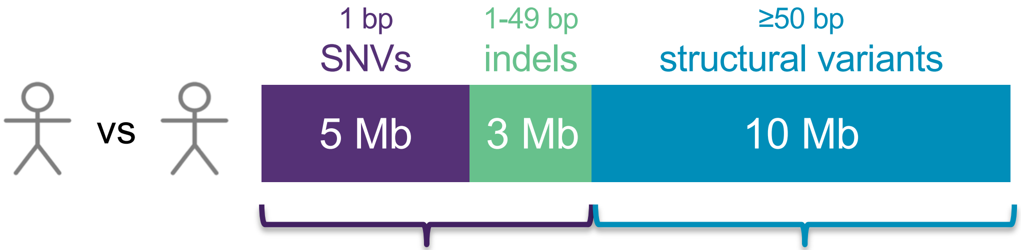
2x250 bp

IMPROVED MAPPING IN SEGMENTAL DUPLICATIONS



Variant Calling

GENOME VARIATION COMES IN ALL SIZES



vs

1 bp
SNVs
5 Mb

1-49 bp
indels
3 Mb

≥50 bp
structural variants
10 Mb

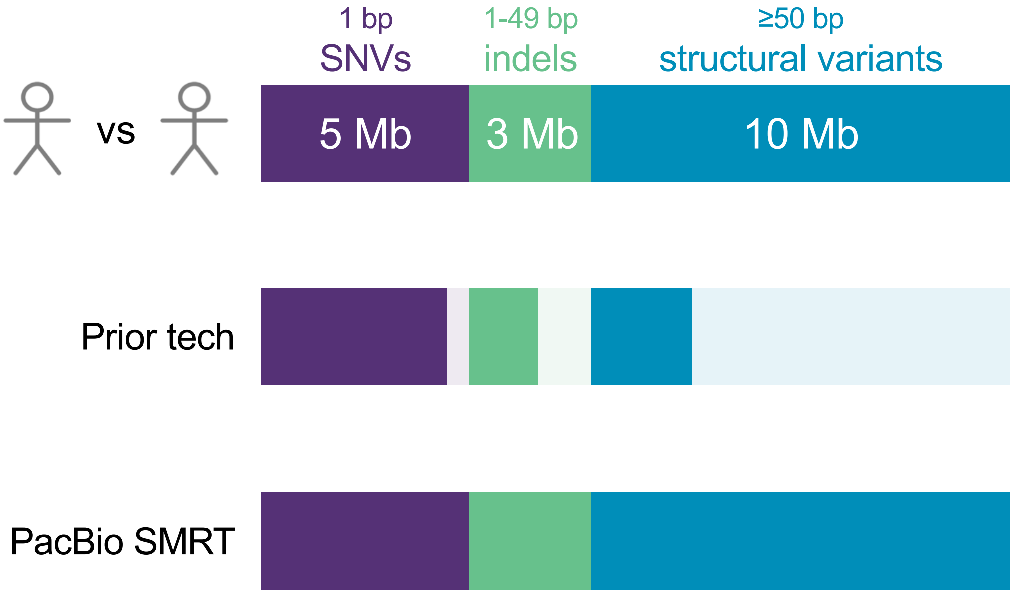
"Small variants":

- Single Nucleotide Variants (SNVs)
- Indels <50 bp

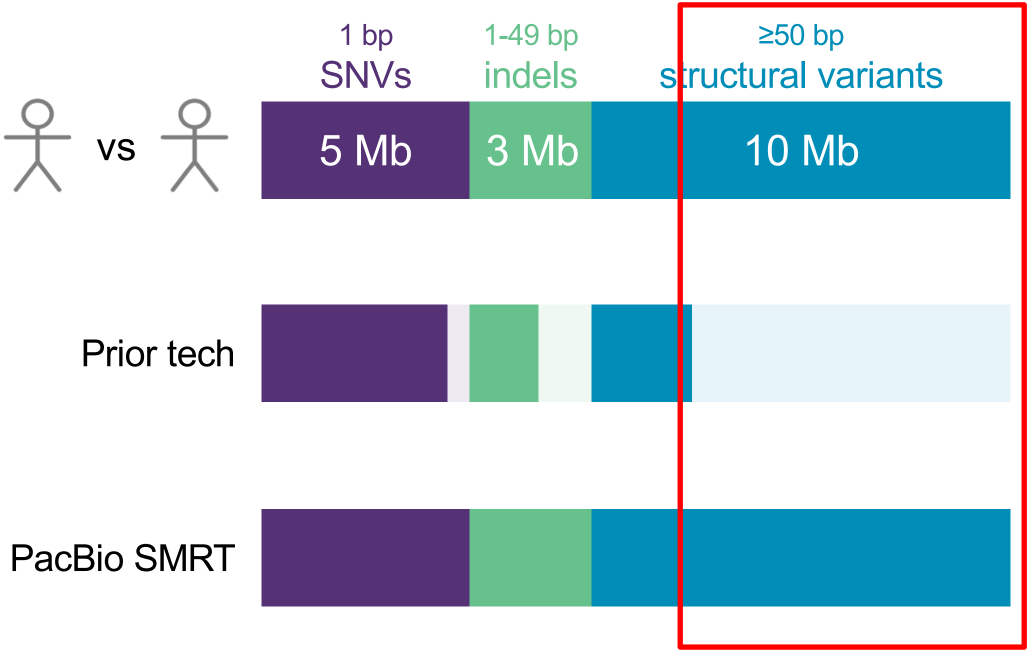
Structural Variants (SVs):

- Indels ≥50 bp
  - Duplications
  - Copy Number Variants (CNVs)
- Translocations
- Inversions

OTHER TECHNOLOGIES MISS VARIANTS

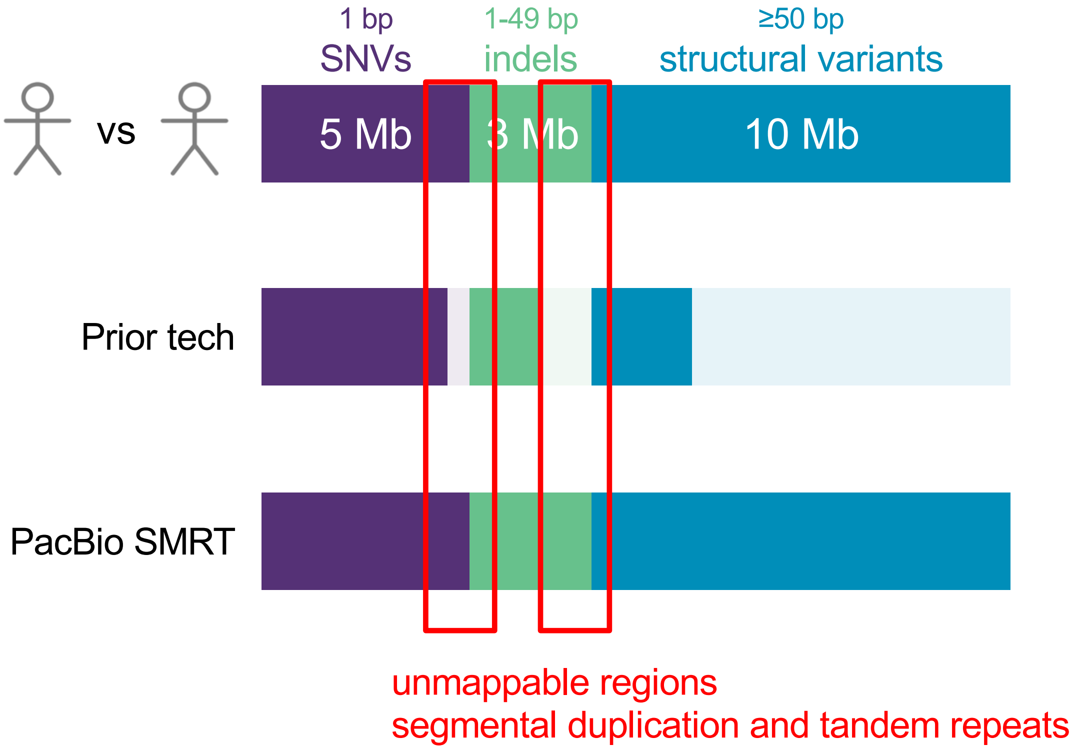


PACBIO ENABLES STRUCTURAL VARIANT DETECTION



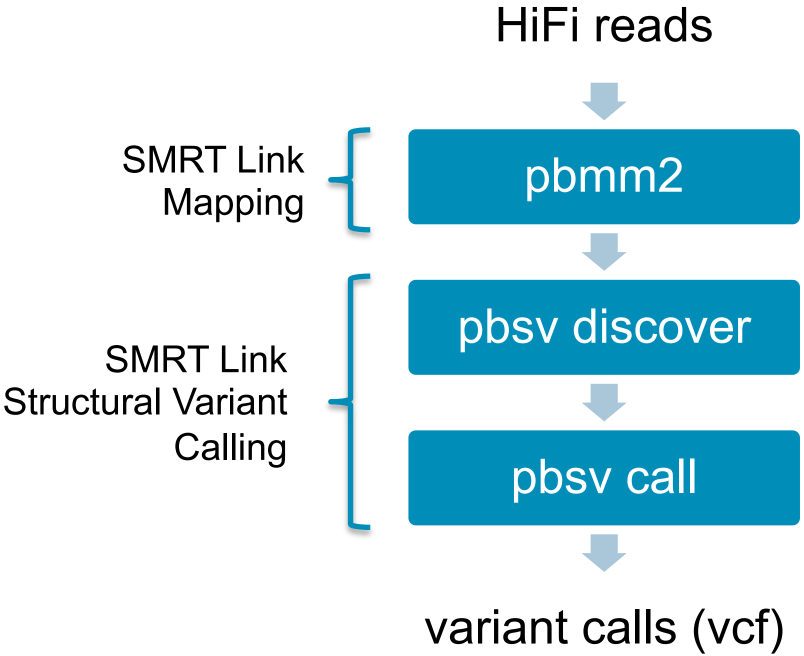
long insertions
events in repeat regions

