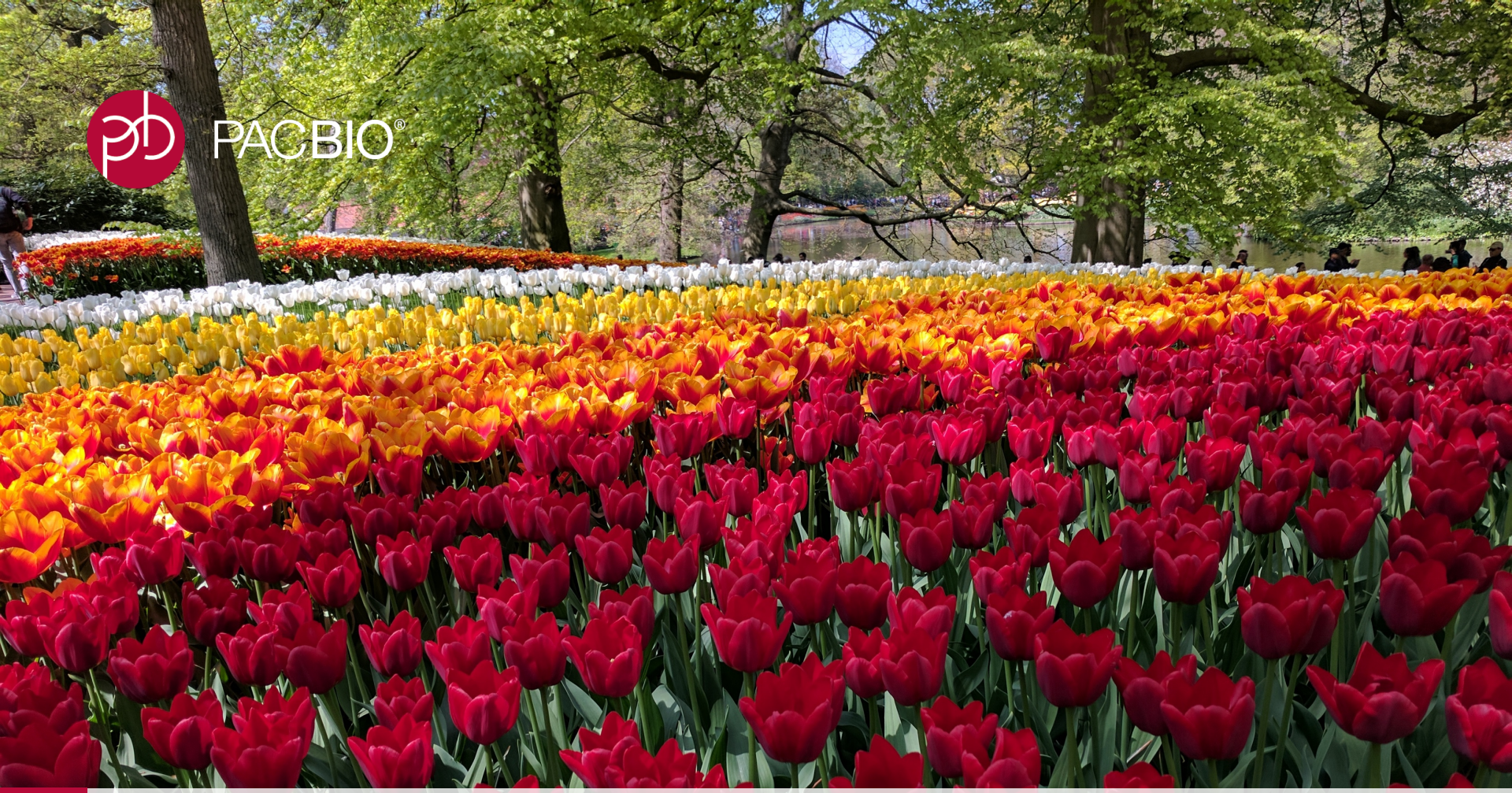


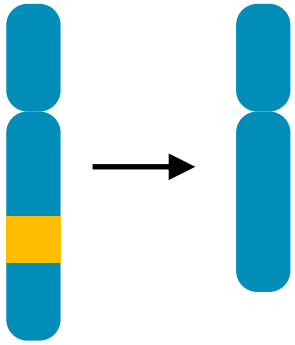
Calling all variants: fast, accurate, population-scale structural variant analysis

Dr. Armin Töpfer

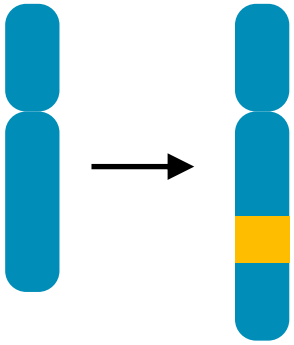
2018-06-13

STRUCTURAL VARIANT = DIFFERENCE ≥50 BP

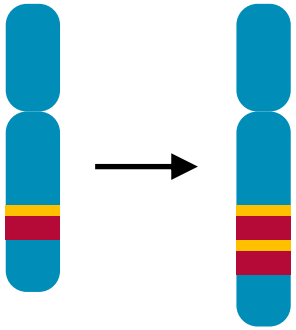
deletion



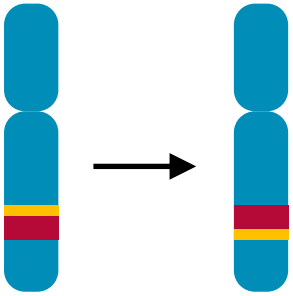
insertion



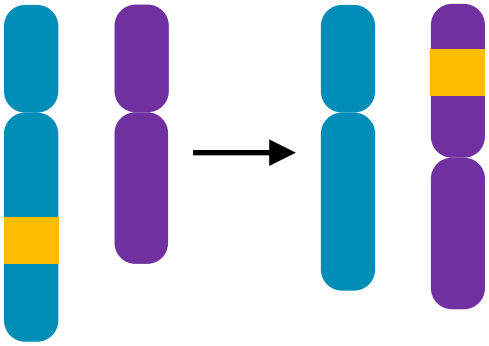
duplication



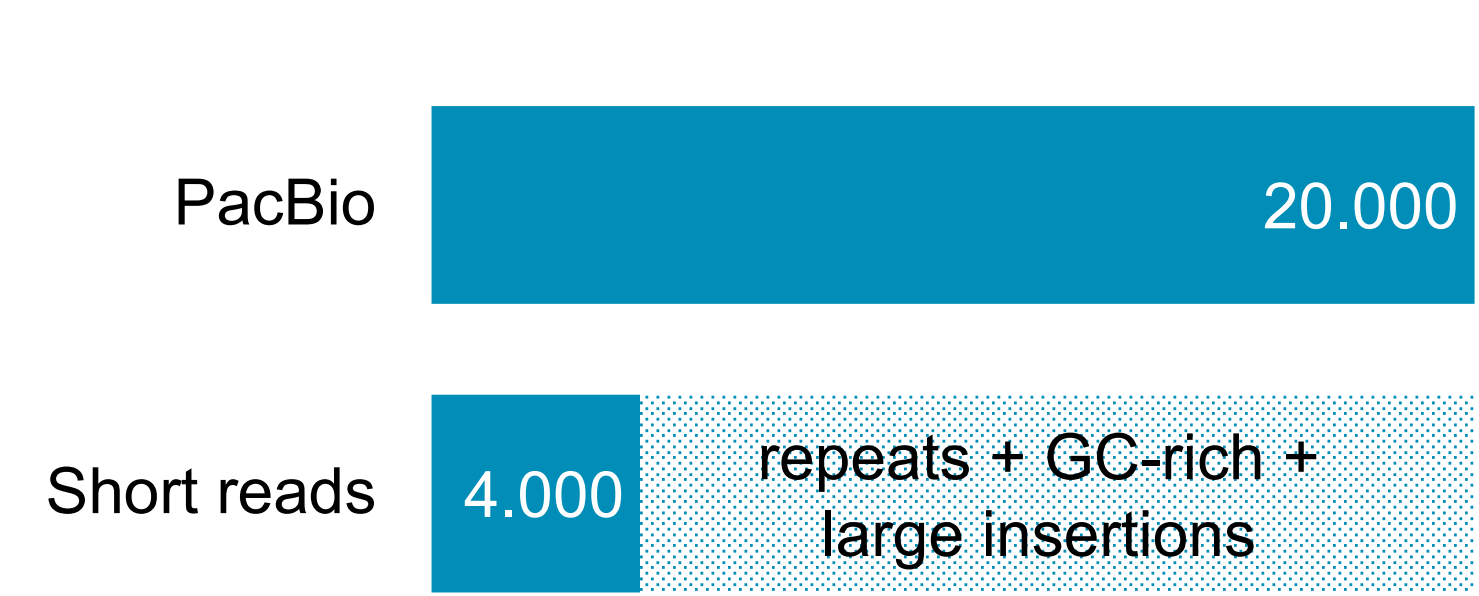
inversion

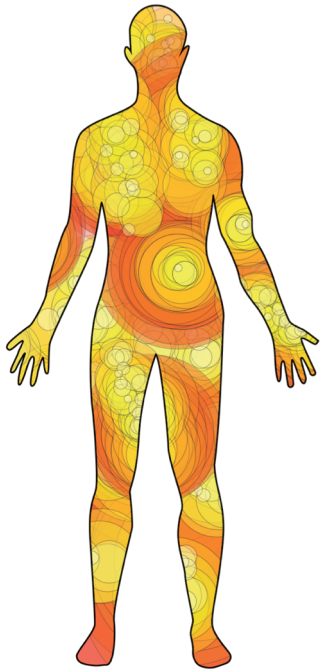


translocation



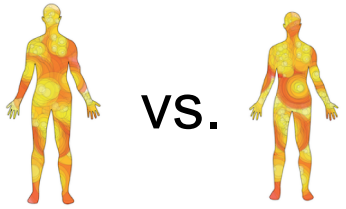
STRUCTURAL VARIANTS DETECTED IN A HUMAN GENOME

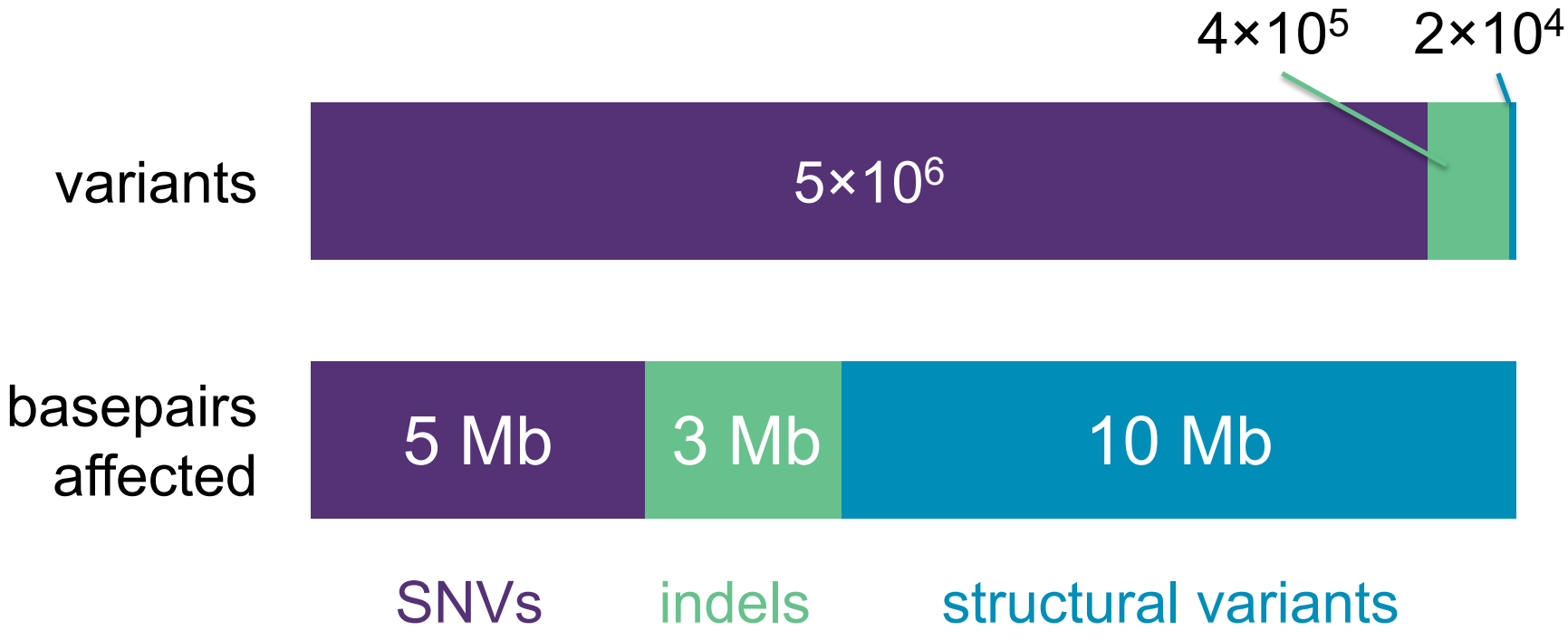




Huddleston et al. (2017) *Genome Research* 27(5):677-85.
Seo et al. (2016) *Nature* 538:243-7.
Sudmant et al. (2016) *Nature* 526:75-81.

VARIATION BETWEEN TWO HUMAN GENOMES





Huddleston et al. (2017) *Genome Research* 27(5):677-85.

