Disclosure Slide

Financial Disclosure for:
William J. Rowell
Staff Scientist, Bioinformatics

Pacific Biosciences
Accurate, comprehensive variant calling in difficult-to-map genes using HiFi reads

William J. Rowell¹, Aaron M. Wenger¹, Gunjan Baid², Maria Nattestad², Andrew Carroll², Pi-Chuan Chang²

1. Pacific Biosciences, Menlo Park, CA, USA
2. Google Inc., Mountain View, CA, USA

Introduction:
• ~5% (1,168) of human protein-coding genes contain an exon that is difficult-to-map with typical next-generation sequencing (NGS) read lengths
• Among the difficult-to-map genes are 193 with known medical relevance
• Long-read DNA sequencing provides increased mappability, accessing many of the difficult-to-map regions by connecting the homologous exon to neighboring unique sequence
• HiFi reads from the PacBio Sequel II System provide both long read length (15 kb - 25 kb) for mappability and high read quality (>99%) for accurate variant calling, expanding the regions of the genome that can be characterized to high precision and recall.
Accurate, comprehensive variant calling in difficult-to-map genes using HiFi reads

William J. Rowell¹, Aaron M. Wenger¹, Gunjan Baid², Maria Nattestad², Andrew Carroll², Pi-Chuan Chang²

Materials and Methods: Genome in a Bottle (GIAB) human reference samples HG002, HG003, and HG004 were sequenced to 35-fold HiFi read coverage on the PacBio Sequel II System. Matched 35-fold coverage with NGS reads was obtained on the Illumina NovaSeq 6000. Reads were mapped to the GRCh38 reference genome using pbmm2 for HiFi reads and BWA for NGS reads. Small variants were called using DeepVariant models for PrecisionFDA Truth V2 competition submissions. The variant callsets were compared to each other and to the GIAB v4.2 benchmark within exons previously reported to be problematic for NGS.
Accurate, comprehensive variant calling in difficult-to-map genes using HiFi reads

William J. Rowell¹, Aaron M. Wenger¹, Gunjan Baid², Maria Nattestad², Andrew Carroll², Pi-Chuan Chang²

Results: Figure 2. For difficult-to-map exons within the GIAB benchmark, PacBio HiFi reads detect an average of 1,188 true benchmark variants per sample, 22% more than are detected with NGS reads (975). Figure 3. Small variant precision in difficult-to-map exons is on average 96.6% for HiFi reads, markedly higher than 90.0% for NGS reads.
Accurate, comprehensive variant calling in difficult-to-map genes using HiFi reads

William J. Rowell¹, Aaron M. Wenger¹, Gunjan Baid², Maria Nattestad², Andrew Carroll², Pi-Chuan Chang²

Results: Figure 4. Extending outside of the benchmark, PacBio HiFi reads detect many additional variants compared to NGS. In HG002, 222 additional small variants are detected across 110 difficult-to-map exons of 41 medically relevant genes.
Accurate, comprehensive variant calling in difficult-to-map genes using HiFi reads

William J. Rowell$^1$, Aaron M. Wenger$^1$, Gunjan Baid$^2$, Maria Nattestad$^2$, Andrew Carroll$^2$, Pi-Chuan Chang$^2$

Results: Figure 5a. HiFi reads detect 14 additional small variants in difficult-to-map exons of $C4A$ for HG002.
Accurate, comprehensive variant calling in difficult-to-map genes using HiFi reads

William J. Rowell¹, Aaron M. Wenger¹, Gunjan Baid², Maria Nattestad², Andrew Carroll², Pi-Chuan Chang²

Results: Figure 5b. HiFi reads detect 5 additional small variants in difficult-to-map exons of SMN1 for HG002.
Accurate, comprehensive variant calling in difficult-to-map genes using HiFi reads

William J. Rowell¹, Aaron M. Wenger¹, Gunjan Baid², Maria Nattestad², Andrew Carroll², Pi-Chuan Chang²

Results: Figure 5c. HiFi reads detect 2 additional small variants in difficult-to-map exons of STRC for HG002.
Accurate, comprehensive variant calling in difficult-to-map genes using HiFi reads

William J. Rowell¹, Aaron M. Wenger¹, Gunjan Baid², Maria Nattestad², Andrew Carroll², Pi-Chuan Chang²

1. Pacific Biosciences, Menlo Park, CA, USA
2. Google Inc., Mountain View, CA, USA

Conclusion: HiFi reads have both high mappability and high read quality, which enables accurate small variant calling in difficult-to-map genes that are challenging for NGS. We predict that large-scale use of HiFi reads in disease cohort studies will discover additional disease genes and variants that have remained beyond the reach of NGS.

Data availability:
• 35-fold PacBio Sequel II System HiFi and 35-fold Illumina NovaSeq reads for GIAB samples HG002, HG003, and HG004 are available from PrecisionFDA.

Analysis tools:
• Reference alignment by pbmm2 and BWA. Variant detection by DeepVariant.
• Variant benchmarking with hap.py and the GIAB v4.2 small variant benchmarks.