PACBIO HIFI READS ENABLE SMALL VARIANT DETECTION IN DIFFICULT-TO-MAP REGIONS



unmappable regions
segmental duplication and tandem repeats

WGS HIFI STRUCTURAL VARIANT CALLING OVERVIEW



HiFi reads

SMRT Link Mapping

pbmm2

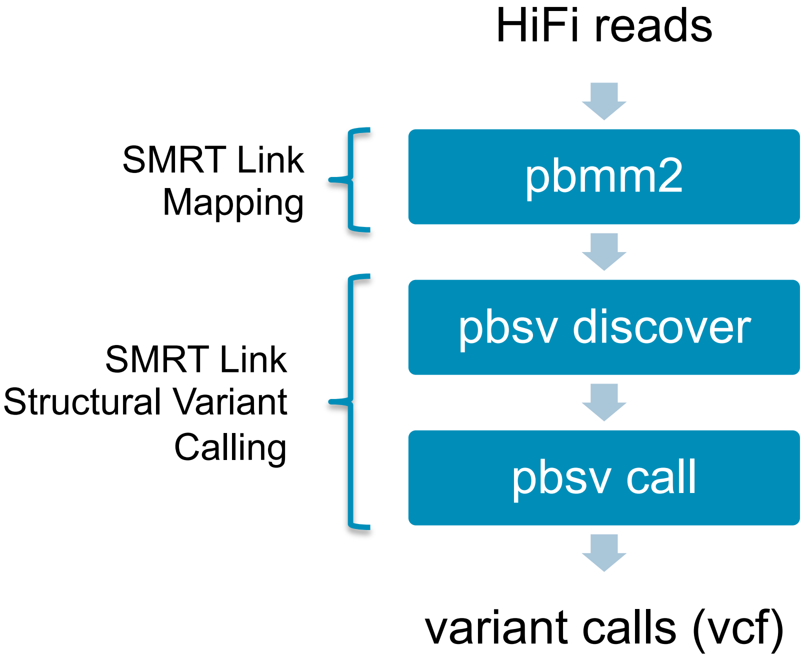
SMRT Link Structural Variant Calling

pbsv discover

pbsv call

variant calls (vcf)

WGS HIFI STRUCTURAL VARIANT CALLING OVERVIEW



HiFi reads

SMRT Link Mapping

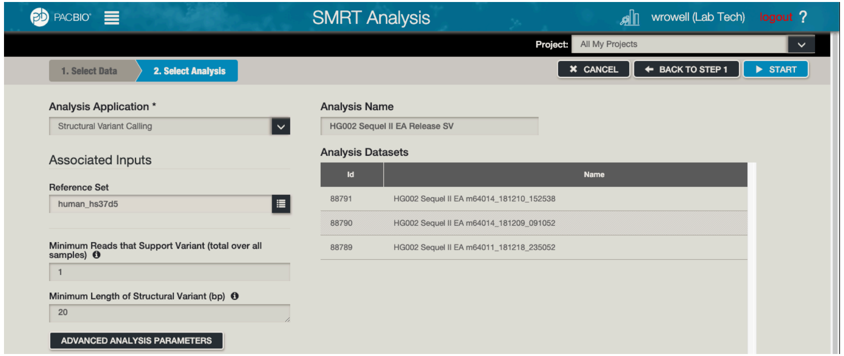
pbmm2

SMRT Link Structural Variant Calling

pbsv discover

pbsv call

variant calls (vcf)



Project: All My Projects

Analysis Application *: Structural Variant Calling

Associated Inputs

Reference Set: human_hs37d5

Minimum Reads that Support Variant (total over all samples): 1

Minimum Length of Structural Variant (bp): 20

Analysis Name: HG002 Sequel II EA Release SV

Analysis Datasets

Id	Name
88791	HG002 Sequel II EA m64014_181210_152538
88790	HG002 Sequel II EA m64014_181209_091052
88789	HG002 Sequel II EA m64011_181218_235052

OR

PACBIO® & BIOCONDA®

15-FOLD HIFI READS PROVIDE A COMPREHENSIVE VIEW OF STRUCTURAL VARIANTS ≥20BP

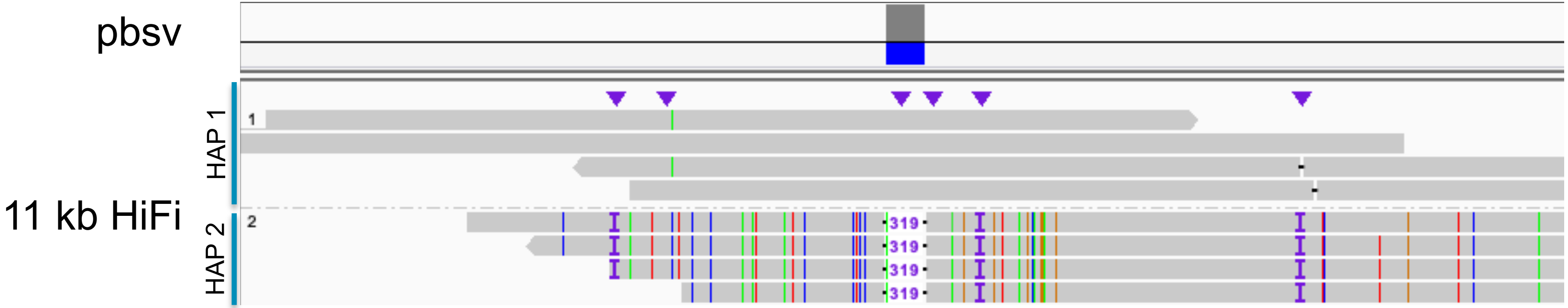
SVTYPE	HG001	HG002	HG005
BND	752	712	708
CNV	108	107	97
DEL	24,192	24,471	24,353
DUP	11,523	11,472	11,451
INS	20,638	20,820	21,066
INV	51	47	50

15-FOLD HIFI READS PROVIDE A COMPREHENSIVE VIEW OF STRUCTURAL VARIANTS ≥20BP

SVTYPE	HG001	HG002	HG005
BND	752	712	708
CNV	108	107	97
DEL	24,192	24,471	24,353
DUP	11,523	11,472	11,451
INS	20,638	20,820	21,066
INV	51	47	50

	Recall	Precision
3 SMRT Cells 8M HG002 (16-fold, 11kb)	96.8%	95.4%

HETEROZYGOUS *ALU* DELETION IN HG001

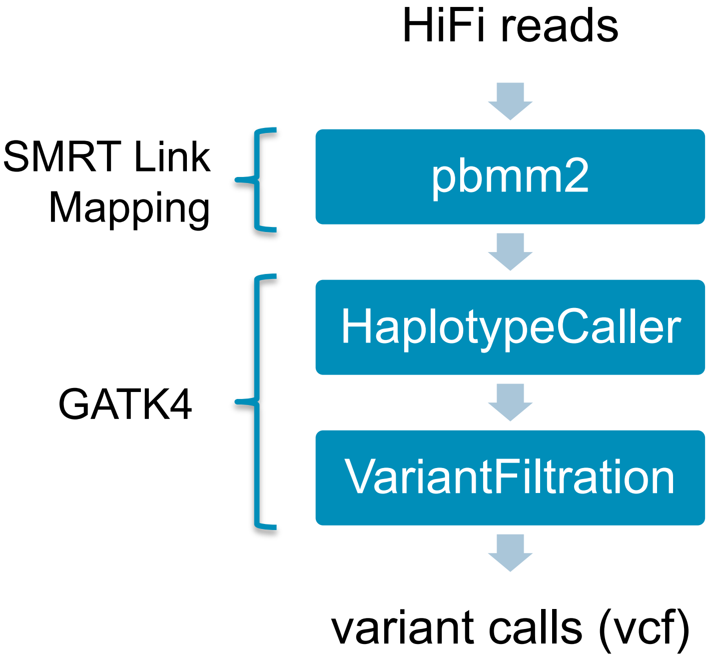


chrX:116,449,107-116,459,909

HOMOZYGOUS *APP* INTRONIC INVERSION IN HG001

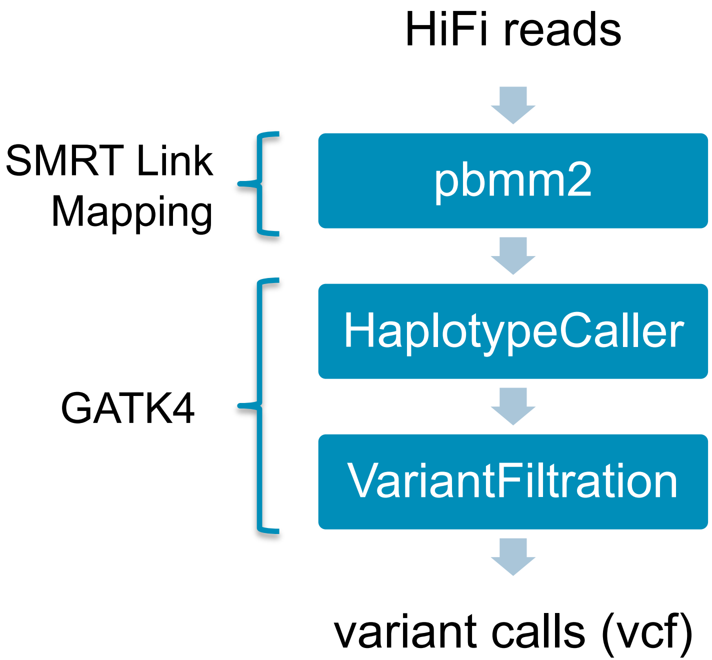
chr21:27,373,479-27,375,496

SMALL VARIANTS CAN BE DETECTED BY GATK HAPLOTYPECALLER



A framework for variation discovery and genotyping using next-generation DNA sequencing data DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M, 2011 *NATURE GENETICS 43:491-498*

SMALL VARIANTS CAN BE DETECTED BY GATK HAPLOTYPECALLER



HiFi reads

SMRT Link Mapping

pbmm2

GATK4

HaplotypeCaller

VariantFiltration

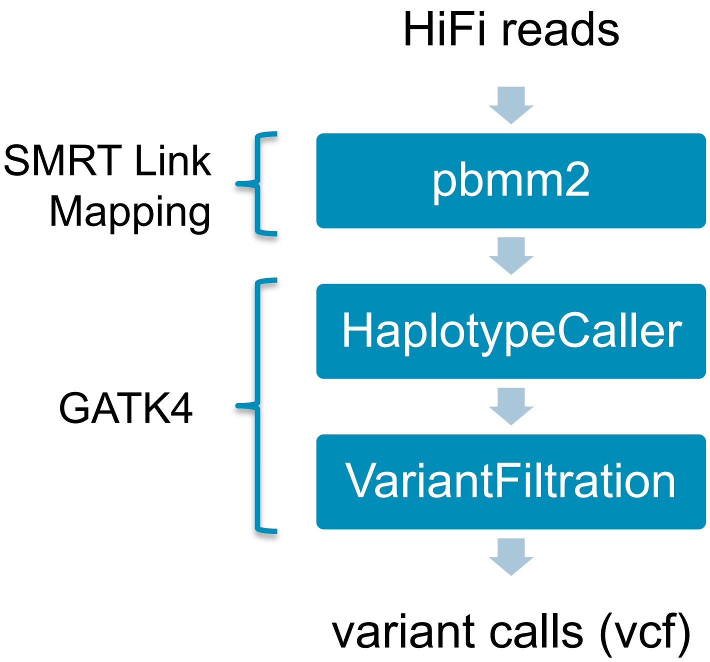
variant calls (vcf)

15-fold HiFi HG002 against GIAB v3.3.2 benchmark

	Precision	**Recall**
SNVs	99.6%	99.7%
Indels	85.0%	82.3%

A framework for variation discovery and genotyping using next-generation DNA sequencing data DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M, 2011 *NATURE GENETICS 43:491-498*