PROOF OF CONCEPT – WHOLE HUMAN GENOME

Short reads for small variants

2008

Vol 456 | 6 November 2008 | doi:10.1038/nature07517

nature

ARTICLES

Accurate whole human genome sequencing using reversible terminator chemistry

A list of authors and their affiliations appears at the end of the paper

DNA sequence information underpins genetic research, enabling discoveries of important biological or medical benefit. Sequencing projects have traditionally used long (400–800 base pair) reads, but the existence of reference sequences for the human and many other genomes makes it possible to develop new, fast approaches to re-sequencing, whereby shorter reads are compared to a reference to identify intraspecies genetic variation. Here we report an approach that generates several billion bases of accurate nucleotide sequence per experiment at low cost. Single molecules of DNA are attached to a flat surface, amplified *in situ* and used as templates for synthetic sequencing with fluorescent reversible terminator deoxyribonucleotides. Images of the surface are analysed to generate high-quality sequence. We demonstrate application of this approach to human genome sequencing on flow-sorted X chromosomes and then scale the approach to determine the genome sequence of a male Yoruba from Ibadan, Nigeria. We build an accurate consensus sequence from >30× average depth of paired 35-base reads. We characterize four million single-nucleotide polymorphisms and four hundred thousand structural variants, many of which were previously unknown. Our approach is effective for accurate, rapid and economical whole-genome re-sequencing and many other biomedical applications.

DNA sequencing yields an unrivalled resource of genetic information. We can characterize individual genomes, transcriptional states and genetic variation in populations and disease. Until recently, the scope of sequencing projects was limited by the cost and throughput of Sanger sequencing. The raw data for the three billion base (3 gigabase (Gb)) human genome sequence, completed in 2004 (ref. 1), was generated over several years for ~$300 million using several hundred capillary sequencers. More recently an individual human genome sequence has been determined for ~$10 million by capillary sequencing[2]. Several new approaches at varying stages of development aim to increase sequencing throughput and reduce cost[3–6]. They increase parallelization markedly by imaging many DNA molecules simultaneously. One instrument run produces typically thousands or millions of sequences that are shorter than capillary reads. Another human genome sequence was recently determined using one of these approaches[7]. However, much bigger improvements are necessary to enable routine whole human genome sequencing in genetic research.

We describe a massively parallel synthetic sequencing approach that transforms our ability to use DNA and RNA sequence information in biological systems. We demonstrate utility by re-sequencing an individual human genome to high accuracy. Our approach delivers data at very high throughput and low cost, and enables extraction of genetic information of high biological value, including single-nucleotide polymorphisms (SNPs) and structural variants.

DNA sequencing using reversible terminators

We generated high-density single-molecule arrays of genomic DNA fragments attached to the surface of the reaction chamber (the 'flow cell') and used isothermal 'bridging' amplification to form DNA 'clusters' from each fragment. We made the DNA in each cluster single-stranded and added a universal primer for sequencing. For paired read sequencing, we then converted the templates to double-stranded DNA and removed the original strands, leaving the complementary strand as template for the second sequencing reaction (Fig. 1a–c). To obtain paired reads separated by larger distances, we circularized DNA fragments of the required length (for example, 2 ± 0.2 kb) and obtained short junction fragments for paired end sequencing (Fig. 1d).

We sequenced DNA templates by repeated cycles of polymerase-directed single base extension. To ensure base-by-base nucleotide incorporation in a stepwise manner, we used a set of four reversible terminators, 3′-*O*-azidomethyl 2′-deoxynucleoside triphosphates (A, C, G and T), each labelled with a different removable fluorophore (Supplementary Fig. 1a)[8]. The use of 3′-modified nucleotides allowed the incorporation to be driven essentially to completion without risk of over-incorporation. It also enabled addition of all four nucleotides simultaneously rather than sequentially, minimizing risk of misincorporation. We engineered the active site of 9°N DNA polymerase to improve the efficiency of incorporation of these unnatural nucleotides[9]. After each cycle of incorporation, we determined the identity of the inserted base by laser-induced excitation of the fluorophores and imaging. We added tris (2-carboxyethyl) phosphine (TCEP) to remove the fluorescent dye and side arm from a linker attached to the base and simultaneously regenerate a 3′ hydroxyl group ready for the next cycle of nucleotide addition (Supplementary Fig. 1b). The Genome Analyzer (GA1) was designed to perform multiple cycles of sequencing chemistry and imaging to collect the sequence data automatically from each cluster on the surface of each lane of an eight-lane flow cell (Supplementary Fig. 2).

To determine the sequence from each cluster, we quantified the fluorescent signal from each cycle and applied a base-calling algorithm. We defined a quality (Q) value for each base call (scaled as by the phred algorithm[10]) that represents the likelihood of each call being correct (Supplementary Fig. 3). We used the Q-values in subsequent analyses to weight the contribution of each base to sequence alignment and detection of sequence variants (for example, SNP

PacBio long reads for SVs

2015

LETTER

doi:10.1038/nature13907

Resolving the complexity of the human genome using single-molecule sequencing

Mark J. P. Chaisson[1], John Huddleston[1,2], Megan Y. Dennis[1], Peter H. Sudmant[1], Maika Malig[1], Fereydoun Hormozdiari[1], Francesca Antonacci[1], Urvashi Surti[4], Richard Sandstrom[1], Matthew Boitano[5], Jane M. Landolin[5], John A. Stamatoyannopoulos[1], Michael W. Hunkapiller[5], Jonas Korlach[5] & Evan E. Eichler[1,2]

The human genome is arguably the most complete mammalian reference assembly[1–4], yet more than 160 euchromatic gaps remain[2–4] and aspects of its structural variation remain poorly understood ten years after its completion[2–9]. To identify missing sequence and genetic variation, here we sequence and analyse a haploid human genome (CHM1) using single-molecule, real-time DNA sequencing[10]. We close or extend 55% of the remaining interstitial gaps in the human GRCh37 reference genome—78% of which carried long runs of degenerate short tandem repeats, often several kilobases in length, embedded within (G+C)-rich genomic regions. We resolve the complete sequence of 26,079 euchromatic structural variants at the base-pair level, including inversions, complex insertions and long tracts of tandem repeats. Most have not been previously reported, with the greatest increases in sensitivity occurring for events less than 5 kilobases in size. Compared to the human reference, we find a significant insertional bias (3:1) in regions corresponding to complex insertions and long short tandem repeats. Our results suggest a greater complexity of the human genome in the form of variation of longer and more complex repetitive DNA that can now be largely resolved with the application of this longer-read sequencing technology.

Data generated by single-molecule, real-time (SMRT) sequencing technology differ drastically from most sequencing platforms because native DNA is sequenced without cloning or amplification, and read lengths typically exceed 5 kilobases (kb). Despite overall lower individual read accuracy (~85%), longer read length facilitates high confidence mapping across a greater percentage of the genome[11,12]. We generated ~40-fold sequence coverage from a human CHM1 hydatidiform mole using long-read SMRT sequence technology (average mapped read length = 5.8 kb; Supplementary Table 1). We selected a complete hydatidiform mole to sequence because it is haploid, lacking allelic variation, and provides higher effective sequence coverage. We aligned 93.8% of all sequence reads to the human reference genome (GRCh37) using a modified version of BLASR[11] (Supplementary Information) and generated local assemblies of the mapped reads using Celera[13] and Quiver[14], the latter of which leverages estimates of insertion, deletion and substitution probabilities to determine consensus sequences accurately. We compared the consensus sequences of regions with previously sequenced and assembled large-insert bacterial artificial chromosome (BAC) clones generated from CHM1 tert (ref. 15). The comparison shows a consensus sequencing concordance of >99.97% (phred quality = 37.5), with 72% of the errors confined to indels within homopolymer stretches (Supplementary Table 3).

We initially assessed whether the mapped reads could facilitate closure of any of the 164 interstitial euchromatic gaps within the human reference genome (GRCh37). We extended into gap regions using a reiterative map-and-assemble strategy, in which SMRT whole-genome sequencing (WGS) reads mapping to each edge of a gap were assembled into a new high-quality consensus, which, in turn, served as a template for recruiting additional sequence reads for assembly (Supplementary Information). Using this approach, we closed 50 gaps and extended into 40 others (60 bound aries), adding 398 kb and 721 kb of novel sequence to the genome, respectively (Supplementary Table 4). The closed gaps in the human genome were enriched for simple repeats, long tandem repeats, and high (G+C) content (Fig. 1) but also included novel exons (Supplementary Table 20) and putative regulatory sequences based on DNase I hypersensitivity and chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) analysis (Supplementary Information). We identified a significant 15-fold enrichment of short tandem repeats (STRs) when compared to a random sample ($P < 0.00001$) (Fig. 1a). A total of 78% (39 out of 50) of the closed gap sequences were composed of 10% or more of STRs. The STRs were frequently embedded in longer, more complex, tandem arrays of degenerate repeats reaching up to 8,000 bp in length (Extended Data Fig. 1a–c), some of which bore resemblance to sequences known to be toxic to *Escherichia coli*[16]. Because most human reference sequences[17,18] have been derived from clones propagated in *E. coli*, it is perhaps not surprising that the application of a long-read sequence technology to uncloned DNA would resolve such gaps. Moreover, the length and complex degeneracy of these STRs embedded within (G+C)-rich DNA probably thwarted efforts to follow up most of these by PCR amplification and sequencing.

Next, we developed a computational pipeline (Extended Data Fig. 2) to characterize structural variation systematically (structural variation defined here as differences ≥50 bp in length, including deletions, duplications, insertions and inversions[3]). Structural variants were discovered by mapping SMRT sequencing reads to the human reference genome[11]

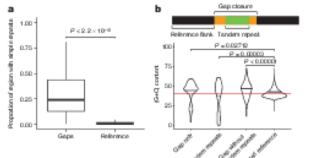


Figure 1 | Sequence content of gap closures. a, Gap closures are enriched for simple repeats compared to equivalently sized regions randomly sampled from GRCh37. **b**, Human genome gaps typically consist of (G+C)-rich sequence (yellow) flanking complex (A+T)-rich STRs (green) (empirical *P* value; Supplementary Information). Red line indicates genomic (G+C) content.

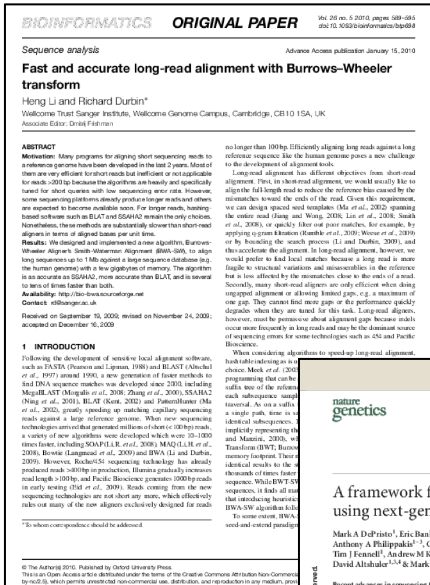
[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA. [2]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. [3]Dipartimento di Biologia, Università degli Studi di Bari 'Aldo Moro', Bari 70126, Italy. [4]Department of Pathology, University of Pittsburgh, Pennsylvania 15261, USA. [5]Pacific Biosciences of California, Inc., Menlo Park, California 94025, USA.

DEVELOPMENT OF ANALYSIS TOOLS

Short reads for small variants

PacBio long reads for SVs

BWA (2010)



BIOINFORMATICS ORIGINAL PAPER

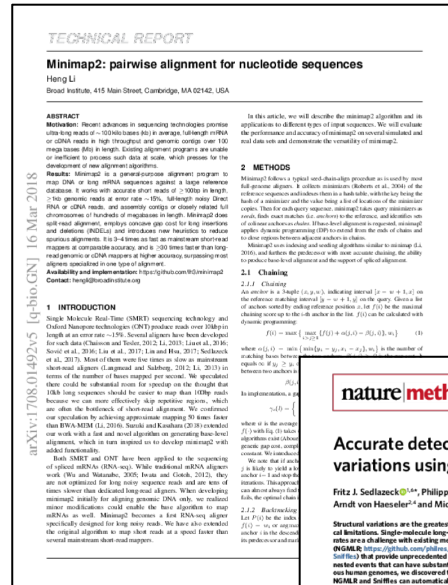
Sequence analysis

Fast and accurate long-read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SA, UK

minimap2 (2017)



TECHNICAL REPORT

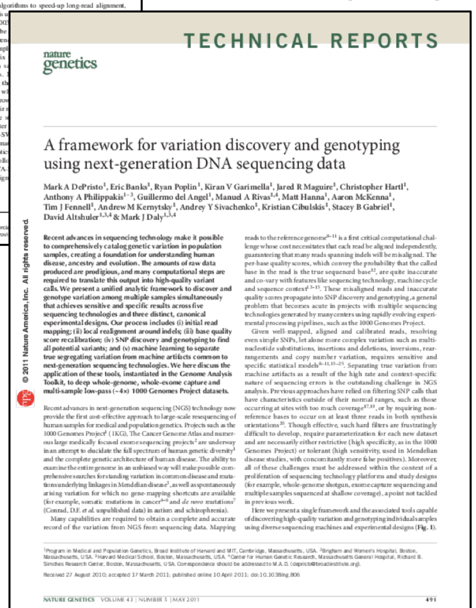
Minimap2: pairwise alignment for nucleotide sequences

Heng Li

Broad Institute, 415 Main Street, Cambridge, MA 02142, USA

arXiv:1708.01492v5 [q-bio.GN] 16 Mar 2018

GATK (2011)

