HiPhase: Jointly phasing small and structural variants from HiFi sequencing

J. Matthew Holt, Christopher T. Saunders, William J. Rowell, Zev Kronenberg, Aaron M. Wenger, and Michael Eberle
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Why do we phase?

In diploid organisms, phasing is the problem of assigning heterozygous variants to one of two haplotypes. Phasing has many applications in human genetics:

- **Rare disease:** Detecting if pathogenic variation is in *cis* or *trans* in autosomal recessive diseases (e.g., Tay-Sachs disease)
- **Methylation:** Identifying regions with allele-specific methylation associated with imprinting disorders (e.g., Prader-Willi and Angelman syndromes)
- **Pharmacogenomics:** Fully phased haplotypes can provide accurate dosing guidelines for many pharmaceuticals

HiPhase – a read-backed phaser for PacBio HiFi datasets

Read-backed phasing allows all detected variants, including *de novo* variants, to be phased using read-level evidence. PacBio HiFi sequencing provides long, accurate observations that are ideal for phasing when researching both inherited and *de novo* variation. HiPhase leverages HiFi sequencing to provide:

- **Better phasing:** Phase blocks are longer and with fewer errors than previous approach
- **Additional phased variants:** HiPhase can *jointly* phase SNVs, indels, structural variants, and tandem repeats
- **No downsampling:** All provided HiFi reads are used
- **Gap spanning:** Logic for constructing phase blocks across reference gaps and homozygous deletions
- **Usability features:** simultaneous haplotagging, innate multi-threading, and built-in statistics gathering

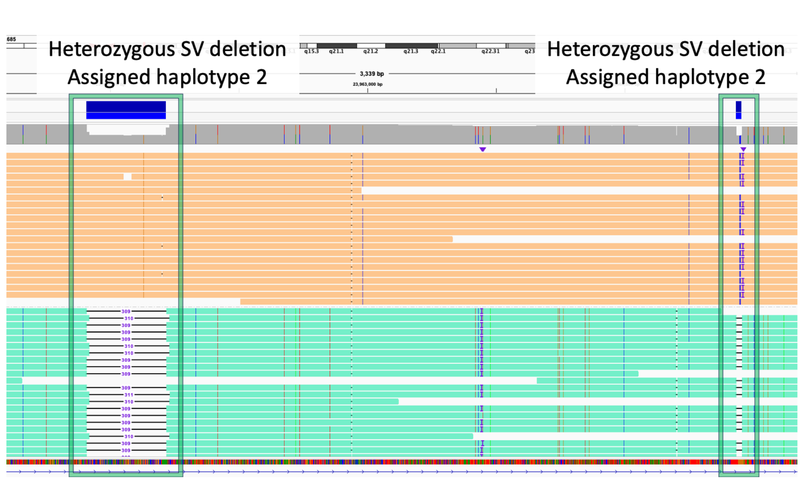

Figure 1. Example of joint phasing. This region contains two heterozygous structural variants on the same haplotype. HiPhase reports the correct haplotype assignments for both variants, as well as the small variants located in-between them.

HiPhase performance

We evaluated HiPhase with three HG002 replicates sequenced to ~30-fold coverage on the Revio system. When compared to the Genome in a Bottle phased benchmark we observed:

- **Longer phase blocks** with **fewer errors** and **more fully phased genes** relative to previous approach
- Over **3 million phased variants** with **11K phased structural variants** and **70K phased short tandem repeats** per dataset