SMALL VARIANTS CAN BE DETECTED BY GATK HAPLOTYPECALLER


HiFi reads

SMRT Link Mapping

pbmm2

GATK4

HaplotypeCaller

VariantFiltration

variant calls (vcf)

15-fold HiFi HG002 against GIAB v3.3.2 benchmark

	Precision	Recall
SNVs	99.6%	99.7%
Indels	85.0%	82.3%

- High SNP Recall and Precision
- Lower Indel Recall and Precision

- HaplotypeCaller optimized for error mode of short reads:
  - [mismatch error] >> [indel error]

A framework for variation discovery and genotyping using next-generation DNA sequencing data DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M, 2011 *NATURE GENETICS 43:491-498*

DEEPVARIANT CAN BE TRAINED TO LEARN NEW ERROR MODELS

New sequence data type (HiFi)

Known Genotypes (GIAB)

https://rdcu.be/7Dhl - Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987 (2018)

DEEPVARIANT CAN BE TRAINED TO LEARN NEW ERROR MODELS

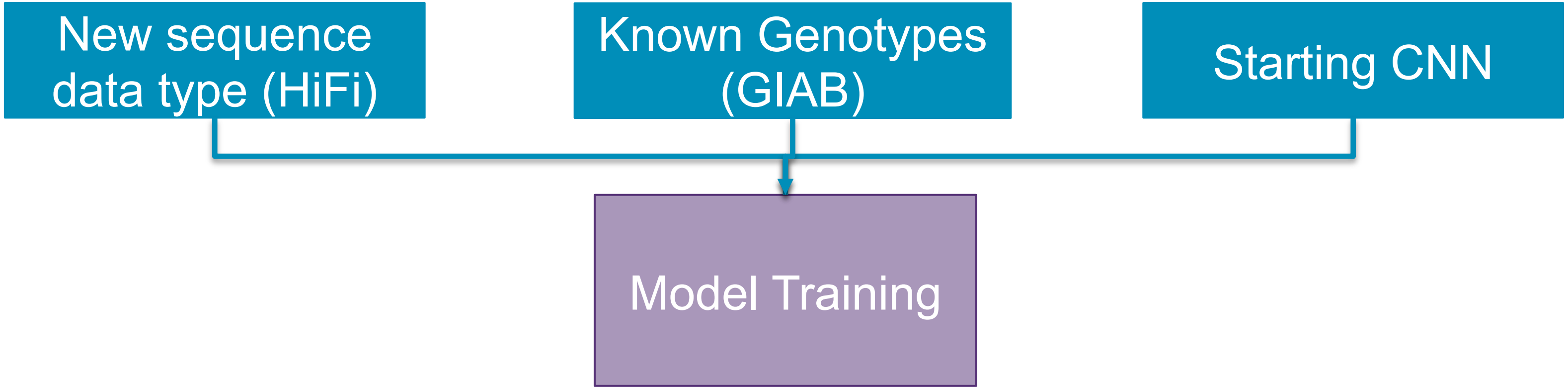
New sequence data type (HiFi)

Known Genotypes (GIAB)

Starting CNN

https://rdcu.be/7Dhl - Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987 (2018)

DEEPVARIANT CAN BE TRAINED TO LEARN NEW ERROR MODELS


New sequence data type (HiFi)

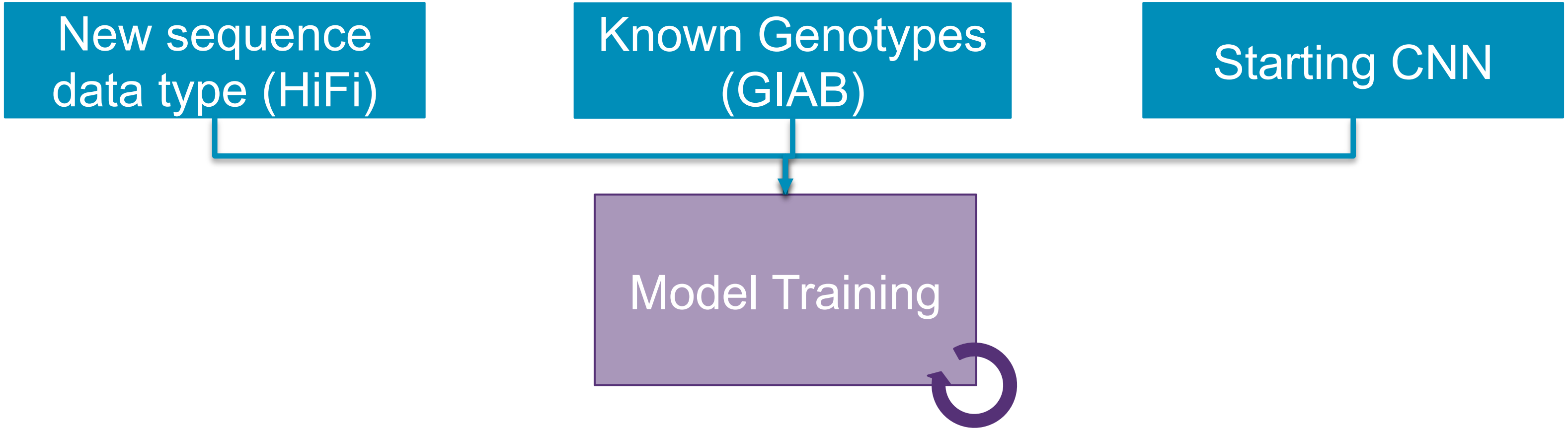
Known Genotypes (GIAB)

Starting CNN

Model Training

https://rdcu.be/7Dhl - Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987 (2018)

DEEPVARIANT CAN BE TRAINED TO LEARN NEW ERROR MODELS



New sequence data type (HiFi)

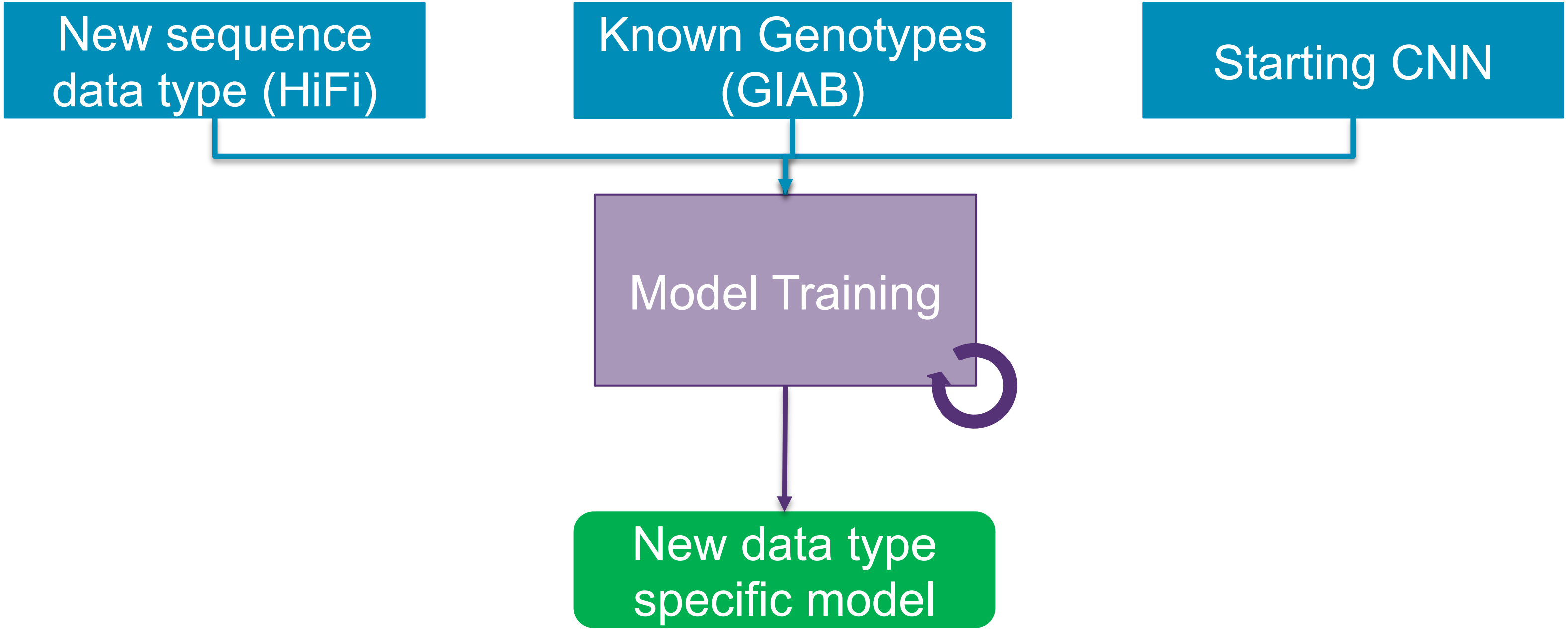
Known Genotypes (GIAB)

Starting CNN

Model Training

https://rdcu.be/7Dhl - Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987 (2018)

DEEPVARIANT CAN BE TRAINED TO LEARN NEW ERROR MODELS



New sequence data type (HiFi)

Known Genotypes (GIAB)

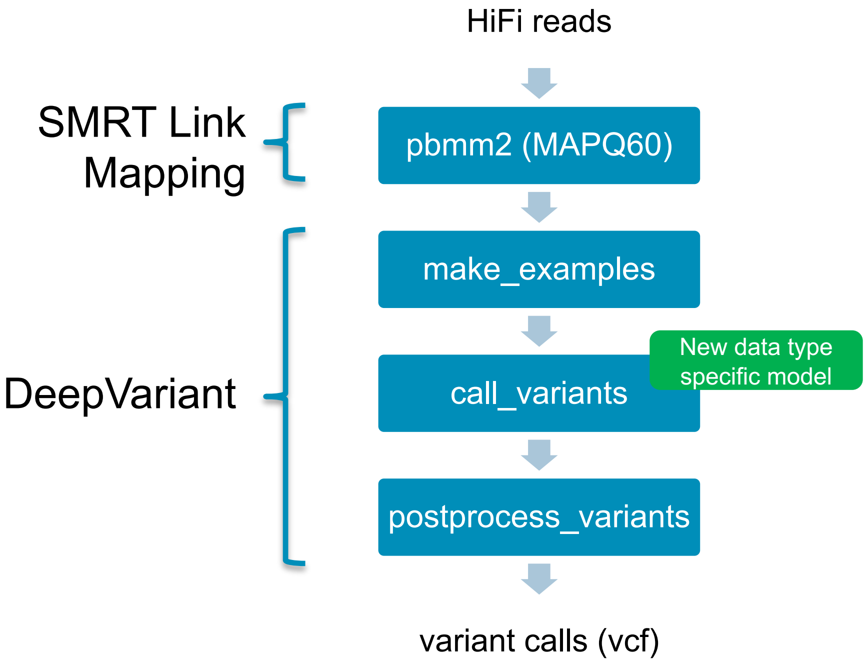
Starting CNN

Model Training

New data type specific model

https://rdcu.be/7Dhl - Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987 (2018)