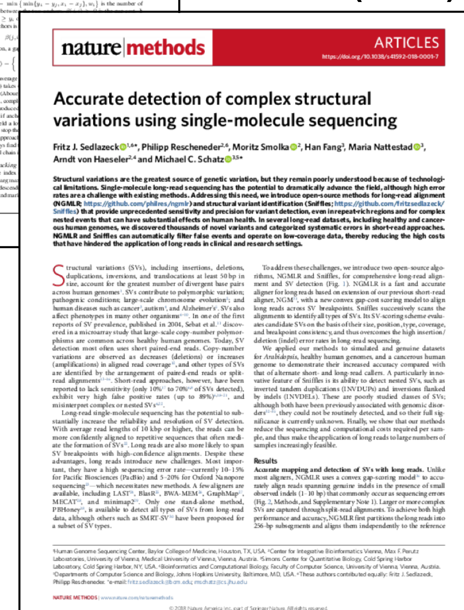


TECHNICAL REPORTS

nature genetics

A framework for variation discovery and genotyping using next-generation DNA sequencing data

Sniffles (2018)



nature methods ARTICLES

Accurate detection of complex structural variations using single-molecule sequencing

RARE DISEASE CASES

Short reads for small variants

2010

BRIEF COMMUNICATIONS

nature genetics

De novo mutations of *SETBP1* cause Schinzel-Giedion syndrome

Alexander Hoischen[1,14], Bregje W M van Bon[1,14], Christian Gilissen[1,14], Peer Arts[1], Bart van Lier[1], Markues Steehouwer[1], Petra de Vries[1], Rick de Reuver[1], Nienke Wieskamp[1], Geert Mortier[2], Koen Devriendt[3], Marta Z Amorim[4], Nicole Revencu[5], Alexa Kidd[6], Mafalda Barbosa[7], Anne Turner[8], Janine Smith[9], Christina Oley[10], Alex Henderson[11], Ian M Hayes[12], Elizabeth M Thompson[13], Han G Brunner[1], Bert B A de Vries[1] & Joris A Veltman[1]

Schinzel-Giedion syndrome is characterized by severe mental retardation, distinctive facial features and multiple congenital malformations; most affected individuals die before the age of ten. We sequenced the exomes of four affected individuals (cases) and found heterozygous *de novo* variants in *SETBP1* in all four. We also identified *SETBP1* mutations in eight additional cases using Sanger sequencing. All mutations clustered to a highly conserved 11-bp exonic region, suggesting a dominant-negative or gain-of-function effect.

Schinzel-Giedion syndrome (MIM%269150) is a highly recognizable syndrome (**Fig. 1a**) characterized by severe mental retardation, distinctive facial features, multiple congenital malformations (including skeletal abnormalities, genitourinary and renal malformations, and cardiac defects) and a higher-than-normal prevalence of tumors, notably neuroepithelial neoplasia[1,2]. In almost all subjects, the disease phenotype occurs sporadically, suggesting heterozygous *de novo* mutations in a single gene as the underlying mechanism. Rare recurrences of this syndrome may be due to gonadal mosaicism. Traditional disease-gene identification approaches have so far failed to identify the gene associated with this disease or those responsible for the majority of this class of rare sporadic disorder. Microarray-based copy number variation screening has been successful for a number of disorders[3], but this method may fail unless the underlying disease mechanism is haploinsufficiency. Recently, whole-exome sequencing was shown to be effective for disease-gene identification[4] and was successfully used to determine the genetic basis of Miller syndrome, a recessive Mendelian disorder[5].

We sequenced the exomes (37 Mb of genomic sequence, targeting ~18,000 genes) of four unrelated individuals with Schinzel-Giedion syndrome to a mean coverage of 43-fold (**Supplementary Table 1**, **Supplementary Figs. 1** and **2**). The exomes of all four individuals were enriched using the SureSelect human exome kit (Agilent) and were subsequently sequenced using one quarter of a SOLiD sequencing slide (Life Technologies). A total of 2.7–3.0 gigabases of mappable sequence data were generated per individual, with 65–72% of bases mapping to the targeted exome (**Supplementary Table 1**). On average, 85% of the exome was covered at least tenfold, and 21,800 genetic variants were identified per individual, including 5,351 nonsynonymous changes. A number of prioritization steps were applied to reduce this number and to identify the potentially pathogenic mutations, similar to the methods used in previous studies[4,5] (**Supplementary Table 2**). A comparison with the NCBI dbSNP build 130 as well as with recently released SNP data from other groups and in-house SNP data (see **Supplementary Note**) showed that >95% of all variants investigated here were previously reported SNPs and cannot explain a genetically dominant disease. We focused on the 12 genes for which all four individuals studied carried variants and found that only two genes showed variants at different genomic positions, strengthening the likelihood that these variants are causative and not simply unidentified SNPs. One of these two candidate genes, *CTBP2*, was excluded from further analysis because it contained numerous variants found during different in-house exome sequencing experiments (data not shown), which may be due to highly homologous sequences from other genomic loci.

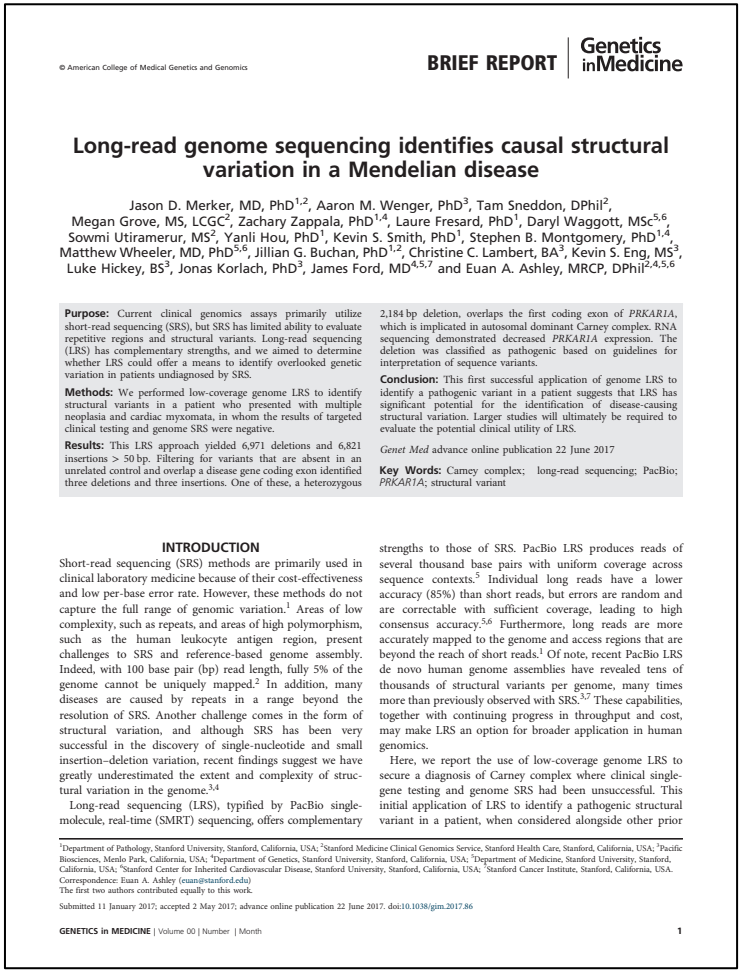
The second candidate was *SETBP1*, which encodes SET binding protein 1. Validation of all four variants in this gene by Sanger sequencing confirmed that these variants were indeed present in a heterozygous state in all four affected individuals (**Supplementary Fig. 3**). Moreover, we tested the DNA of the parents of the affected individuals, which showed that all mutations occurred *de novo*. Using Sanger sequencing, we also identified *SETBP1* mutations in eight out of nine additional individuals with a clinical diagnosis of Schinzel-Giedion syndrome. In total, all 13 affected individuals fulfilled previously suggested diagnostic criteria[2] (**Table 1** and **Supplementary Table 3**); all are of European descent, living in various regions: Europe ($n = 7$), New Zealand ($n = 3$), Australia ($n = 2$) and the United States ($n = 1$). For six of the eight follow-up cases, parental DNA was available, and the mutations present in the affected individuals were again shown to have occurred

[1]Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [2]Centre for Medical Genetics, Antwerp University Hospital, Antwerp, Belgium. [3]Centre for Human Genetics, Leuven University Hospital, Leuven, Belgium. [4]Serviço de Genética de Coimbra, Hospital Pediátrico de Coimbra, Coimbra, Portugal. [5]Centre for Human Genetics, Cliniques Universitaires St. Luc, Université Catholique de Louvain, Brussels, Belgium. [6]Canterbury Health Laboratories, Christchurch Hospital, Christchurch, New Zealand. [7]Centro de Genética Médica Doutor Jacinto Magalhães, Instituto Nacional de Saúde Doutor Ricardo, Porto, Portugal. [8]Department of Medical Genetics, Sydney Children's Hospital, Sydney, Australia. [9]Department of Clinical Genetics, The Children's Hospital at Westmead, Sydney, Australia. [10]Clinical Genetics Unit, Birmingham Women's Hospital, Birmingham, UK. [11]Institute of Human Genetics, Newcastle upon Tyne Hospitals National Health Service Foundation Trust, Newcastle upon Tyne, UK. [12]Northern Regional Genetics Service, Auckland, New Zealand. [13]South Australian Clinical Genetics Service, South Australian Pathology, Women's and Children's Hospital, North Adelaide, South Australia, Australia. [14]These authors contributed equally to this work. Correspondence should be addressed to J.A.V. (j.veltman@antrg.umcn.nl).

Received 9 February; accepted 8 April; published online 2 May 2010; doi:10.1038/ng.581

PacBio long reads for SVs

2017



BRIEF REPORT | Genetics inMedicine

Long-read genome sequencing identifies causal structural variation in a Mendelian disease

Jason D. Merker, MD, PhD[1,2], Aaron M. Wenger, PhD[3], Tam Sneddon, DPhil[2], Megan Grove, MS, LCGC[2], Zachary Zappala, PhD[1,4], Laure Fresard, PhD[1], Daryl Waggott, MSc[5,6], Sowmi Utiramerur, MS[2], Yanli Hou, PhD[1], Kevin S. Smith, PhD[1], Stephen B. Montgomery, PhD[1,4], Matthew Wheeler, MD, PhD[5,6], Jillian G. Buchan, PhD[1,2], Christine C. Lambert, BA[3], Kevin S. Eng, MS[3], Luke Hickey, BS[3], Jonas Korlach, PhD[3], James Ford, MD[4,5,7] and Euan A. Ashley, MRCP, DPhil[2,4,5,6]

Purpose: Current clinical genomics assays primarily utilize short-read sequencing (SRS), but SRS has limited ability to evaluate repetitive regions and structural variants. Long-read sequencing (LRS) has complementary strengths, and we aimed to determine whether LRS could offer a means to identify overlooked genetic variation in patients undiagnosed by SRS.

Methods: We performed low-coverage genome LRS to identify structural variants in a patient who presented with multiple neoplasia and cardiac myxomata, in whom the results of targeted clinical testing and genome SRS were negative.

Results: This LRS approach yielded 6,971 deletions and 6,821 insertions > 50 bp. Filtering for variants that are absent in an unrelated control and overlap a disease gene coding exon identified three deletions and three insertions. One of these, a heterozygous 2,184 bp deletion, overlaps the first coding exon of *PRKAR1A*, which is implicated in autosomal dominant Carney complex. RNA sequencing demonstrated decreased *PRKAR1A* expression. The deletion was classified as pathogenic based on guidelines for interpretation of sequence variants.

Conclusion: This first successful application of genome LRS to identify a pathogenic variant in a patient suggests that LRS has significant potential for the identification of disease-causing structural variation. Larger studies will ultimately be required to evaluate the potential clinical utility of LRS.

Genet Med advance online publication 22 June 2017

Key Words: Carney complex; long-read sequencing; PacBio; *PRKAR1A*; structural variant

INTRODUCTION

Short-read sequencing (SRS) methods are primarily used in clinical laboratory medicine because of their cost-effectiveness and low per-base error rate. However, these methods do not capture the full range of genomic variation.[1] Areas of low complexity, such as repeats, and areas of high polymorphism, such as the human leukocyte antigen region, present challenges to SRS and reference-based genome assembly. Indeed, with 100 base pair (bp) read length, fully 5% of the genome cannot be uniquely mapped.[2] In addition, many diseases are caused by repeats in a range beyond the resolution of SRS. Another challenge comes in the form of structural variation, and although SRS has been very successful in the discovery of single-nucleotide and small insertion–deletion variation, recent findings suggest we have greatly underestimated the extent and complexity of structural variation in the genome.[3,4]

Long-read sequencing (LRS), typified by PacBio single-molecule, real-time (SMRT) sequencing, offers complementary strengths to those of SRS. PacBio LRS produces reads of several thousand base pairs with uniform coverage across sequence contexts.[5] Individual long reads have a lower accuracy (85%) than short reads, but errors are random and are correctable with sufficient coverage, leading to high consensus accuracy.[5,6] Furthermore, long reads are more accurately mapped to the genome and access regions that are beyond the reach of short reads.[1] Of note, recent PacBio LRS de novo human genome assemblies have revealed tens of thousands of structural variants per genome, many times more than previously observed with SRS.[3,7] These capabilities, together with continuing progress in throughput and cost, may make LRS an option for broader application in human genomics.

