Faster Computation of DNA Maps Provides Insight into Genetic Basis of Human Disease

October 26, 2006

Doug Ramsey

High-throughput sequencing of an individual's DNA yields a map of genetic variation which can give clues to the genetic underpinning of human disease. The current technologies collect genotypes, or information from the individual's two chromosomes. Yet many scientists believe that drilling down to the variations between individuals' DNA at the level of each chromosome – so-called haplotypes – will permit more accurate study of genetic differences and their consequences for medical research and the study of evolution.

Experimental methods for deriving these haplotypes are expensive and time-consuming. But now experts in bioinformatics at two California research institutes have used a different, very fast and relatively low-cost computational tool to 'crunch' the world's largest repository of genotypes to predict their haplotypes - and they did so in less than 24 hours, approximately 1,000 times faster than the prevailing technology until now. Their findings are featured in a special issue of the journal *Genome Research*, published today.

"This information provides an invaluable resource for understanding the structure of human genetic variation," said lead author Eleazar Eskin, a professor of computer science and engineering at the University of California, San Diego who is affiliated with the California Institute for Telecommunications and Information Technology (Calit2). "A deeper understanding of the data will improve the design of studies that look for associations between certain genes and disease or inherited conditions."

The team from UCSD and the International Computer Science Institute (ICSI) processed all 286 million human genotypes in the dbSNP database of the National Center for Biotechnology Information (NCBI), part of National Institute of Health's National Library of Medicine. The repository includes all publicly available data on single nucleotide polymorphisms (SNPs), which are sites in the DNA sequence where individuals differ at the level of nucleotides.

These SNPs (pronounced *snips*) are locations in the human DNA sequence where two possible bases occur in the population. SNPs account for the most common type of variation in DNA sequence in humans and due to the recently developed high-throughput genotyping technology, genotype information on an individual's SNPs can be collected very cheaply.

Enter computational biologists around the world who have been devising ways to infer or extrapolate these haplotypes from the flood of genotype data produced by DNA sequencing efforts. Eskin and colleagues Noah Zaitlen and Hyun Min Kang at UCSD, and research scientist Eran Halperin at ICSI, worked with NCBI scientists Michael Feolo and Stephen Sherry to infer haplotypes based on all of the data from genotyping studies deposited in NCBI's dbSNP database. Rather than use standard methods for inferring haplotypes, the computer scientists used HAP, a software tool originally developed at ICSI by Halperin and Richard Karp in collaboration with Eskin.

They ran the HAP algorithm on all dbSNP data sets using a cluster of 30 Intel Xeon processors provided by Calit2's National Science Foundation-funded OptIPuter project, in cooperation with the National Biomedical Computation Resource. Both organizations are based at UCSD. "In under 24 hours we were able to process more

than 286 million haplotypes, partition those haplotypes into blocks, or regions, of limited diversity, and determine a set of 'tag' SNPs that capture the majority of genetic variation," explained Halperin.

The researchers' article appears in a special issue of *Genome Research* on "Human Genetic Variation," and its publication coincides with the release of a wide-ranging genotype study by the International HapMap Consortium in the journal *Nature*. The group's HapMap is a map of haplotype blocks and the tag SNPs that identify the haplotypes from a database of 160 million genotypes of 270 individuals from four different populations with ancestors from parts of Africa, Asia and Europe. The HapMap data is a major resource for understanding the structure of human variation and the genetic basis of human disease.

All of the HapMap data is deposited in NCBI and was made available to the California researchers for their computation, along with more than a dozen other data sets, including the second-largest behind HapMap - 110 million genotypes published earlier this year by a consortium led by Perlegen Sciences.

"The speed with which we are able to compute the entire dbSNP database of genotypes is a combination of the speed of our algorithm and the computational resources that allowed us to do it so quickly," explained Eskin, a professor in UCSD's Jacobs School of Engineering. "We have demonstrated that haplotype phasing can be done routinely every time there is a new release of data deposited in the NCBI database."

"By reducing the waiting time to just 24 hours, NCBI can make it an integral part of the build cycle for dbSNP," said NCBI's Stephen Sherry. "Every time there is a new release of polymorphism and human variation information in our database, our colleagues in California will be able to re-compute the haplotypes and tag SNPs." To underscore that point, in early October the researchers ran another complete computation on an updated version of the NCBI database that has not yet been made public.

ICSI's Halperin notes that working with the entire dbSNP database showed that HAP works well on diverse data sets. "The challenge of analyzing such a large dataset is enormous, since the integration of the different datasets is not a simple task," explained the research scientist. "In particular, different data sets have different characteristics, and one has to take this into account. This project demonstrates the ability of HAP to efficiently deal with different types of data, for instance, unrelated or related individuals". Indeed, for the project, Halperin extended the HAP algorithm to work with 'trios'- where genotypes are available for a mother, father and their child - taking into account that the haplotypes of the children are copies of the haplotypes of the parents.

As a side effect of their research, the computer scientists are now depositing 15 gigabytes of data into dbSNP, and their article in *Genome Research* aims to encourage the research community to use the data depository as a scientific resource. Researchers can use these reference data sets as tools to guide their own studies into the genetic basis of common diseases.

To that end, the team's next collaboration with NCBI researchers will be to help design disease-association studies. "If a researcher is interested in a specific gene, we can use all the available data to come up with how to design the experiment," said Eskin. "We can tell how many individuals' genotypes need to be sequenced - and how many and which SNPs to collect - to minimize the cost and processing power needed for the most effective study correlating genetic data and the incidence of disease."

Disease association research is the main reason why the group from Calit2 and ICSI opted to identify tag SNPs across the entire NCBI database and make all of them available to the research community. Said Halperin: "If you are going to perform a disease association study, it's more economical to use these tag SNPs than the entire data."

Media Contacts: Doug Ramsey, UCSD/Calit2, 858-822-5825 Leah Hitchcock, ICSI, 510-666-2974