DEEPVARIANT IMPROVES SMALL VARIANT DETECTION



HiFi reads

SMRT Link Mapping

pbmm2 (MAPQ60)

DeepVariant

make_examples

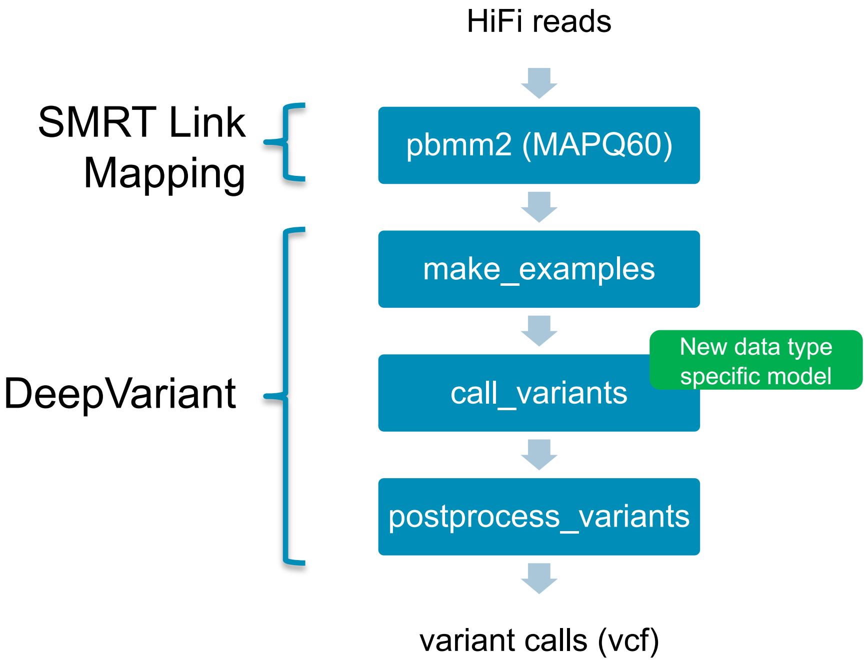
call_variants

New data type specific model

postprocess_variants

variant calls (vcf)

DEEPVARIANT IMPROVES SMALL VARIANT DETECTION



15-fold HiFi against GIAB v3.3.2 benchmarks

Sample	SNV Recall	SNV Precision	Indel Recall	Indel Precision
HG001	99.1%	99.5%	94.1%	95.0%
HG002	99.2%	99.5%	95.4%	96.6%
HG005	99.4%	99.7%	97.0%	97.5%

- DeepVariant learns error model of HiFi reads from training data.
- Improved precision and recall for both SNVs and Indels

DEEPVARIANT CONFIDENTLY CALLS SMALL VARIANTS IN HIFI READS OUTSIDE OF THE GIAB HIGH CONFIDENCE REGION

Expanding the Genome in a Bottle benchmark callsets with high-confidence small variant calls from long and linked read sequencing technologies

Justin Wagner[1], Nathan D. Olson[1], Lesley M. Chapman[1], Marc Salit[1,2,3], Justin M. Zook[1], and the Genome in a Bottle Consortium
1: Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD 20899; 2: Joint Initiative for Metrology in Biology, Stanford, CA 94305, USA; 3: Department of Bioengineering, Stanford University, Stanford, CA 94305

National Institute of Standards and Technology
U.S. Department of Commerce

Genome in a Bottle Consortium

Overview

NIST hosts the Genome in a Bottle Consortium, which develops metrology infrastructure for characterization of human whole genome variant detection. Consortium products include:

- Characterization of seven broadly-consented human genomes including 2 son-mother-father trios released as Reference Materials (RMs)
- Reference data associated with RMs are benchmark variants and genomic regions covering, for example, 87.84% of assembled bases in chromosomes 1-22 in GRCh37 for the sample HG002
- Short read variant callers perform poorly in genomic locations with high homology such as segmental duplications and low-complexity repeat-rich regions
- Now utilizing PacBio long read data and 10X Genomics linked reads to expand the GIAB benchmark regions and reduce errors in current regions
- Initial results suggest linked and long reads might be able to add 139,480 benchmark SNPs and 16,081 insertions/deletions, mostly in regions difficult to map with short reads

Integration Pipeline Process

- Find sensitive variant calls and callable regions for each dataset, excluding difficult regions/SVs that are problematic for each type of data and variant caller
- Find "consensus" calls with support from 2+ technologies (and no other technologies disagree) using callable regions
- Use "consensus" calls to train simple one-class model for each dataset and find "outliers" that are less trustworthy for each dataset
- Find benchmark calls by using callable regions and "outliers" to arbitrate between datasets when they disagree
- Find benchmark regions by taking union of callable regions and subtracting uncertain variants

Integration data for HG002 with GRCh37

In addition to the Genome in a Bottle v3.3.2 input data that consisted of Illumina, Complete Genomics, Ion, 10X, and Solid technologies v4α includes PacBio CCS and new 10X linked read data.

Platform	Characteristics	Alignment; Variant Calling
PacBio	~15Kbp reads; ~28x coverage	minimap2; GATK4
PacBio	~15Kbp reads; ~28x coverage	minimap2; DeepVariant
10X	Linked reads; ~84x coverage	LongRanger Pipeline

Difficult Region Description	Method Excluded From
All candidate structural variant regions from the Son-Mother-Father Trio	All methods
All tandem repeats < 51bp in length	All methods except GATK from Illumina PCR-free and Complete Genomics
All tandem repeats > 51bp and < 200bp in length	All methods except GATK from Illumina PCR-free
All tandem repeats > 200bp in length	All methods
Perfect or imperfect homopolymers > 10bp	All methods except GATK from Illumina PCR-free
Segmental duplications from Eichler *et al.*	**All methods except 10X Genomics linked reads and PacBio CCS**
Segmental duplications > 10Kbp from self-chain mapping	**All methods except 10X Genomics linked reads and PacBio CCS**
Regions homologous to contigs in hs37d5 decoy	**All methods except 10X Genomics linked reads and PacBio CCS**

Results from Adding Long and Linked Reads

Benchmark includes more bases, variants, and segmental duplications in v4α

	v3.3.2	v4α	Increase in v4α
Number of bases covered	2,358,060,765	2,442,494,334	84,433,569
Percent of GRCh37 covered	87.84%	90.98%	3.14%
SNPs	3,046,933	3,186,536	139,603
Indels	465,670	482,172	16,502
Number of bases covered in Segmental Duplications	269,887	269,589,673	269,319,786

Comparison of Illumina GATK4 VCF against benchmark sets

- SNP FN rate increases by a factor of 10, almost entirely due to new benchmark variants in difficult to map regions (lowmap) and segmental duplications (segdups)

Subset	v3.3.2 Recall	v4α Recall	v3.3.2 Precision	v4α Precision
All SNPs	0.9995	0.9947	0.9981	0.9933
Lowmap 100 bp	0.9799	0.8464	0.9623	0.8717
Lowmap 250 bp no mismatch	0.9474	0.5522	0.8911	0.7180
Segdups	0.9982	0.9321	0.9910	0.9085

Performance in medically-relevant genes

- Top 5 genes with variants increased from v3.3.2 to v4α benchmark: TSPEAR (37), TNXB (22), CYP21A2 (9), KANSL1 (9), SDHA (8)
- PMS2 from ACMG59 has 2 more variants covered in v4α benchmark

	Variants in Medical Exome (genes from OMIM, HGMD, ClinVar, UniProt)
Benchmark Regions v3.3.2	8,209
Benchmark Regions v4α	8,627

Variant Added in v4α

v3.3.2 Error Excluded in v4α

Ongoing and Future work

- Machine learning
  - Multi-view classification, outlier detection, active learning
- Refine use of genome stratifications
- Adding variant calls from raw PacBio and Oxford Nanopore
- Improve benchmark for larger indels, homopolymers, and tandem repeats
- Explore graph-based methods to characterize MHC region
- Improve normalization of complex variants

Genome in a Bottle Consortium

New members welcome! Sign up for newsletters at www.genomeinabottle.org
Recruiting members to test v4α benchmark please email: justin.zook@nist.gov

- Expands the HG002 small variant high confidence region by >84 Mb (~4%)
- Expands high confidence coverage of segmental duplications by 100-fold
- Adds an additional ~156,000 variants to the benchmark set
- Increases variants in "medically relevant exome" by 5%

CSH Cold Spring Harbor Laboratory

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

Highly-accurate long-read sequencing improves variant detection and assembly of a human genome

doi: https://doi.org/10.1101/519025

Abstract

The major DNA sequencing

https://www.slideshare.net/GenomeInABottle/giab-agbt-smallvar2019

LONG RANGE INFORMATION CAN BE USED TO PHASE SMALL VARIANTS

- WhatsHap phases small variants using long-range information.

Marcel Martin, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schoenhuth, Tobias Marschall. *WhatsHap: fast and accurate read-based phasing.* bioRxiv 085050, doi: 10.1101/085050

LONG RANGE INFORMATION CAN BE USED TO PHASE SMALL VARIANTS

- WhatsHap phases small variants using long-range information.

- PacBio HiFi reads can be used both to generate small variant calls **and** to provide long-range phasing information.

Marcel Martin, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schoenhuth, Tobias Marschall. *WhatsHap: fast and accurate read-based phasing.* bioRxiv 085050, doi: 10.1101/085050

LONG RANGE INFORMATION CAN BE USED TO PHASE SMALL VARIANTS

- WhatsHap phases small variants using long-range information.

- PacBio HiFi reads can be used both to generate small variant calls **and** to provide long-range phasing information.