Here, we report the use of low-coverage genome LRS to secure a diagnosis of Carney complex where clinical single-gene testing and genome SRS had been unsuccessful. This initial application of LRS to identify a pathogenic structural variant in a patient, when considered alongside other prior

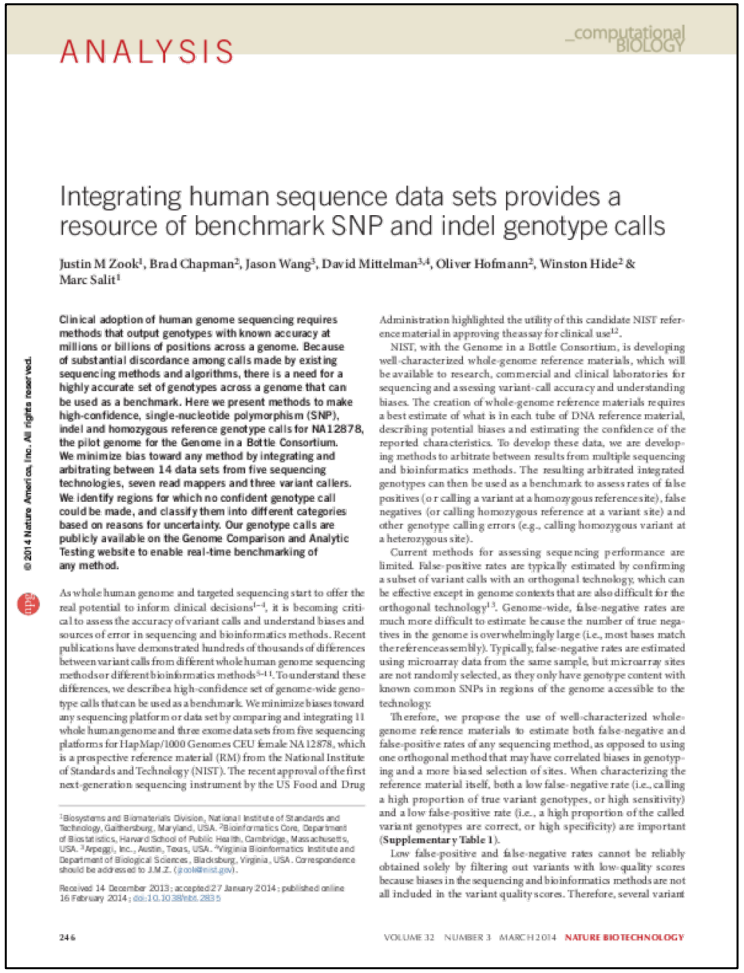
[1]Department of Pathology, Stanford University, Stanford, California, USA; [2]Stanford Medicine Clinical Genomics Service, Stanford Health Care, Stanford, California, USA; [3]Pacific Biosciences, Menlo Park, California, USA; [4]Department of Genetics, Stanford University, Stanford, California, USA; [5]Department of Medicine, Stanford University, Stanford, California, USA; [6]Stanford Center for Inherited Cardiovascular Disease, Stanford University, Stanford, California, USA; [7]Stanford Cancer Institute, Stanford, California, USA. Correspondence: Euan A. Ashley (euan@stanford.edu)
The first two authors contributed equally to this work.

Submitted 11 January 2017; accepted 2 May 2017; advance online publication 22 June 2017. doi:10.1038/gim.2017.86

BENCHMARK STANDARDS

Short reads for small variants

2014



ANALYSIS

_computational BIOLOGY

Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls

Justin M Zook[1], Brad Chapman[2], Jason Wang[3], David Mittelman[3,4], Oliver Hofmann[2], Winston Hide[2] & Marc Salit[1]

Clinical adoption of human genome sequencing requires methods that output genotypes with known accuracy at millions or billions of positions across a genome. Because of substantial discordance among calls made by existing sequencing methods and algorithms, there is a need for a highly accurate set of genotypes across a genome that can be used as a benchmark. Here we present methods to make high-confidence, single-nucleotide polymorphism (SNP), indel and homozygous reference genotype calls for NA12878, the pilot genome for the Genome in a Bottle Consortium. We minimize bias toward any method by integrating and arbitrating between 14 data sets from five sequencing technologies, seven read mappers and three variant callers. We identify regions for which no confident genotype call could be made, and classify them into different categories based on reasons for uncertainty. Our genotype calls are publicly available on the Genome Comparison and Analytic Testing website to enable real-time benchmarking of any method.

As whole human genome and targeted sequencing start to offer the real potential to inform clinical decisions[1–3], it is becoming critical to assess the accuracy of variant calls and understand biases and sources of error in sequencing and bioinformatics methods. Recent publications have demonstrated hundreds of thousands of differences between variant calls from different whole human genome sequencing methods or different bioinformatics methods[5–11]. To understand these differences, we describe a high-confidence set of genome-wide genotype calls that can be used as a benchmark. We minimize biases toward any sequencing platform or data set by comparing and integrating 11 whole human genome and three exome data sets from five sequencing platforms for HapMap/1000 Genomes CEU female NA12878, which is a prospective reference material (RM) from the National Institute of Standards and Technology (NIST). The recent approval of the first next-generation sequencing instrument by the US Food and Drug Administration highlighted the utility of this candidate NIST reference material in approving the assay for clinical use[12].

NIST, with the Genome in a Bottle Consortium, is developing well-characterized whole-genome reference materials, which will be available to research, commercial and clinical laboratories for sequencing and assessing variant-call accuracy and understanding biases. The creation of whole-genome reference materials requires a best estimate of what is in each tube of DNA reference material, describing potential biases and estimating the confidence of the reported characteristics. To develop these data, we are developing methods to arbitrate between results from multiple sequencing and bioinformatics methods. The resulting arbitrated integrated genotypes can then be used as a benchmark to assess rates of false positives (or calling a variant at a homozygous reference site), false negatives (or calling homozygous reference at a variant site) and other genotype calling errors (e.g., calling homozygous variant at a heterozygous site).

Current methods for assessing sequencing performance are limited. False-positive rates are typically estimated by confirming a subset of variant calls with an orthogonal technology, which can be effective except in genome contexts that are also difficult for the orthogonal technology[13]. Genome-wide, false-negative rates are much more difficult to estimate because the number of true negatives in the genome is overwhelmingly large (i.e., most bases match the reference assembly). Typically, false-negative rates are estimated using microarray data from the same sample, but microarray sites are not randomly selected, as they only have genotype content with known common SNPs in regions of the genome accessible to the technology.

Therefore, we propose the use of well-characterized whole-genome reference materials to estimate both false-negative and false-positive rates of any sequencing method, as opposed to using one orthogonal method that may have correlated biases in genotyping and a more biased selection of sites. When characterizing the reference material itself, both a low false-negative rate (i.e., calling a high proportion of true variant genotypes, or high sensitivity) and a low false-positive rate (i.e., a high proportion of the called variant genotypes are correct, or high specificity) are important (**Supplementary Table 1**).

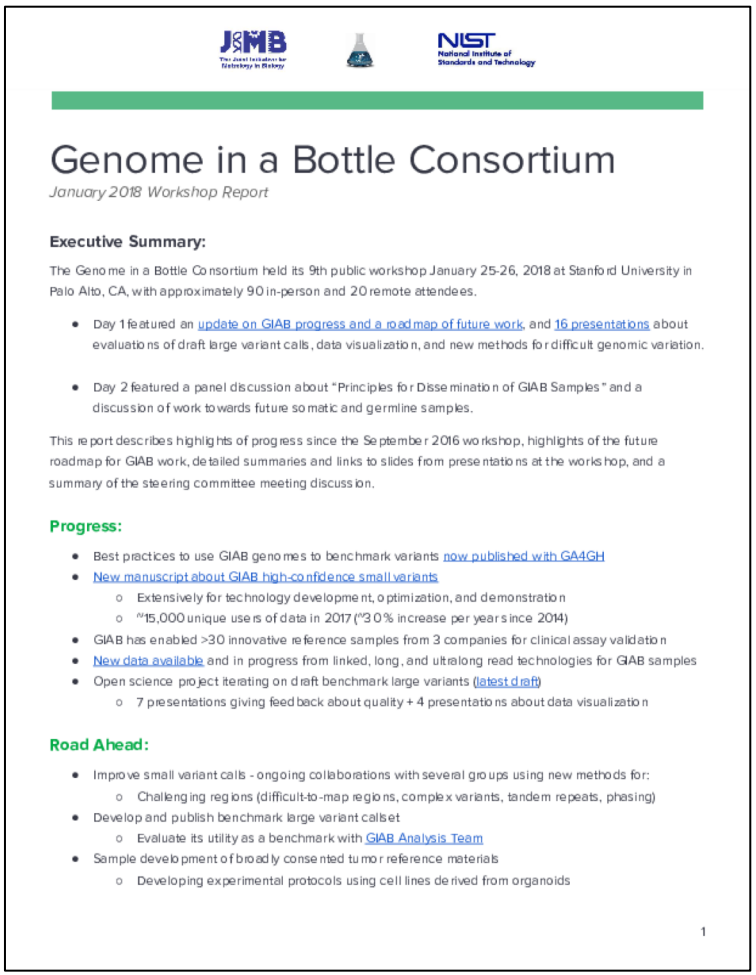
Low false-positive and false-negative rates cannot be reliably obtained solely by filtering out variants with low-quality scores because biases in the sequencing and bioinformatics methods are not all included in the variant quality scores. Therefore, several variant

[1]Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA. [2]Bioinformatics Core, Department of Biostatistics, Harvard School of Public Health, Cambridge, Massachusetts, USA. [3]Arpeggi, Inc., Austin, Texas, USA. [4]Virginia Bioinformatics Institute and Department of Biological Sciences, Blacksburg, Virginia, USA. Correspondence should be addressed to J.M.Z. (jzook@nist.gov).

Received 14 December 2013; accepted 27 January 2014; published online 16 February 2014; doi:10.1038/nbt.2835

PacBio long reads for SVs

in progress



JCVI The J. Craig Venter Institute

NIST National Institute of Standards and Technology

Genome in a Bottle Consortium

January 2018 Workshop Report

Executive Summary:

The Genome in a Bottle Consortium held its 9th public workshop January 25-26, 2018 at Stanford University in Palo Alto, CA, with approximately 90 in-person and 20 remote attendees.

- Day 1 featured an update on GIAB progress and a roadmap of future work, and 16 presentations about evaluations of draft large variant calls, data visualization, and new methods for difficult genomic variation.
- Day 2 featured a panel discussion about "Principles for Dissemination of GIAB Samples" and a discussion of work towards future somatic and germline samples.

This report describes highlights of progress since the September 2016 workshop, highlights of the future roadmap for GIAB work, detailed summaries and links to slides from presentations at the workshop, and a summary of the steering committee meeting discussion.

Progress:

- Best practices to use GIAB genomes to benchmark variants now published with GA4GH
- New manuscript about GIAB high-confidence small variants
  - Extensively for technology development, optimization, and demonstration
  - ~15,000 unique users of data in 2017 (~30% increase per year since 2014)
- GIAB has enabled >30 innovative reference samples from 3 companies for clinical assay validation
- New data available and in progress from linked, long, and ultralong read technologies for GIAB samples
- Open science project iterating on draft benchmark large variants (latest draft)
  - 7 presentations giving feedback about quality + 4 presentations about data visualization

Road Ahead:

- Improve small variant calls - ongoing collaborations with several groups using new methods for:
  - Challenging regions (difficult-to-map regions, complex variants, tandem repeats, phasing)
- Develop and publish benchmark large variant callset
  - Evaluate its utility as a benchmark with GIAB Analysis Team
- Sample development of broadly consented tumor reference materials
  - Developing experimental protocols using cell lines derived from organoids

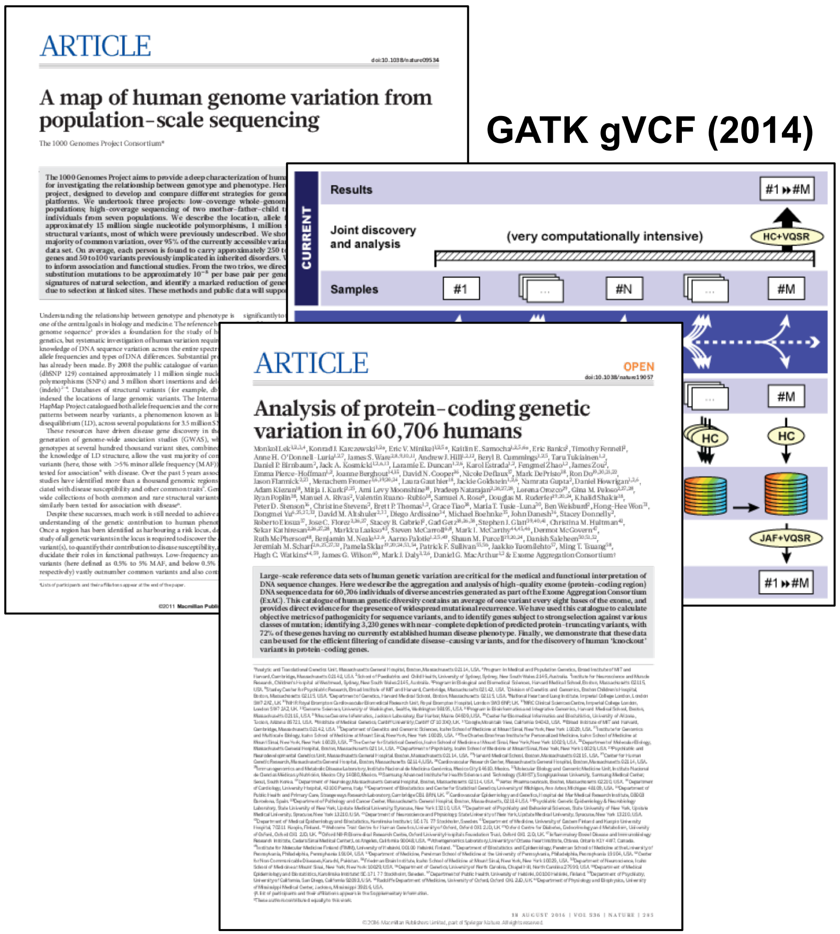
POPULATION STUDIES

Short reads for small variants

PacBio long reads for SVs

1K Genomes Pilot (2010)

