Low-Coverage Whole Genome Sequencing: Learning From Less Data

The genomic research field has been trending towards more data points, so the emergence of low-coverage sequencing.

April 16, 2019 By National Cancer Institute

By Peggy I. Wang

First there were microarrays. Then high-throughput whole-genome sequencing. Average sequencing depth of 20X, 40X, 80X. Deep sequencing of individual cells. And the next new revolution in sequencing is… 0.1X sequencing?

The genomic research field has been trending towards more data points, so the emergence of low-coverage whole-genome sequencing (LC-WGS), also known as ultra low-pass WGS, may be surprising. With the general agreement that sequencing depth should be 10 – 30X to reliably call mutations, what is the value of sequencing merely 0.1 – 3X?

I spoke to two researchers who are using the same low-coverage technique in very different ways in their work: Robert Davies, PhD, research fellow at Hospital for Sick Children, Toronto, who studies genetic architecture of diseases, and Viktor Adalsteinsson, PhD, associate director of the Gerstner Center for Cancer Diagnostics at the Broad Institute, who studies cancer diagnostics.

Both researchers have developed computational methods for learning more from less data, enabling an affordable new class of genomic studies.

What Exactly Is Low-Coverage Whole Genome Sequencing?
Viktor Adalsteinsson: The execution of LC-WGS is simple: instead of loading a single sample onto the sequencer, we load many samples so that each receives a small amount of sequencing. Because it only requires a trivial amount of sequencing, LC-WGS can be applied to many samples at low cost.

Robert Davies: With the decreasing cost of sequencing, LC-WGS has become a promising and affordable alternative to genotyping arrays and other technologies to get molecular information cheaply and accurately.

Enabling More Genomic Studies With LC-WGS
RD: LC-WGS allows a statistical geneticist like me to study a lot more populations. In genome-wide

association studies (GWAS), we collect phenotypes (like height or disease) and genotypes from a large set of individuals and look for associations between the two. For example, genetic loci within nicotine receptor and telomere-related genes have been associated with lung cancer susceptibility through GWAS.

Genotype collection for GWAS is usually done by genotyping arrays, which first became commercially feasible for this type of study around 2005. Today, arrays can obtain genotypes for fixed panels of 500,000 common single-nucleotide polymorphisms, or SNPs, at relatively low cost.

Though affordable when compared to whole-genome sequencing type studies, GWAS are limited: you're restricted to the sites on the array and you need a large reference panel to compare your data with. With LC-WGS, however, you have the whole genome at your disposal and can use imputation to fill in the blanks.

Statistical Imputation to Help Complete LC-WGS Data

RD: Statistical imputation is routinely used to predict the genotypes of positions not measured on an array using external reference panels. E.g., if we have 1,000,000 SNPs on a chromosome but only measured 50,000 with our array, we can use the patterns of variation found in the 1000 Genomes project to help fill in the blanks for the many SNPs not measured.

With LC-WGS, we can go a step further to impute genotypes using the set of LC-WGS data itself and no external reference. So if one sample is missing information on a particular SNP, we can recover it through imputation applied to the data set as a whole.

The process we've used is to pretend our population was founded by some number of people (say 100) some number of generations ago (say 800) and make a first guess of the founding genotypes. We then iteratively 1) impute the missing genotypes of the LC-WGS samples using the founding genotypes and 2) re-estimate the founding genotypes given the imputed data. We stop once we've converged to an acceptable solution.

Imputation makes LC-WGS particularly promising for animal studies, where we would otherwise need to develop new arrays (which can be expensive). As the cost of sequencing continues to drop, it is becoming increasingly advantageous for human studies as well. E.g., LC-WGS has already been used to identify two loci for major depressive disorder in Chinese women.

Gaining Power to Study Minorities and Underrepresented Populations

RD: With less dependence on reference panels, LC-WGS holds promise as a means to facilitate studies specifically in non-European populations. Large-cohort studies are known to lack minorities and thus serve as poor reference panels for studies with minorities. Using LC-WGS and imputation is an affordable solution to the problem.

LC-WGS has great potential for genetic research in Africa, where the genetic diversity is greater than out-of-African populations. An initiative has been launched to study psychiatric genetics in roughly 35,000 African individuals, for which it is considering LC-WGS.

LC-WGS of fetal cell-free DNA (cfDNA) from non-invasive prenatal testing is another promising avenue of research. For example, a study with 140,000 Chinese women has identified associations with maternal traits, such as height and BMI, as well as insights into the genetic structure and migration history of the Chinese population.

LC-WGS For Assessing Cancer in Patient Samples

VA: My work is in cancer diagnostics, where we are finding LC-WGS to be a cost-effective technique for checking tumor content and quality. Patient blood and tumor biopsies can vary greatly in tumor content and quality—both of which we ideally want to know before proceeding with genomic analysis.

We have used LC-WGS to confirm that cells isolated from blood are indeed cancer cells, that sequencing libraries generated from single cells uniformly represent the genome, and that sequencing libraries generated from cfDNA harbor tumor DNA.

In this manner, LC-WGS can help us identify patient samples eligible for more extensive genomic profiling. But the technique is even more versatile; we've developed a computational method to extract clinically-relevant features directly from LC-WGS data.

Identifying Prognostic Copy Number Alterations from Cell-Free DNA

VA: Our team's method simultaneously detects somatic copy number alterations (SCNAs) and estimates the tumor content of cfDNA in blood. So from a minimally-invasive blood test, we can profile SCNAs from large numbers of cancer patients and determine how they change over time.

The process for identifying SCNAs from so little data is similar to existing methods that work on deep sequencing, but uses larger genomic windows (or bins) to compensate for lower coverage and probabilistic models to extract information. In this way, we've recovered SCNA landscapes from LC-WGS of cfDNA similar to that derived from whole-exome and whole-genome sequencing of tumor biopsies in patients with metastatic prostate or breast cancer.

Other ways to assess SCNAs, such as arrays, tend to require much more sample and have less resolution than LC-WGS. Resolution for arrays are constrained by the probe sequences used, whereas LC-WGS provides a genome-wide (unbiased) sampling.

Measuring Tumor Fraction to Predicting Patient Outcomes

VA: Profiling SCNAs in a sample also allows us to quantify the fraction of tumor-derived cfDNA in blood. We've now studied tumor fraction and how it changes during therapy in patients with metastatic breast and prostate cancer.

This post was originally published by the National Cancer Institute. It is republished with permission.