- Phase block size is driven by:
    - insert length
    - heterozygosity

Marcel Martin, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schoenhuth, Tobias Marschall. *WhatsHap: fast and accurate read-based phasing.* bioRxiv 085050, doi: 10.1101/085050

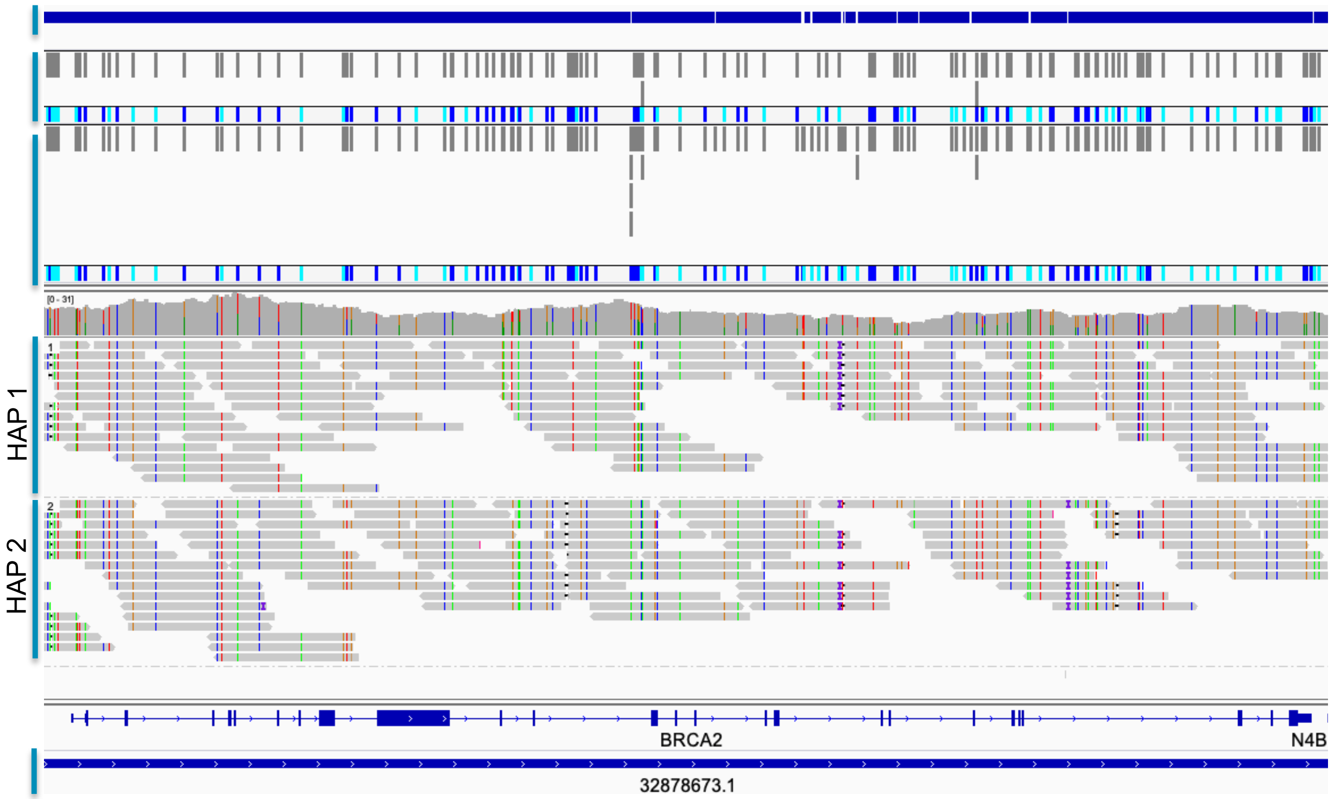
LONG RANGE INFORMATION CAN BE USED TO PHASE SMALL VARIANTS

- WhatsHap phases small variants using long-range information.
- PacBio HiFi reads can be used both to generate small variant calls **and** to provide long-range phasing information.
- Phase block size is driven by:
  - insert length
  - heterozygosity

	autosomal phase blocks			
	mean	**median**	**N50**	**sum**
3 SMRT Cells 8M HG002 (16-fold, 11kb)	76 kb	20 kb	94 kb	1.8 Gb

Marcel Martin, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schoenhuth, Tobias Marschall. *WhatsHap: fast and accurate read-based phasing.* bioRxiv 085050, doi: 10.1101/085050