GATK gVCF (2014)

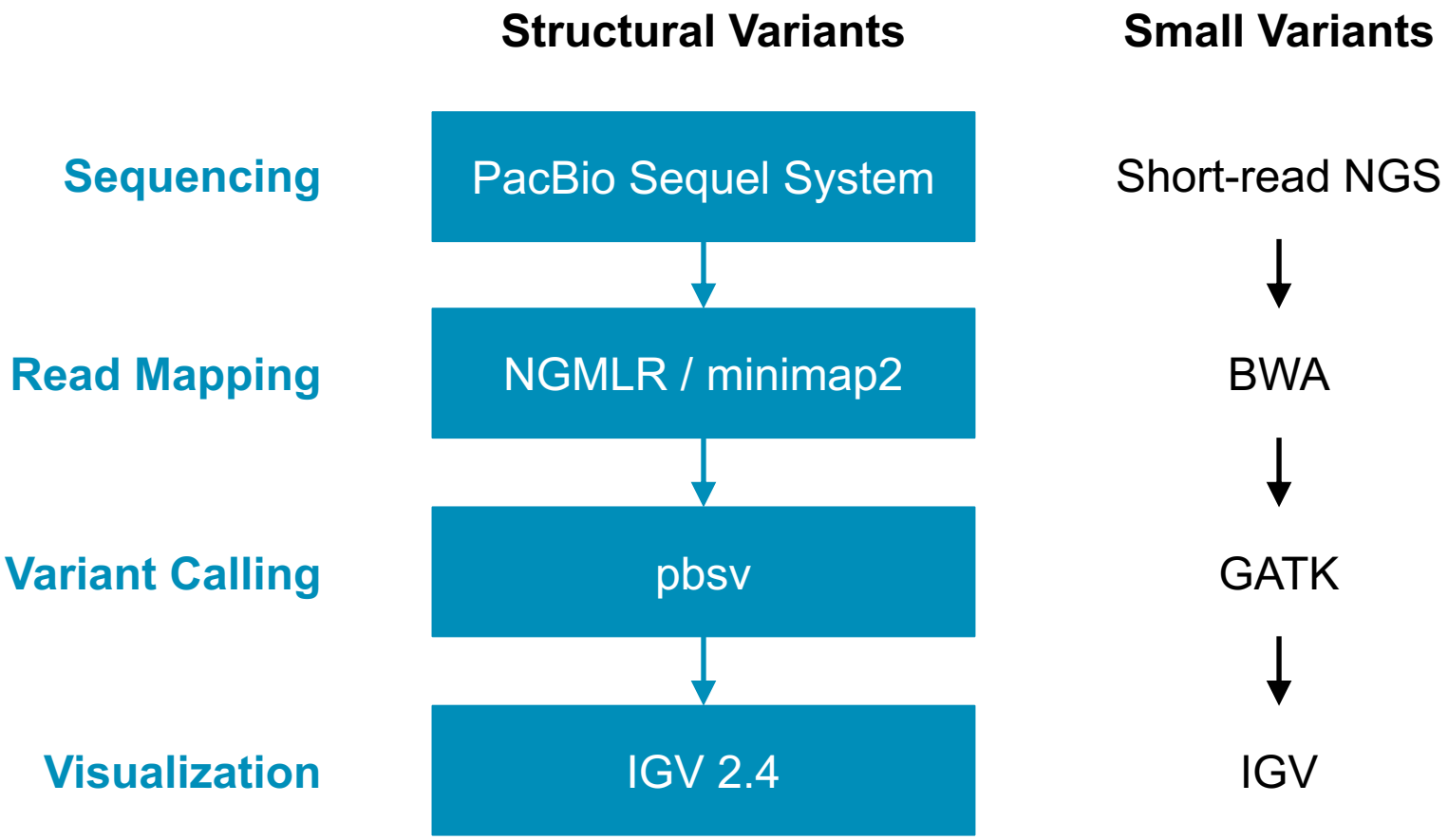


ExAC (2016)

HGSVC (2017)

in progress

PBSV WORKFLOW



PBSV – STRENGTHS AND LIMITATIONS

Strengths

+ Simple to use
+ Low false discovery rate
+ High sensitivity at low coverage
+ Joint calling

Limitations

– Only insertions and deletions
– Approximate breakpoints
– Slow for large cohorts

PBSV 2.0 ADDRESSES LIMITATIONS

Strengths

+ Simple to use
+ Low false discovery rate
+ High sensitivity at low coverage
+ Joint calling
+ Translocations and inversions
+ Indels under 50 bp
+ Polished breakpoints
+ Scalable workflow

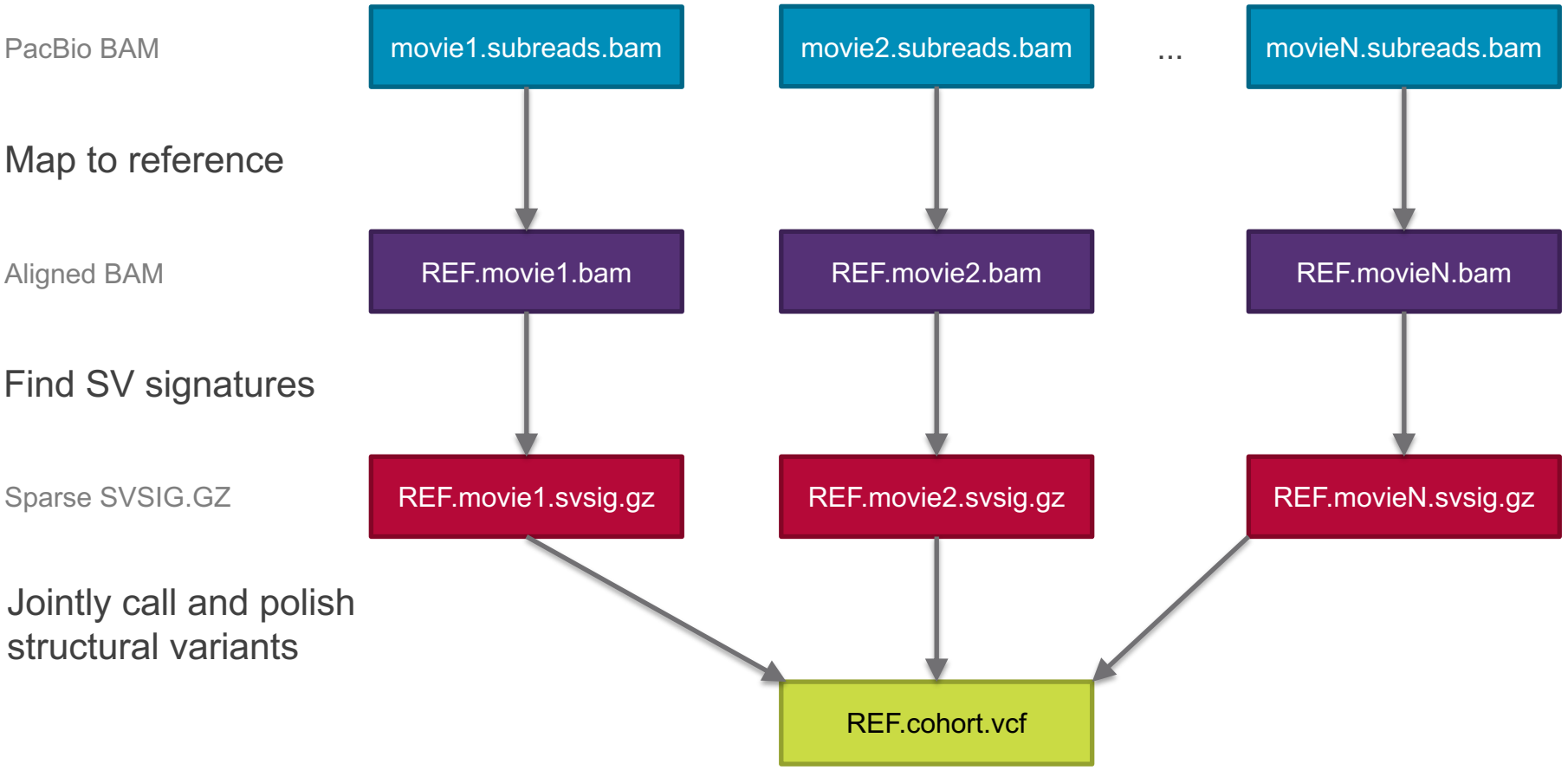
Limitations

– Only insertions and deletions

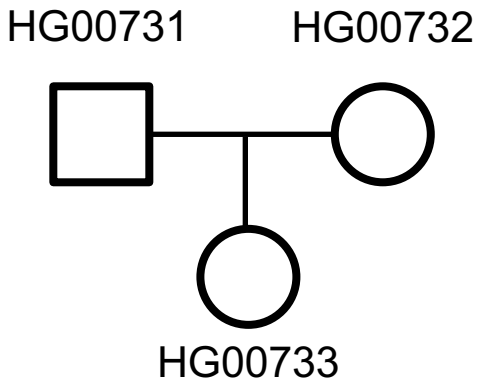
– Approximate breakpoints

– Slow for large cohorts

THREE STAGE WORKFLOW

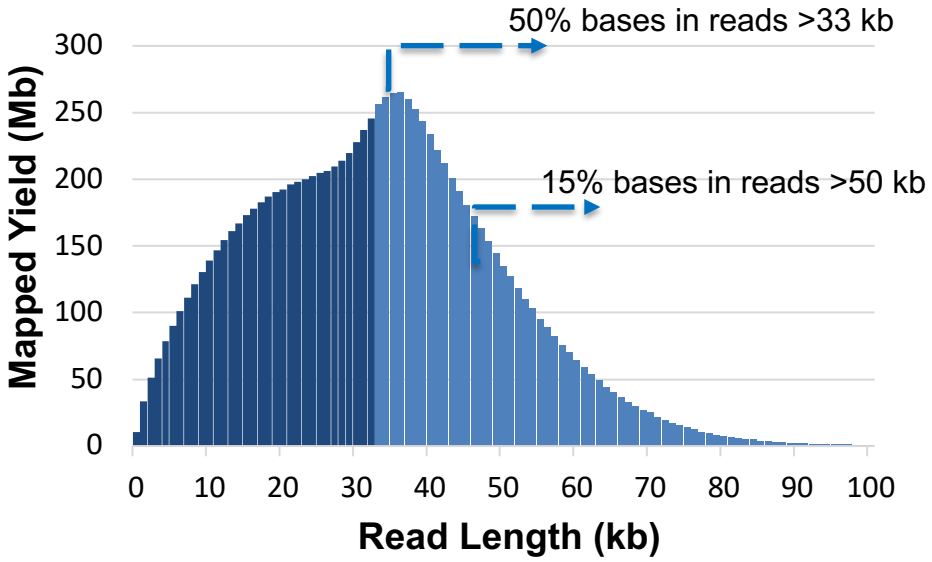


HG00733 – PUERTO RICAN CHILD



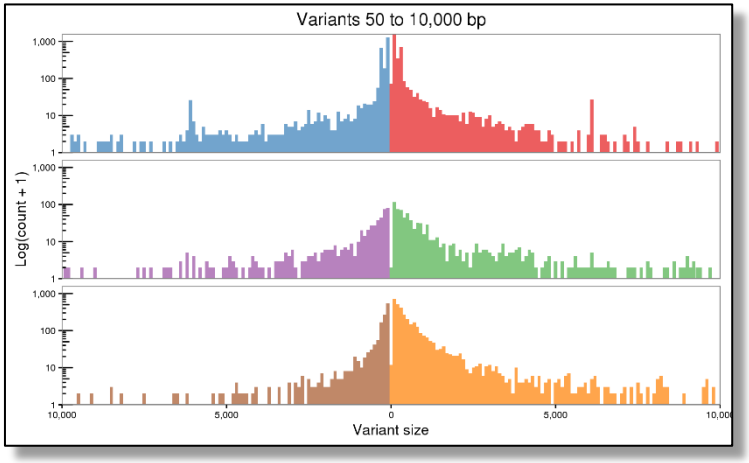
- Trio from 1000 Genomes Project and HGSVC
- SMRTbell Express Template Prep Kit
- Sequel System 2.1 chemistry and 5.1 software
- 28 Sequel SMRT Cells 1M

