Docket #: S11-488

Phasing Algorithm That Incorporates Sequencing Read Data To Improve Haplotype Inference

Researchers in Dr. Carlos D. Bustamante's lab have developed a phasing algorithm that incorporates sequencing read information, population and individual genotype data to provide more accurate haplotype reconstruction. Humans carry two copies of every chromosome in their genome: one from each parent. Determining which alleles were inherited together from one parent or the other (phasing) is a critical component of many downstream medical and population studies. Methods have been developed to phase unrelated individuals in the absence of parental information but these methods restrict themselves to genotype, population, or read data alone when reconstructing haplotypes and obtaining better performance has become quite challenging. This algorithm overcomes this limitation by incorporating sequence read information to reconstruct more accurate haplotypes from the genomes of related or unrelated individuals using population level genotype and/or haplotype data (when available).

Stage of Research
A comparative study has demonstrated that this approach yields significantly more accurate results over existing methods. The inclusion of paired end read data has also been shown to be critical for the phasing of rare variants.

Applications

- Generation of high quality haplotypes for:
    - Demographic inferences for a population.
    - Identity by Descent (IBD) studies.
    - Cryptic relatedness studies.
    - Phased local ancestry deconvolution in admixed populations.

    - Haplotype-based association studies, such as those conducted in medical genetic studies.

Advantages

- First algorithm designed to interlace sequencing read data with genotype data.
- Superior accuracy over existing phasing algorithms.
- Enables the phasing of rare variants that would normally be impossible to accurately haplotype.
- Can be generalized to benefit from other sources of phasing information.
- Efficiently implemented in C++ and compiled for a wide range of operating system environments.
- Employs a parallelized, multithreaded software architecture yielding much faster results than comparable software packages such as fastPHASE.

Publications

- U.S. Published Patent Application 20140045705, "Techniques for Determining Haplotype by Population Genotype and Sequence Data".

Patents

- Published Application: 20140045705
- Issued: 10,847,248 (USA)

Innovators

- Carlos Bustamante
- Fouad Zakharia

Licensing Contact

Hyunjin Kim

Licensing Manager, Life Sciences

Email