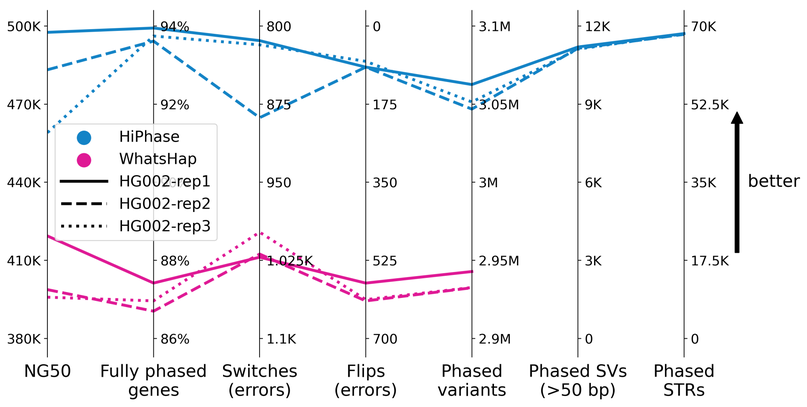

Figure 2. Summary comparison metrics. HiPhase generates longer phase blocks with fewer phase errors than the previous approach. Additionally, ~11K structural variants and ~70K short tandem repeat variants per dataset are jointly phased with SNVs and smaller indels.

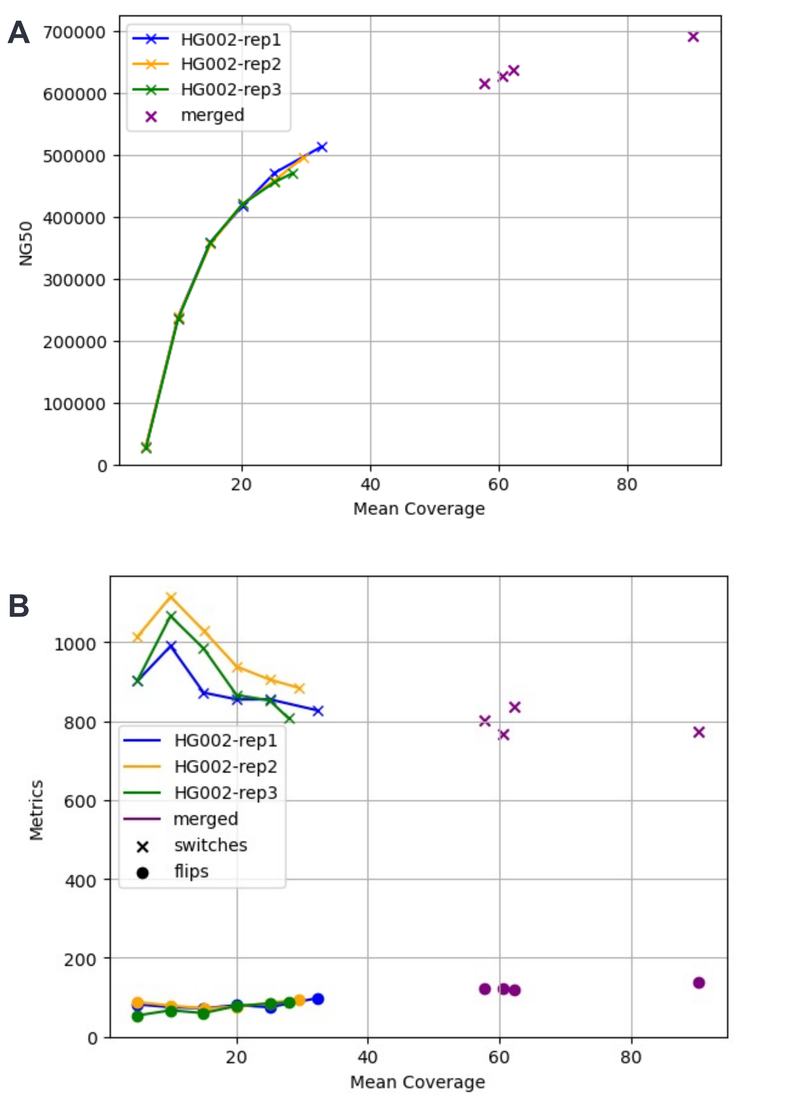

Figure 3. Effect of coverage on phasing performance. Labeled lines indicate downsampling of a single replicate, and "merged" datasets represent artificial merging of two or three replicates into a single dataset. All provided reads are used by HiPhase, providing gains in phase block length at high coverage **(A)** with negligible impact on errors **(B)**.

Pharmacogenomic applications

Pharmacogenomic (PGx) diplotypes can be derived from unphased data, but often generate ambiguous solutions. Phase blocks from HiPhase can resolve ambiguous solutions by grouping the variants into long haplotypes. We used a phase-aware PGx diplotyper, pb-StarPhase, to resolve the diplotypes for known PGx genes in CEPH pedigree 1463.