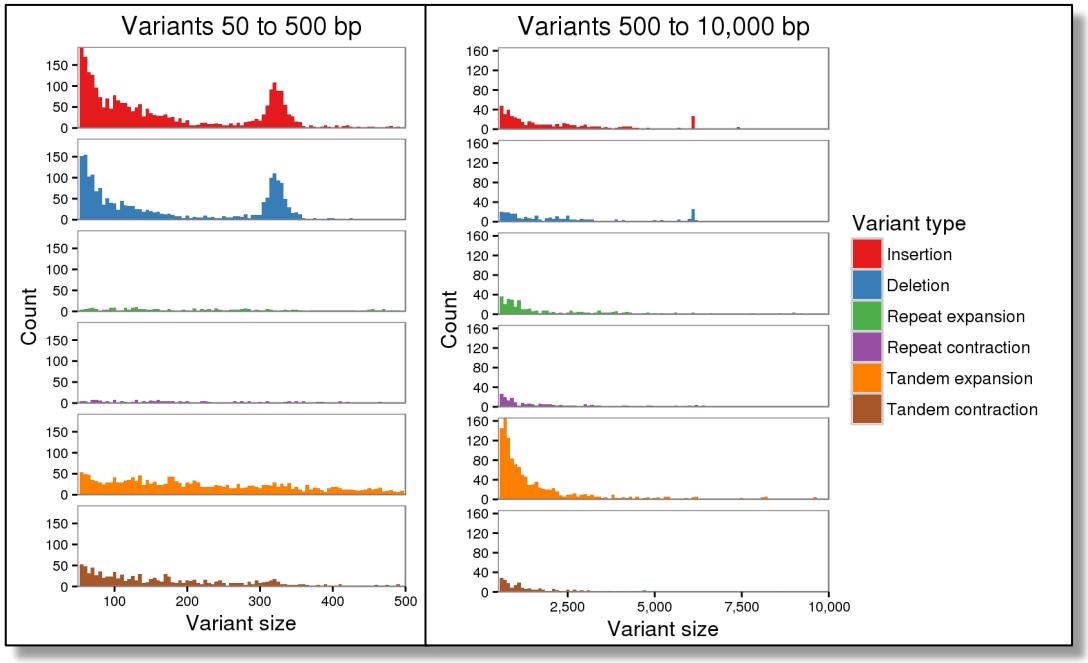
- 263 Gb raw yield (82-fold human)
  = 9.3 Gb / SMRT Cell
- 21 kb average read length



HG00733 – FALCON ASM REVEALS STRUCTURAL VARIATION

Reference	hg38
Sequences	415
Total Length	3.21 Gbp
Mean	7.73 Mbp
Max	248.96 Mbp
N50	145.14 Mbp





Query	HG00733
Sequences	947
Total Length	2.87 Gbp
Mean	3.03 Mbp
Max	86.08 Mbp
N50	31.43 Mbp

http://assemblytics.com

HG00733 – PBSV

28 servers each 16 cores (448c)

PBSV

Stage	CPU	Wall
Map to reference	9d	0h45m
Discover SV signatures	5h	0h5m
Joint call and polish variants	14h	0h23m
sum	**9d19h**	**1h16m**

16 servers each 64 cores (1024c)

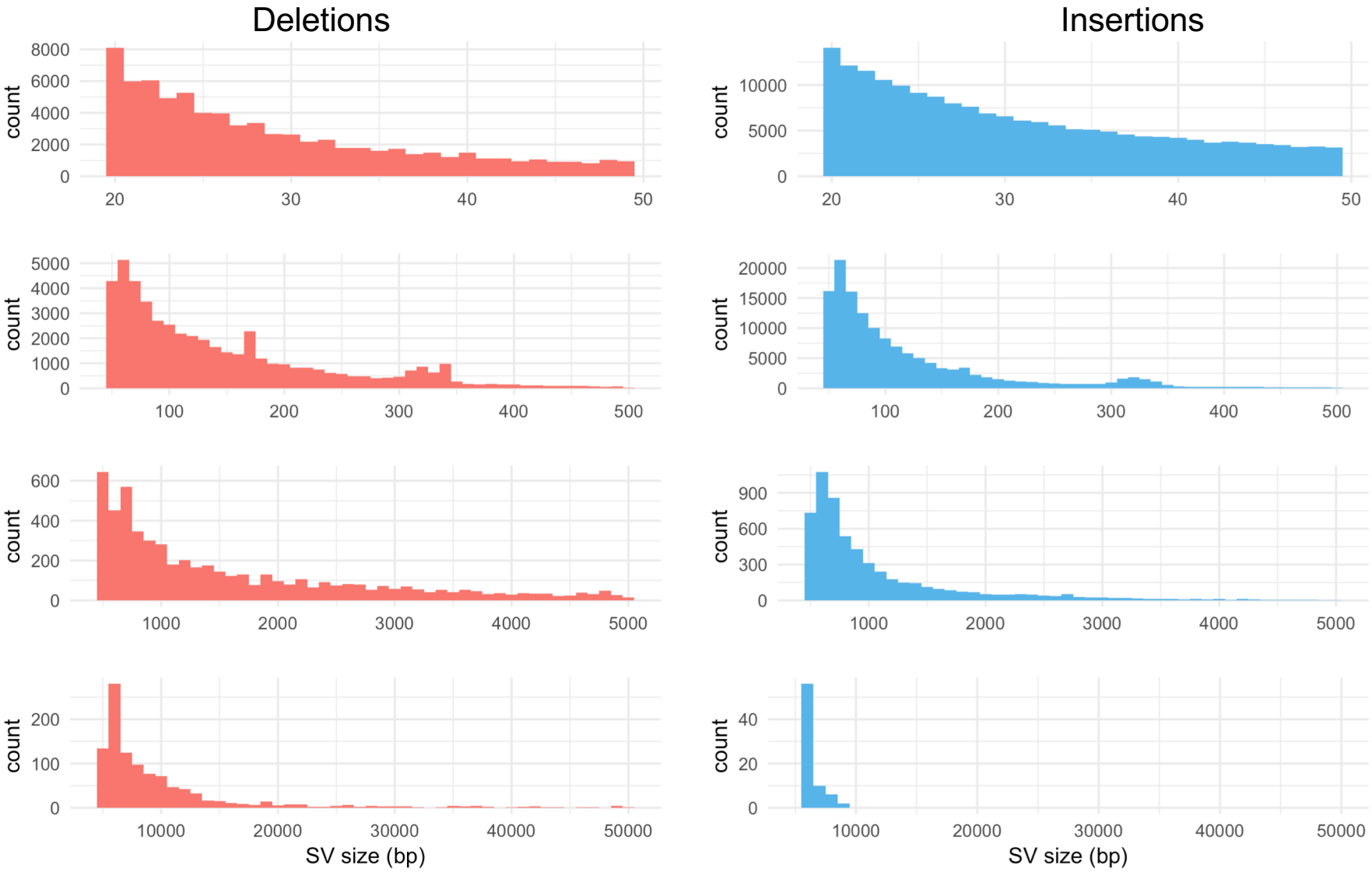
Assembly

Stage	CPU	Wall
Raw read overlap	862d	
Pread consensus	312d	
Pread overlap	845d	
sum	**5y194d**	**2d12h**