HIGHLY CONCORDANT IN GIAB HIGH CONFIDENCE REGION



GIAB High Confidence

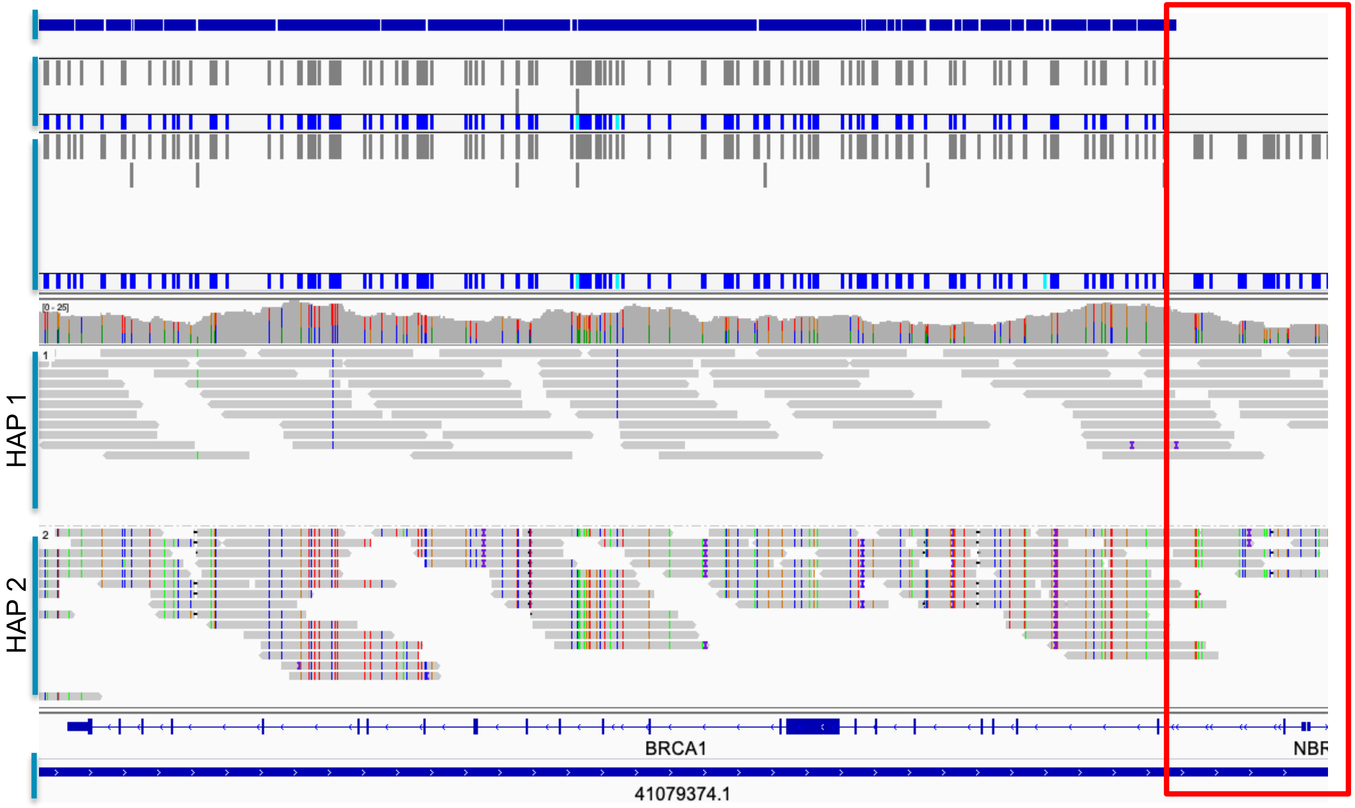
GIAB variants

DeepVariant
small variants

11 kb HiFi

WhatsHap
phase blocks

DETECT VARIANTS OUTSIDE OF HIGH CONFIDENCE REGION



GIAB High Confidence

GIAB variants

DeepVariant
small variants

11 kb HiFi

WhatsHap
phase blocks

DETECT VARIANTS AND PHASE ACROSS DIFFICULT REGIONS

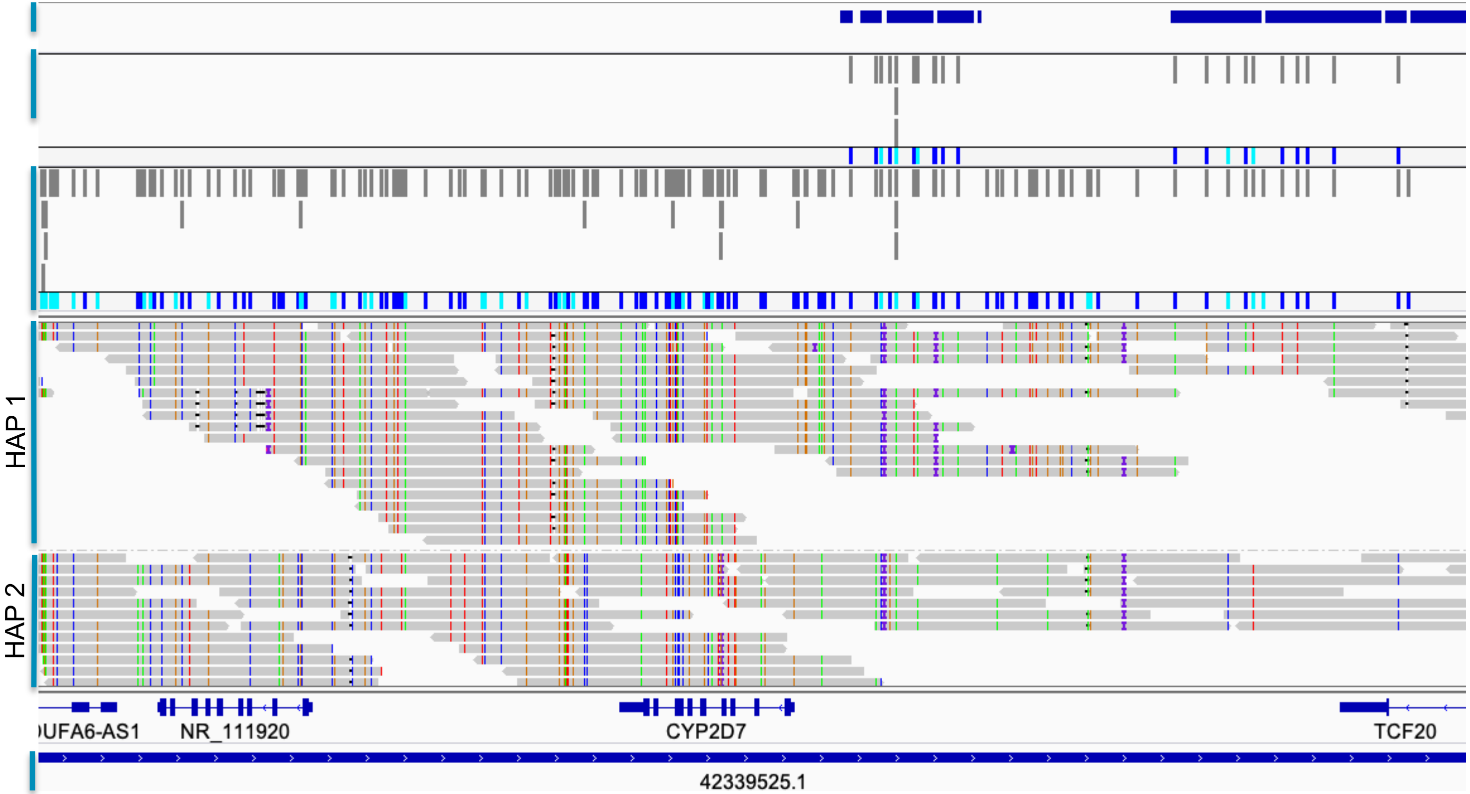
GIAB High Confidence

GIAB variants

DeepVariant
small variants

11 kb HiFi

WhatsHap
phase blocks



De Novo Assembly

CCS ASSEMBLIES ARE HIGHLY CONCORDANT

Cold Spring Harbor Laboratory

bioRχiv

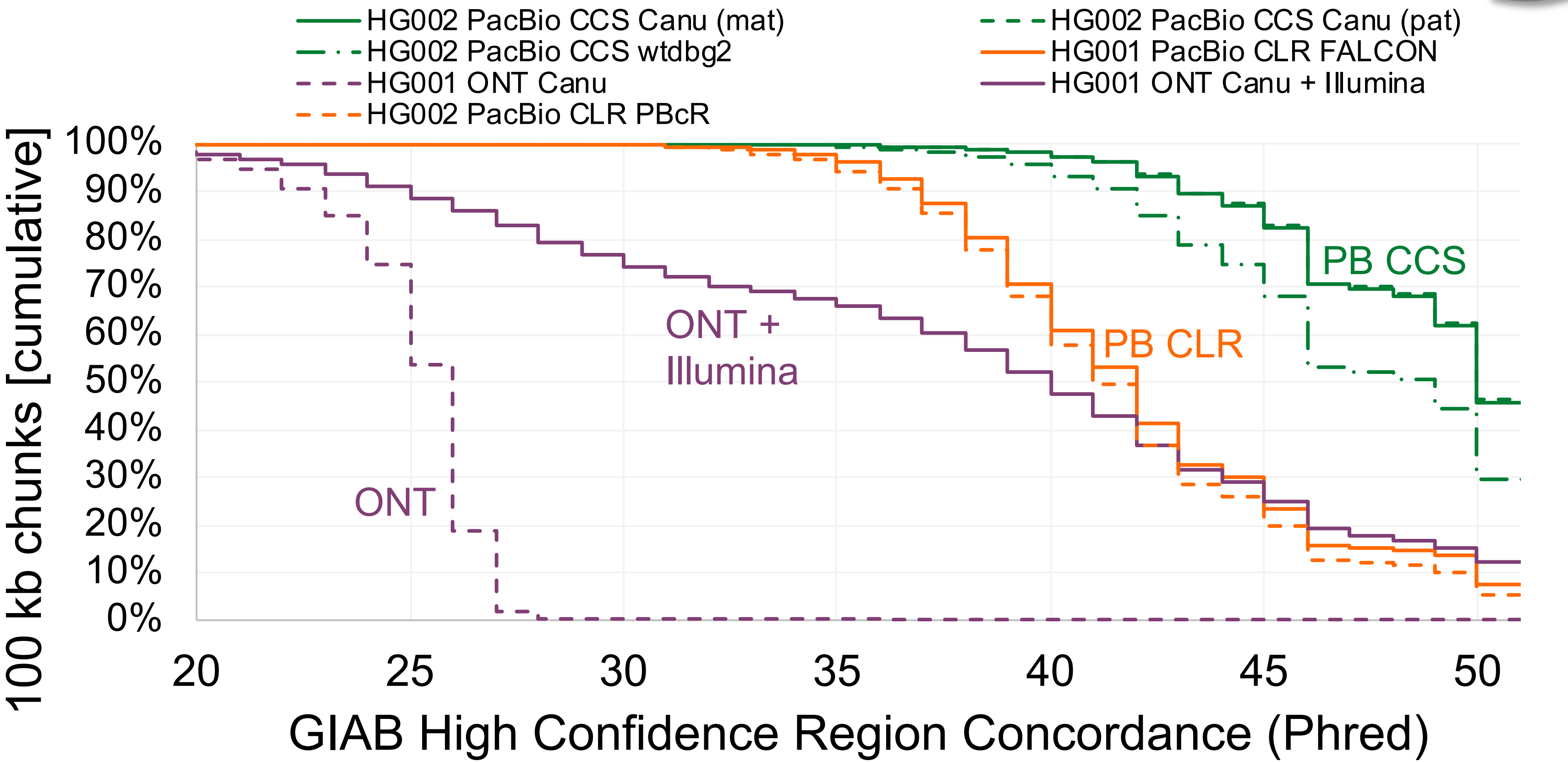
THE PREPRINT SERVER FOR BIOLOGY

Highly-accurate long-read sequencing improves variant detection and assembly of a human genome

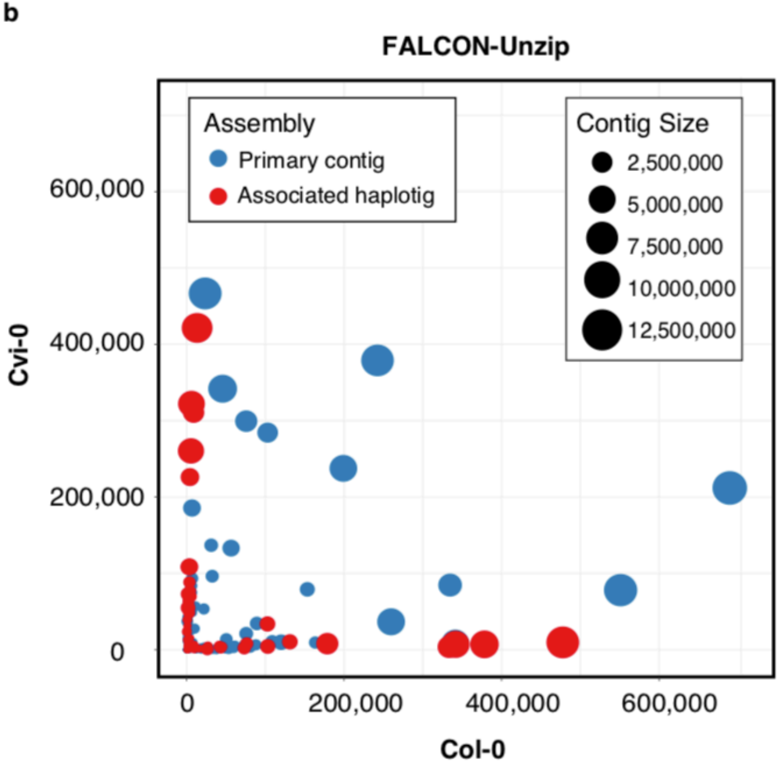
doi: https://doi.org/10.1101/519025

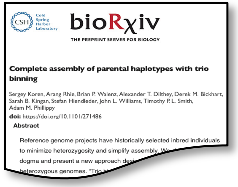
Abstract

The major DNA sequencing



TRIO INFORMATION CAN BE USED TO UNZIP ASSEMBLIES




Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

Complete assembly of parental haplotypes with trio binning

Sergey Koren, Arang Rhie, Brian P. Walenz, Alexander T. Dilthey, Derek M. Bickhart, Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P. L. Smith, Adam M. Phillippy

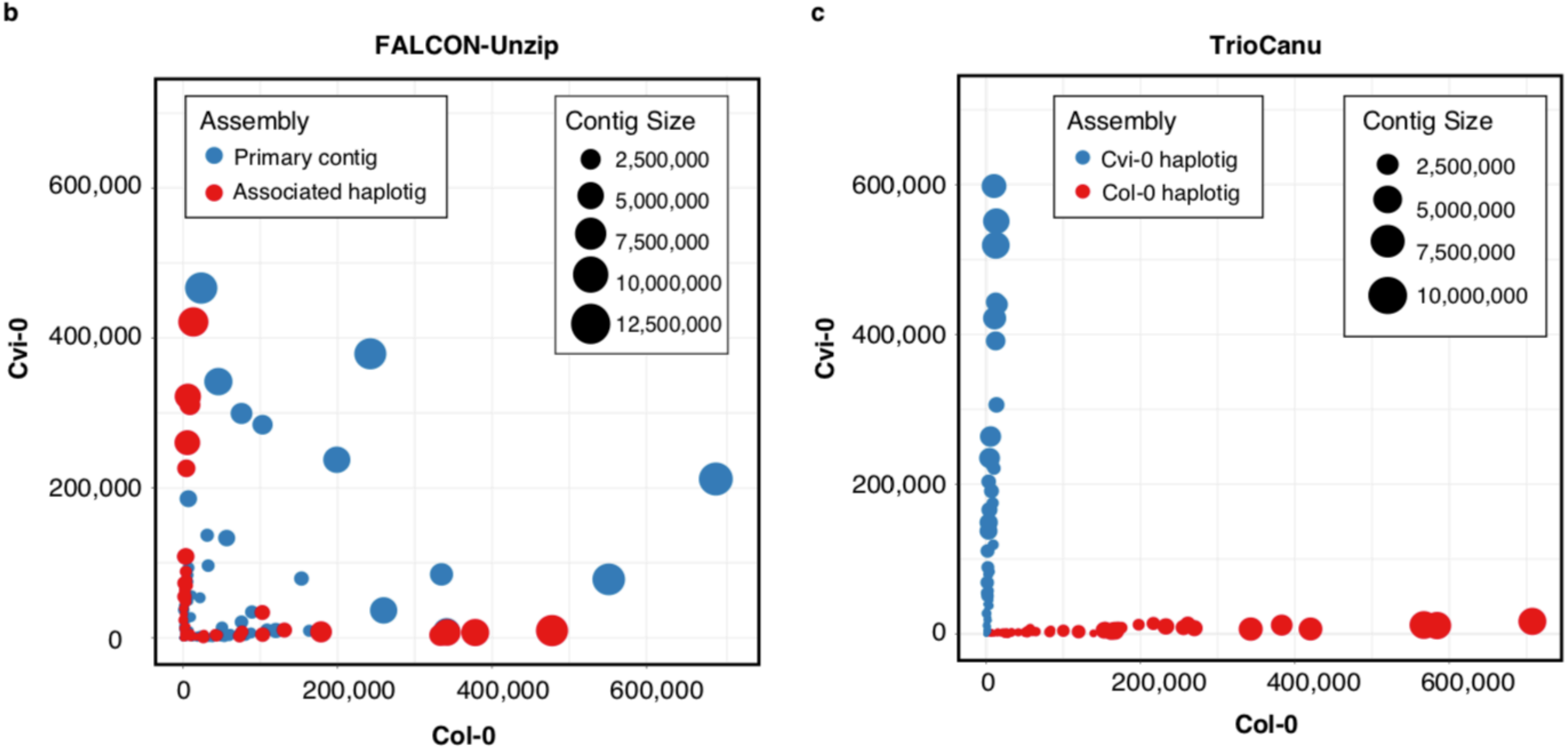
doi: https://doi.org/10.1101/271486

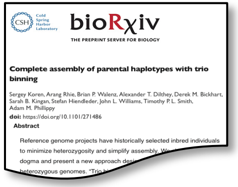
Abstract

Reference genome projects have historically selected inbred individuals to minimize heterozygosity and simplify assembly. We
dogma and present a new approach desi
heterozygous genomes. "Trio

Sergey Koren, Arang Rhie, Brian P. Walenz, Alexander T. Dilthey, Derek M. Bickhart, Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P. L. Smith, Adam M.Phillippy. Complete assembly of parental haplotypes with trio binning. bioRxiv 271486; doi: https://doi.org/10.1101/271486