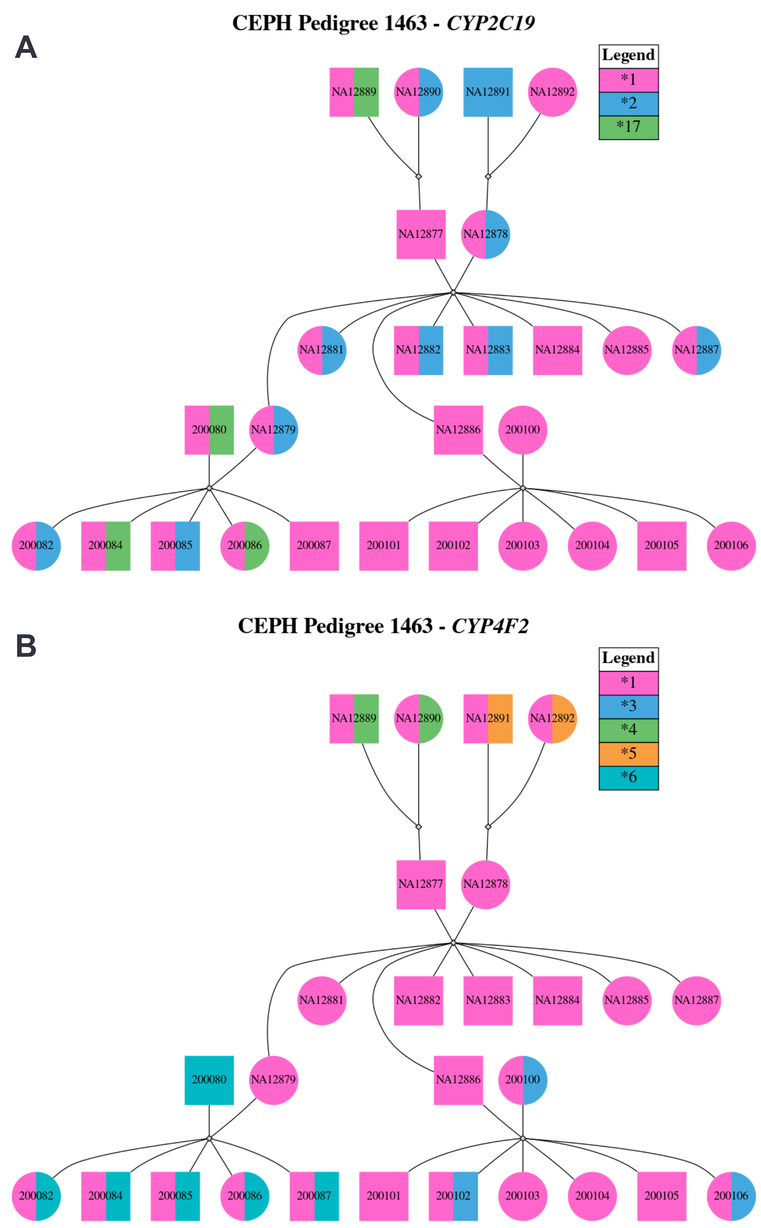

Figure 4. Example phased pharmacogenomic pedigrees. The phased diplotypes for genes *CYP2C19* **(A)** and *CYP4F2* **(B)** on CEPH pedigree 1463. HiPhase was used to generate phased VCF files from HiFi whole genome sequencing. These phased VCFs were used as input to pb-StarPhase to generate diplotype calls for CPIC PGx genes. Pedigrees were drawn using the diplotype calls from the tool. Datasets that are homozygous show up as a single color, and heterozygous as two colors. The diplotype calls in both example pedigrees are consistent with expected inheritance patterns.

Read more about HiPhase in our pre-print or download HiPhase at:
https://www.github.com/PacificBiosciences/HiPhase

Visit more PacBio posters for information on:
STR calling with TRGT-denovo – poster PB3070
CEPH Pedigree 1463 – poster PB3397