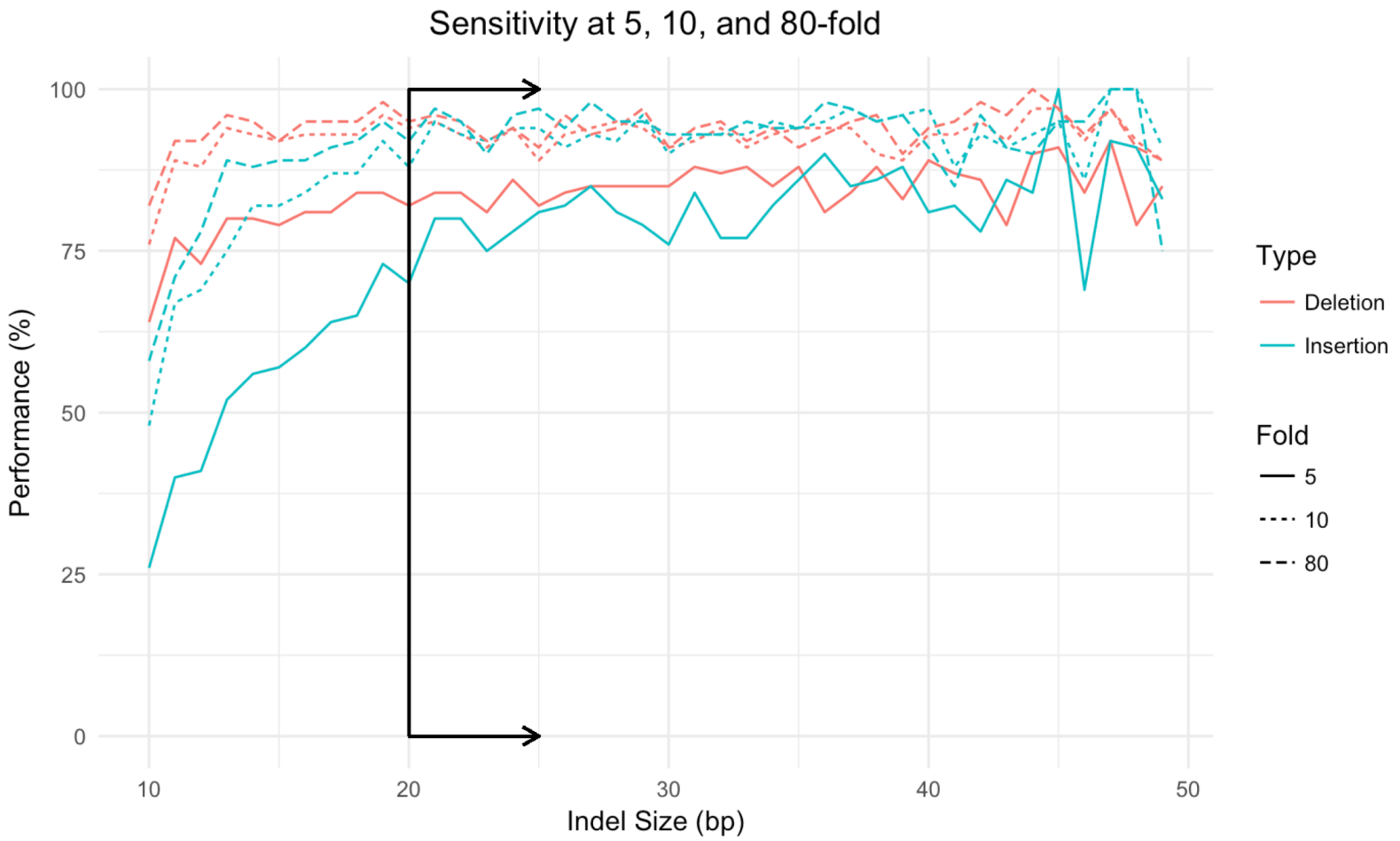
Low-fold calling

Fold	CPU	Wall
10-fold	6h	8m
5-fold	3h	6m

HG00733 – PBSV – INS/DEL LENGTH OVERVIEW

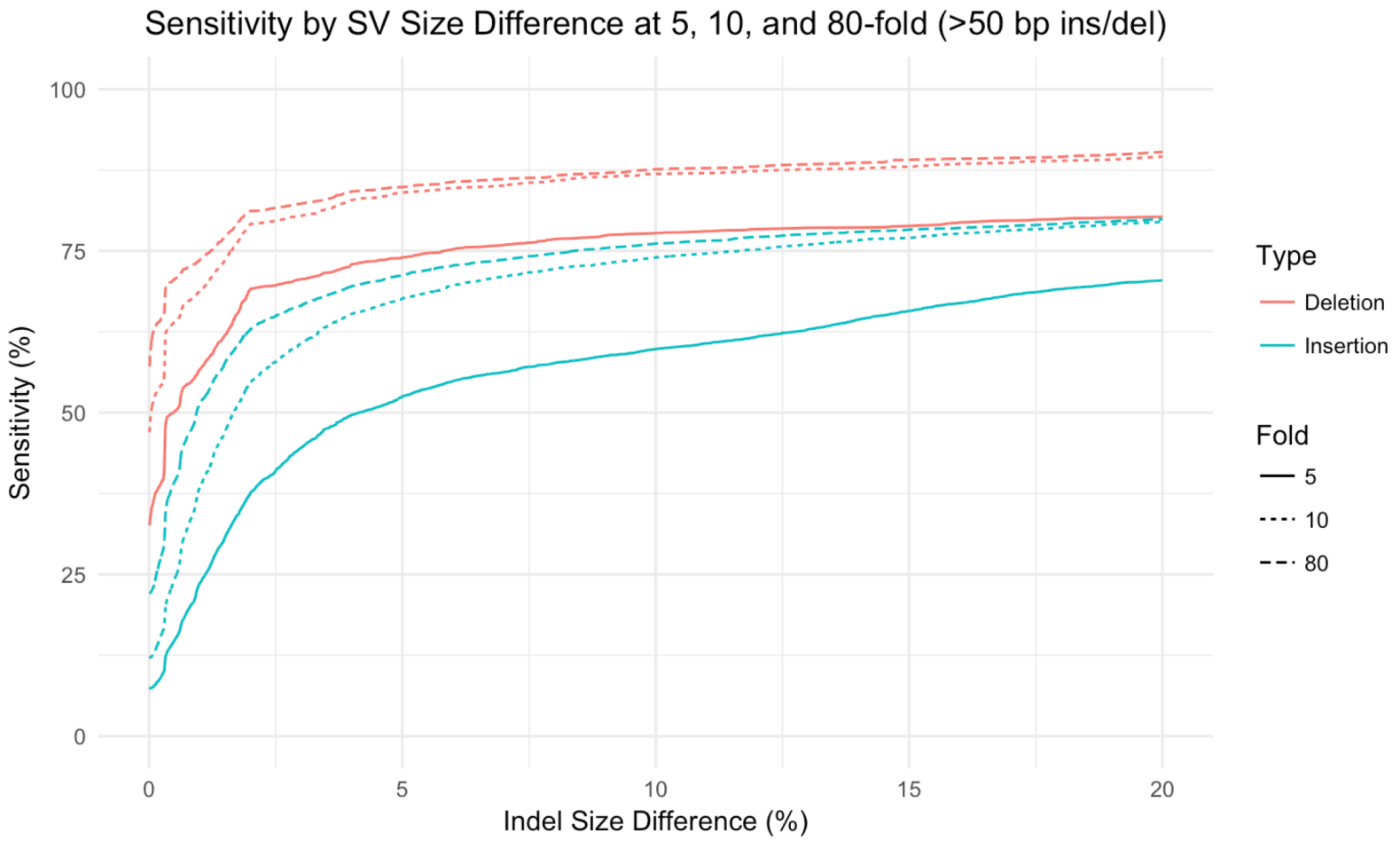


HG00733 – PBSV – INDEL SENSITIVITY



90% sensitivity > 20bp at 10-fold

HG00733 – PBSV – INS/DEL SIZE SENSITIVITY

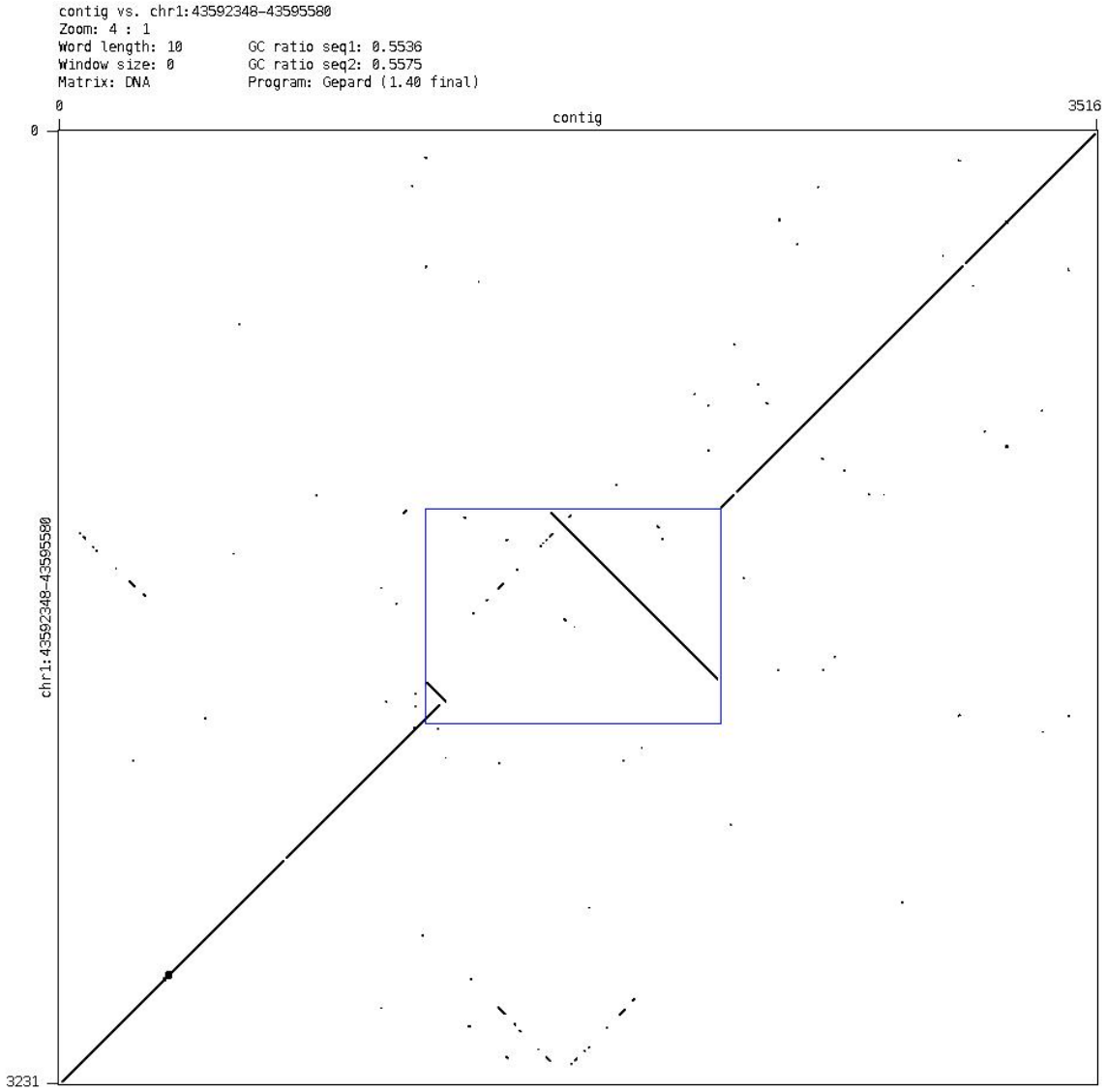


50% Deletions are base pair perfect at 10-fold

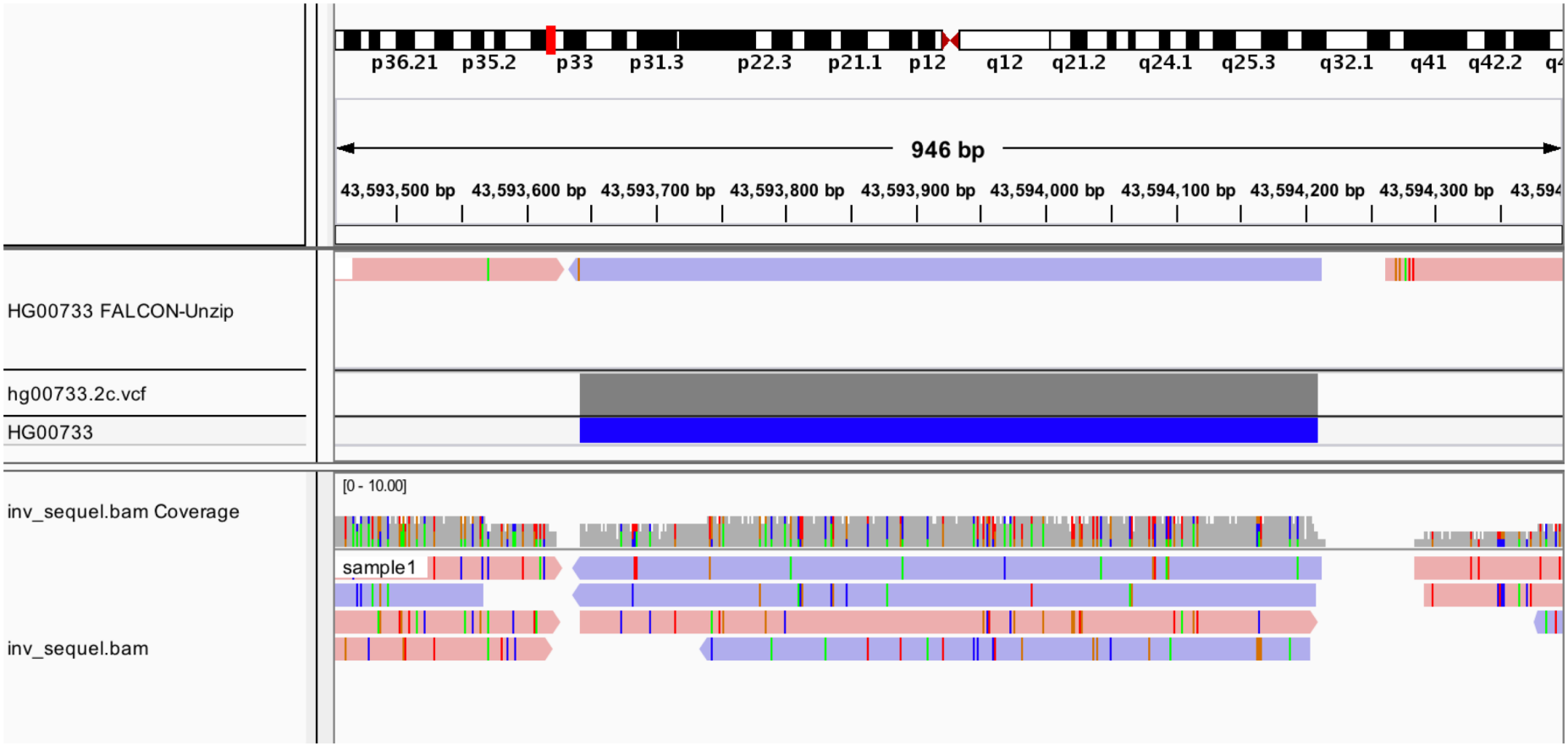
HG00733 – INVERSION EXAMPLE

~600bp inversion on chr1

Polished falcon contig vs hg38



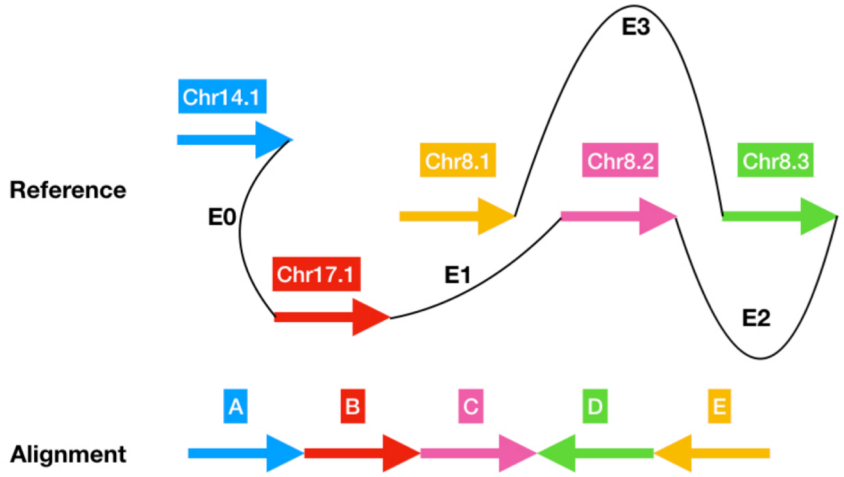
HG00733 – INVERSION EXAMPLE



5-fold coverage with precise breakpoints

(2 SMRT Cells)

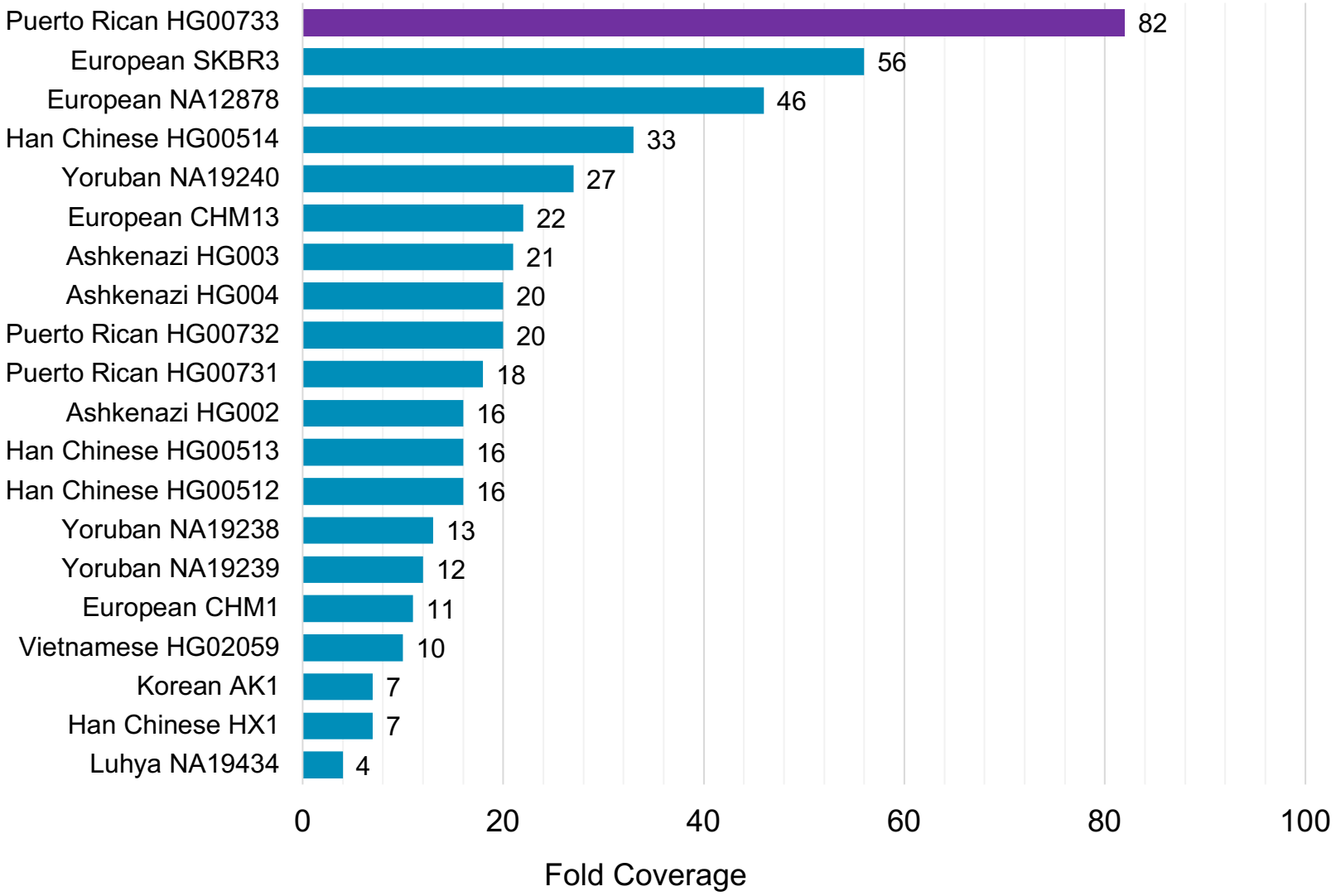
PBSV – TRANSLOCATIONS



Enable chaining of translocations via IDs: **`bnd_<chain>_<edge>`**

```
chr14    49789856    bnd_0_0    C    C[chr17:66046141[    .    PASS    SVTYPE=BND;CIPOS=0,13;MATEID=bnd_0_1
chr17    66046141    bnd_0_1    A    ]chr14:49789856]A    .    PASS    SVTYPE=BND;CIPOS=0,17;MATEID=bnd_0_0
chr17    66047636    bnd_0_2    C    C[chr8:125468594[    .    PASS    SVTYPE=BND;CIPOS=0,15;MATEID=bnd_0_3
chr8    125468594    bnd_0_3    C    ]chr17:66047636]C    .    PASS    SVTYPE=BND;CIPOS=0,23;MATEID=bnd_0_2
chr8    125468758    bnd_0_4    G    G]chr8:122895219]    .    PASS    SVTYPE=BND;CIPOS=0,13;MATEID=bnd_0_5
chr8    122895219    bnd_0_5    A    A]chr8:125468758]    .    PASS    SVTYPE=BND;CIPOS=0,35;MATEID=bnd_0_4
```