TRIO INFORMATION CAN BE USED TO UNZIP ASSEMBLIES




CSH Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

Complete assembly of parental haplotypes with trio binning

Sergey Koren, Arang Rhie, Brian P. Walenz, Alexander T. Dilthey, Derek M. Bickhart, Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P. L. Smith, Adam M. Phillippy

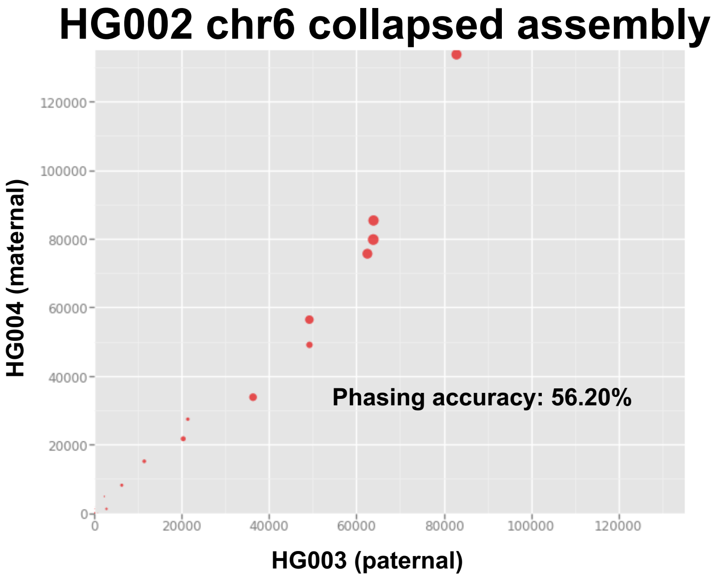
doi: https://doi.org/10.1101/271486

Abstract

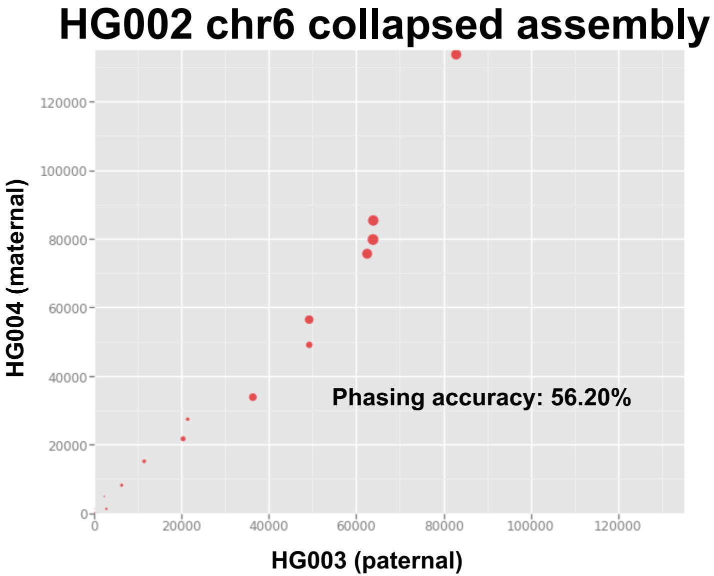
Reference genome projects have historically selected inbred individuals to minimize heterozygosity and simplify assembly. We
dogma and present a new approach desi
heterozygous genomes. “Trio

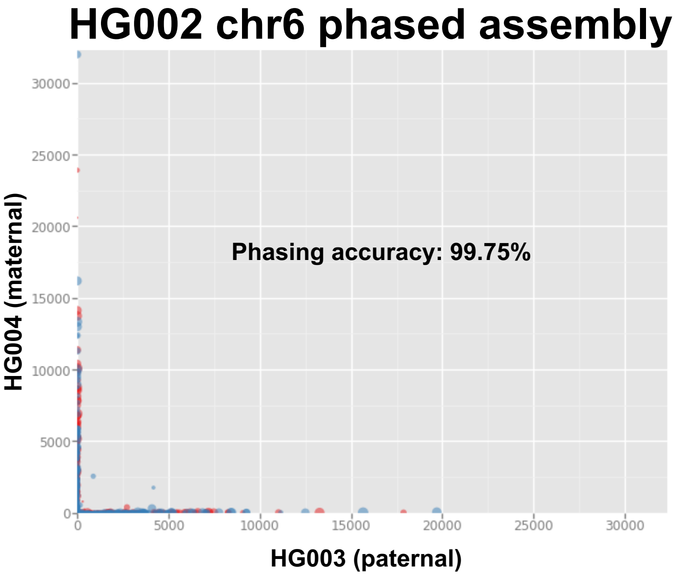
Sergey Koren, Arang Rhie, Brian P. Walenz, Alexander T. Dilthey, Derek M. Bickhart, Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P. L. Smith, Adam M.Phillippy. Complete assembly of parental haplotypes with trio binning. bioRxiv 271486; doi: https://doi.org/10.1101/271486

HIFI ASSEMBLIES CAN BE PHASED WITHOUT PARENTAL DATA

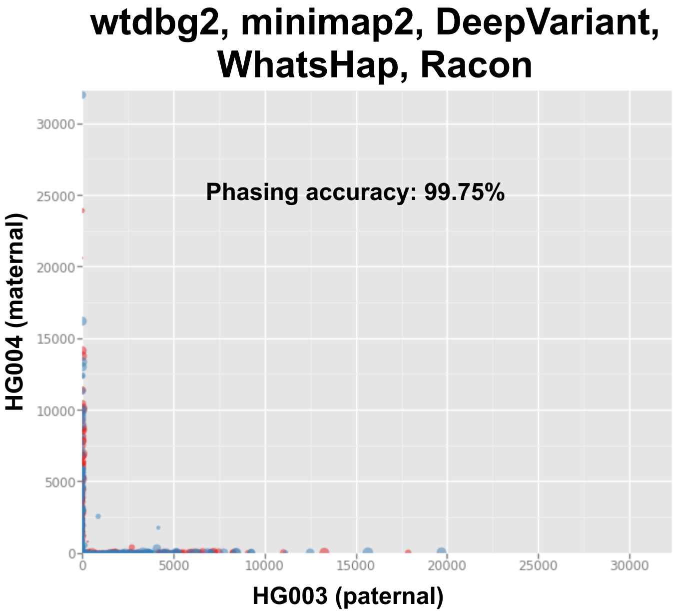


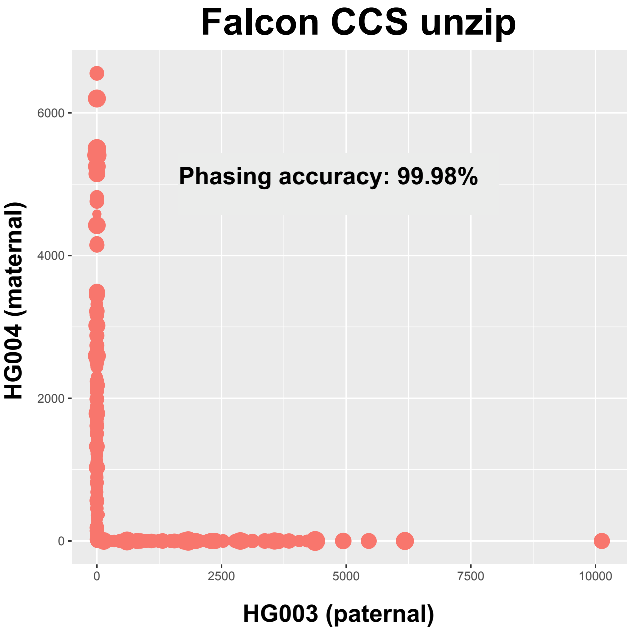
HIFI ASSEMBLIES CAN BE PHASED WITHOUT PARENTAL DATA





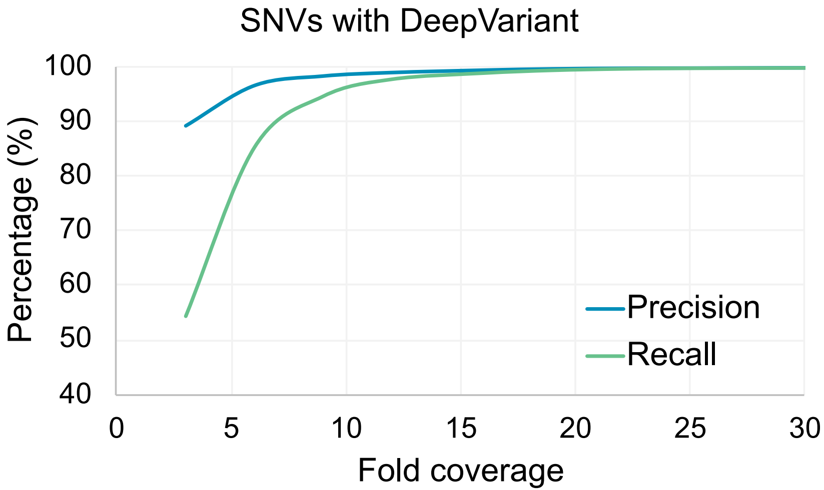
WORKING ON TWO APPROACHES TO HIFI ASSEMBLY PHASING

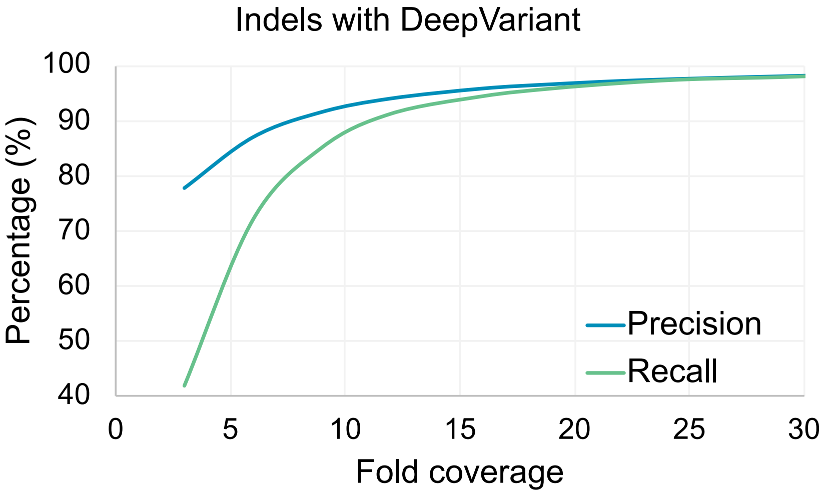




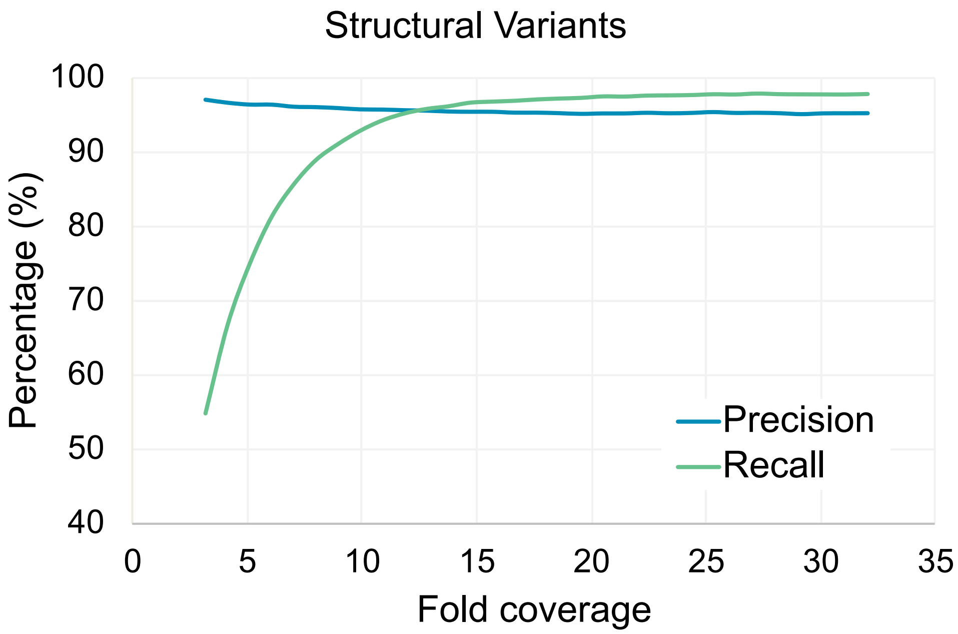
Coverage Recommendations

VARIANT DETECTION COVERAGE TITRATION FOR HG002 ON SEQUEL II SYSTEM

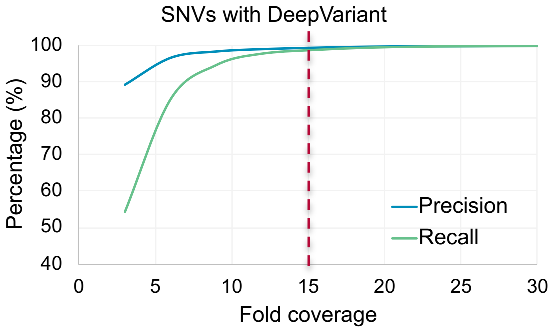


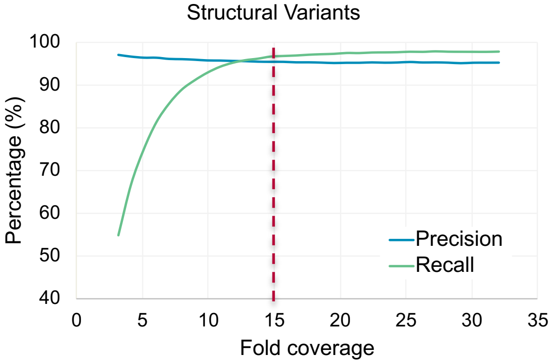


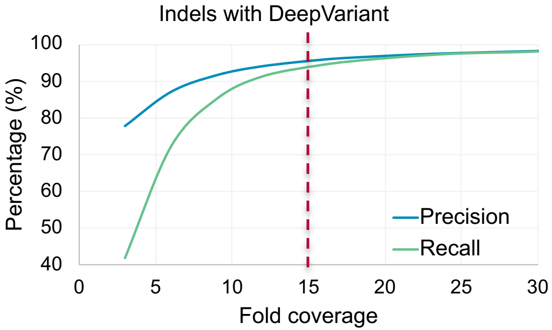
VARIANT DETECTION COVERAGE TITRATION FOR HG002 ON SEQUEL II SYSTEM



VARIANT DETECTION COVERAGE TITRATION FOR HG002 ON SEQUEL II SYSTEM

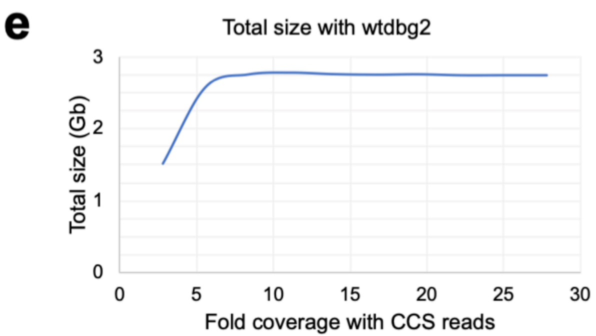


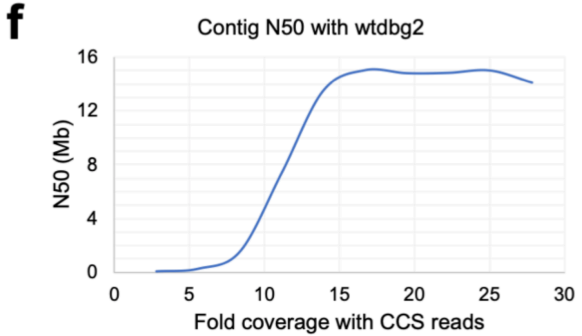


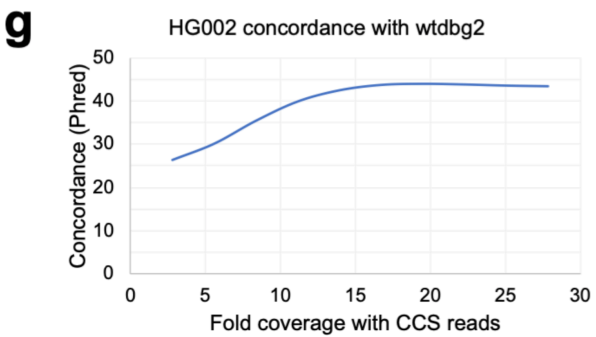


15-fold HiFi Coverage (2-3 SMRT Cells 8M) provides a good trade-off between costs and results

DE NOVO HUMAN ASSEMBLY COVERAGE TITRATION FOR HIFI READS







Highly-accurate long-read sequencing improves variant detection and assembly of a human genome

doi: https://doi.org/10.1101/519025

Abstract

The major DNA sequencing tec

PACBIO®

Public HiFi datasets

for Sequel II System

WE HAVE MANY GIAB DATASETS AVAILABLE FOR TESTING

- HG002/NA24385, 11 kb fraction, 15-fold coverage (3 SMRT Cells):
  - reads, alignments, analysis: https://downloads.pacbcloud.com/public/dataset/HG002_SV_and_SNV_CCS/

- HG002/NA24385 Ashkenazi son, 11 kb fraction, ~30-fold coverage (6 SMRT Cells)
  - reads: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA527278
  - alignments: ftp://ftp.ncbi.nlm.nih.gov//giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_SequelII_CCS_11kb

- HG001/NA12878 CEU female, ~30-fold coverage (6 SMRT Cells)
  - reads: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA540705
  - alignments: ftp://ftp.ncbi.nlm.nih.gov//giab/ftp/data/NA12878/PacBio_SequelII_CCS_11kb
- HG005/NA24631 Han Chinese son, ~30-fold coverage (6 SMRT Cells)
  - reads: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA540706
  - alignments: ftp://ftp.ncbi.nlm.nih.gov//giab/ftp/data/ChineseTrio/HG005_NA24631_son/PacBio_SequelII_CCS_11kb

SUMMARY

- With a single data type, PacBio HiFi reads, you can accurately call small variants and structural variants over >90% of the human genome.
- CCS assemblies are highly concordant and can be highly phased without parental data.

10.1101/519025

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

Baylor – Medhat Mahmoud, Fritz Sedlazeck

Dana-Farber – Heng Li

Chinese Academy of Agricultural Sciences – Jue Ruan

DNAnexus – Chen-Shan Chin, Arkarachai Fungtammasan

Google – Andrew Carroll, Pi-Chuan Chang, Mark DePristo, Alexey Kolesnikov

Johns Hopkins – Michael Alonge, Michael Schatz

Max Planck Dresden – Gene Myers

NIH/NHGRI – Sergey Koren, Adam Phillippy

NIST – Nathan Olson, Justin Zook

PacBio – Greg Concepcion, Richard Hall, Paul Peluso, Yufeng Qian, David Rank, William Rowell, Armin Töpfer, Aaron Wenger, Mike Hunkapiller

Saarland University – Jana Ebler, Tobias Marschall

SUMMARY

- With a single data type, PacBio HiFi reads, you can accurately call small variants and structural variants over >90% of the human genome.
- CCS assemblies are highly concordant and can be highly phased without parental data.

10.1101/519025

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

Baylor – Medhat Mahmoud, Fritz Sedlazeck

Dana-Farber – Heng Li

Chinese Academy of Agricultural Sciences – Jue Ruan

DNAnexus – Chen-Shan Chin, Arkarachai Fungtammasan

Google – Andrew Carroll, Pi-Chuan Chang, Mark DePristo, Alexey Kolesnikov

Johns Hopkins – Michael Alonge, Michael Schatz

Max Planck Dresden – Gene Myers

NIH/NHGRI – Sergey Koren, Adam Phillippy

NIST – Nathan Olson, Justin Zook

PacBio – Greg Concepcion, Richard Hall, Paul Peluso, Yufeng Qian, David Rank, William Rowell, Armin Töpfer, Aaron Wenger, Mike Hunkapiller

Saarland University – Jana Ebler, Tobias Marschall

CCS improvements – Jim Drake, Chris Dunn, David Seifert, Ivan Sovic, Armin Töpfer

CCS Assembly Application Team – Greg Concepcion, Jim Drake, Chris Dunn, Richard Hall, Tzvetana Kerelska, Sarah Kingan, Jonas Korlach, Zev Kronenberg, Ivan Sovic, Michell Vierra, Alicia Yang

7 WORD SUMMARY

Calling all variants with long, accurate reads.

10.1101/519025

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

Baylor – Medhat Mahmoud, Fritz Sedlazeck

Dana-Farber – Heng Li

Chinese Academy of Agricultural Sciences – Jue Ruan

DNAnexus – Chen-Shan Chin, Arkarachai Fungtammasan

Google – Andrew Carroll, Pi-Chuan Chang, Mark DePristo, Alexey Kolesnikov

Johns Hopkins – Michael Alonge, Michael Schatz

Max Planck Dresden – Gene Myers

NIH/NHGRI – Sergey Koren, Adam Phillippy

NIST – Nathan Olson, Justin Zook

PacBio – Greg Concepcion, Richard Hall, Paul Peluso, Yufeng Qian, David Rank, William Rowell, Armin Töpfer, Aaron Wenger, Mike Hunkapiller

Saarland University – Jana Ebler, Tobias Marschall

CCS improvements – Jim Drake, Chris Dunn, David Seifert, Ivan Sovic, Armin Töpfer

CCS Assembly Application Team – Greg Concepcion, Jim Drake, Chris Dunn, Richard Hall, Tzvetana Kerelska, Sarah Kingan, Jonas Korlach, Zev Kronenberg, Ivan Sovic, Michell Vierra, Alicia Yang

PACBIO®

www.pacb.com