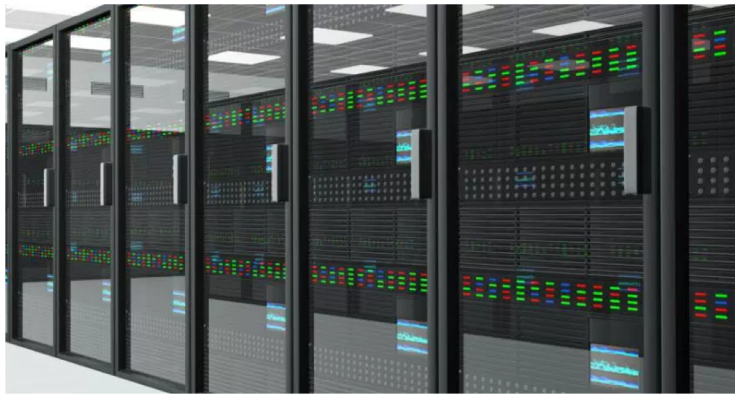
PBSV – HUMAN COHORT



PBSV – HUMAN COHORT

25 servers each 16 cores (400c)

Stage	CPU	Wall
Discover SV signatures	20h	0h17m
Joint call and polish variants	82h	0h51m
sum	**4d6h**	**1h08m**



https://www.tech-coffee.net/understand-failover-cluster-quorum/

Don’t have cluster?

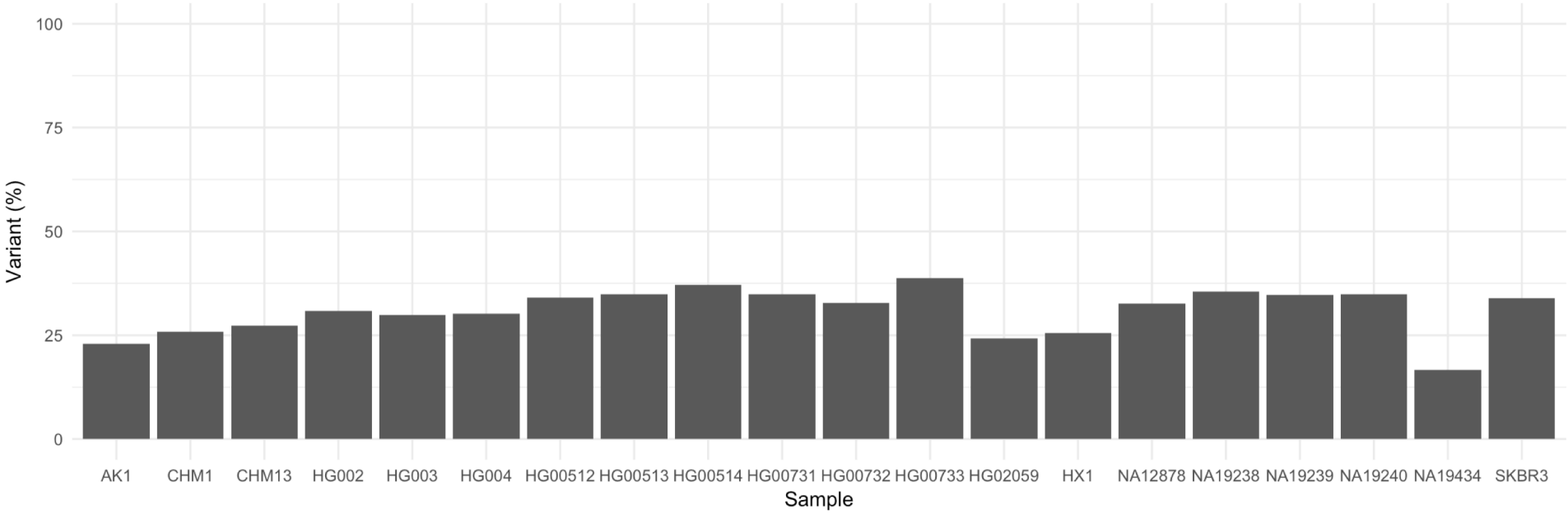
Single 16 core machine

1 day to call **20 humans** with **460-fold**

https://upload.wikimedia.org/wikipedia/commons/f/f1/Ibm_pc_5150.jpg

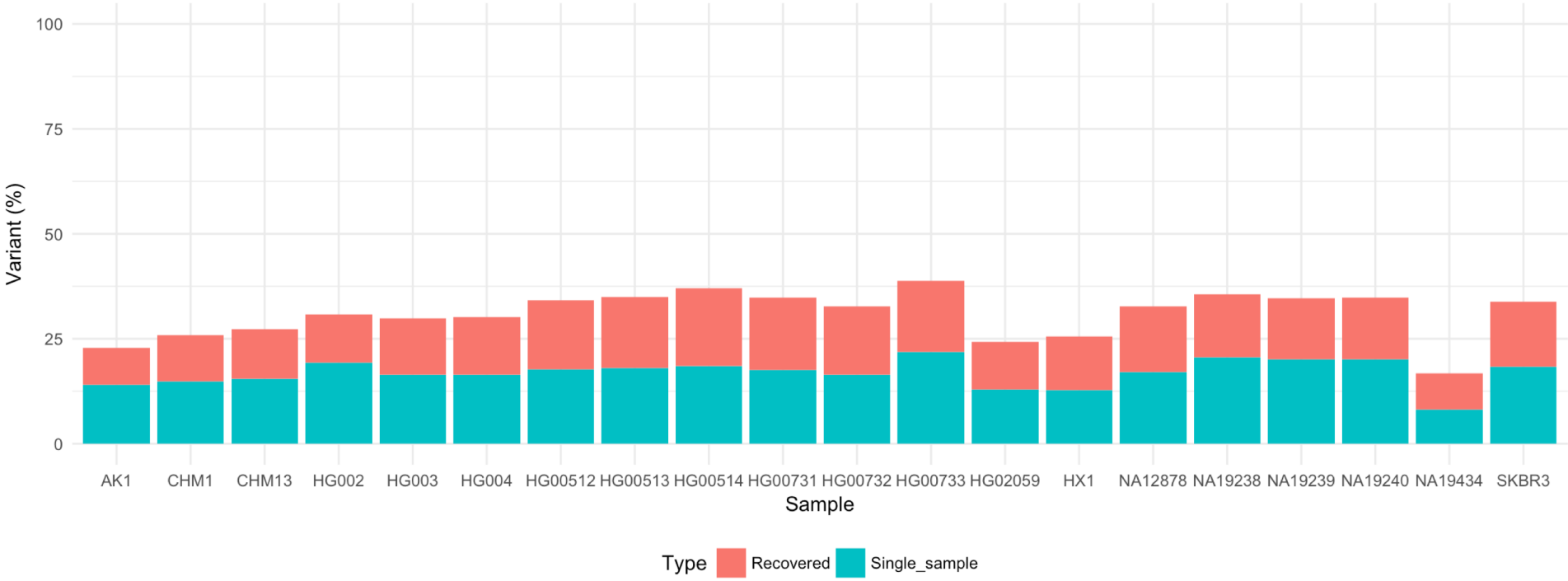
PBSV – HUMAN COHORT

Joint Variants: 832,398 (~700,000 indels)
<1% with no genotype across samples



PBSV – HUMAN COHORT

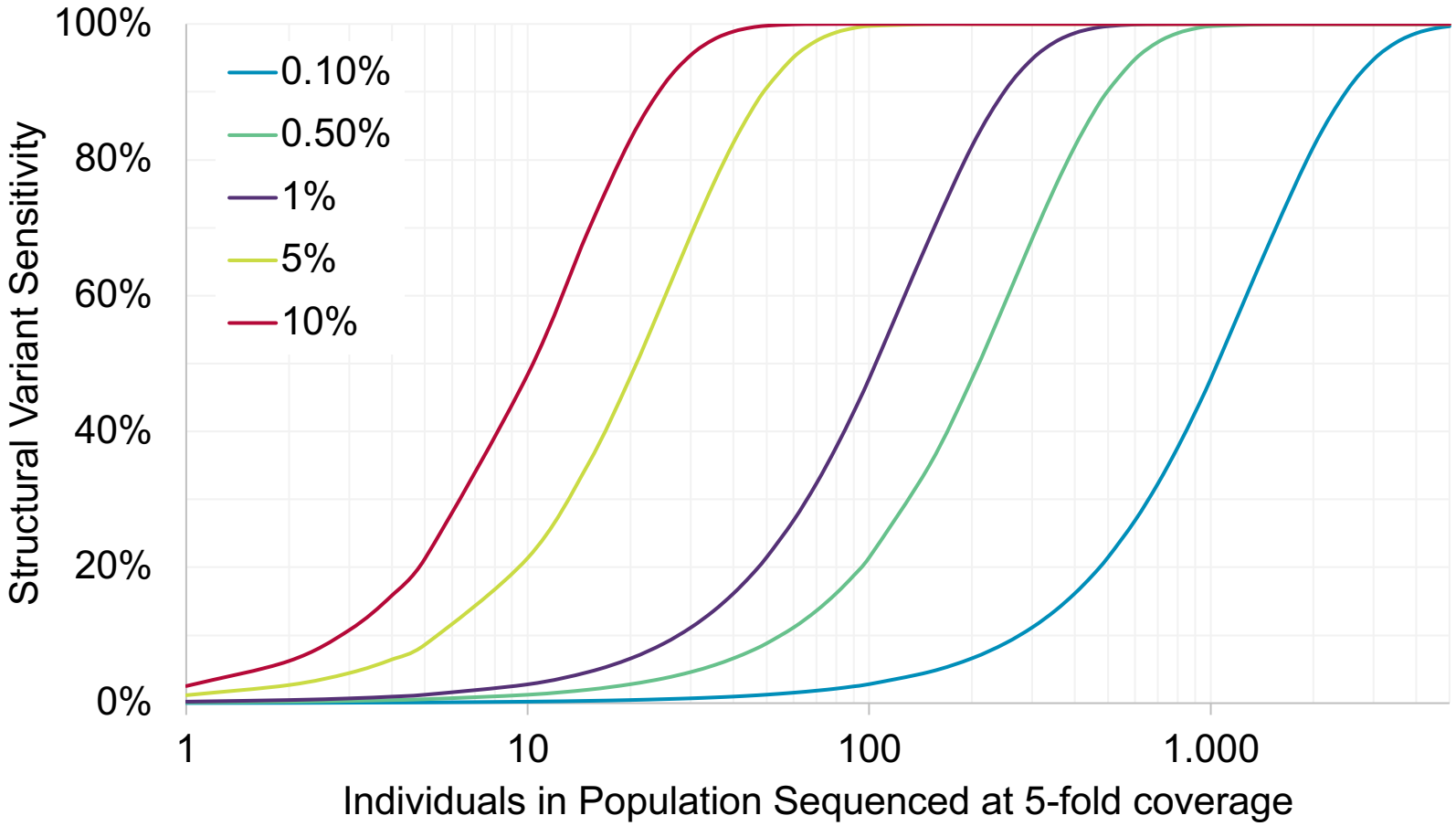
Joint Variants: 832,398 (~700,000 indels)
<1% with no genotype across samples



Example: Recovered SV by joint calling

```
#CHROM   POS       INFO                                HG002(son)  HG003(father)  HG004(mother)
chr1     1145518   SVTYPE=INS;END=1145518;SVLEN=20     0/1:1:12    1/1:11:12      0/1:4:13
```

5-FOLD COVERAGE FOR COMMON VARIANT DISCOVERY



Calculator: **pacb.com/sv**

SUMMARY

- End-to-end solution: library design to structural variants
- Native joint calling capability
- Scales beyond trio calling
- Precise variant size estimates, generate consensus insert sequences
- SV types: Indels down to 20bp, inversions, translocations
- Bioconda support

ACKNOWLEDGEMENTS

Data Sharing

Radboud UMC

HGSVC

GIAB

Open Source Tools

Fritz Sedlazeck and Michael Schatz (NGMLR)

Heng Li (minimap2)

Mark Gerstein (AGE)

PacBio

Aaron Wenger

Yuan Li

Paul Peluso

Greg Concepcion

PACBIO®

www.pacb.com