Molecular pathology selected abstracts

Role of DNA barcodes in discovering species and their role in ecosystems

August 2019—As sequencing technology becomes faster and less expensive and devices become smaller and more portable, the application of DNA sequencing to various branches of science is rapidly expanding. Advances in this technology have, for example, led to the development of portable genomic laboratories that use fast, inexpensive, and portable DNA sequencers for species identification. The Earth is estimated to have between 8.7 million and 20 million types of plants, animals, and fungi, but only about 1.8 million of them have been given formal descriptions. The task of species identification historically has relied heavily on morphologic phenotyping performed by a limited pool of skilled taxonomists. More recently, however, DNA sequencing was introduced as a tool to further characterize species. Various DNA targets, such as ribosomal DNA (rDNA), single nucleotide polymorphisms (SNPs), and evolutionarily conserved genes have been sequenced and ascribed to known species as taxon "barcodes," the name applied to the stretches of DNA sequenced. In 2003, Paul Hebert, et al., at the University of Guelph, in Canada, proposed using the mitochondrial gene cytochrome *c* oxidase I (*COI*) as the basis for a global bioidentification system for animals. By that time, robust universal primers were available for the gene, allowing representatives of most animal phyla to be tested. The later finding that *COI* possessed a greater range of phylogenetic signal than any other mitochondrial gene cemented *COI* as the target of choice for many bioidentification efforts. In 2015, the International Barcode of Life (IBOL) consortium, an alliance of research organizations spearheaded by Hebert, completed its Barcode 500K project by assembling DNA barcode records for 500,000 species in five years. The data from Barcode 500K populate part of the International Barcode of Life Database, which has widespread application in areas ranging from the food industry, where it's used for identifying food allergens and tracking foods and food labeling, to forensic entomology, in which anthropod colonization is used to estimate the postmortem interval. IBOL's Bioscan program, a seven-year effort that was slated to begin this summer, will gather specimens and study species interactions worldwide with the intent of expanding IBOL's reference library by 15 million barcode records. Other teams from around the globe are also adopting this approach to comb samples for new species in their labs or in the field. Rudolf Meier, PhD, and colleagues at the National University of Singapore are using the recently developed MinION nanopore sequencing system to catalog biodiversity in Singapore. Costing less than $1,000 and approximately the size of a smartphone, the sequencing system can be paired with a battery-powered mini polymerase chain reaction thermocycler and a laptop loaded with informatics software to allow mobile sequencing in the field. Duke University scientist Lydia Greene, PhD, and colleagues recently published a field study conducted in the Madagascar dry forest, in which they used the barcoding system on a lemur sample and decided on the spot whether it was likely to be from a new species. With the growth of species barcode databases, IBOL proposes a next-level analysis in which sequence characterization of a single specimen can disclose its commensals, mutualists, and parasites. Taken further, metabarcoding to examine the species composition of 100,000 bulk samples from sites worldwide has been proposed as a first step in compiling comprehensive regional biodiversity baselines. As sequencing costs decrease and bioinformatics optimization progresses, DNA barcoding will serve as the underpinning for global biodiversity monitoring and discovery.

Chimeno C, Moriniere J, Podhorna J, et al. DNA barcoding in forensic entomology—establishing a DNA reference library of potentially forensic relevant arthropod species. *J Forensic Sci.* 2019;64(2):593–601.

Correspondence: Caroline Chimeno at ca_chimeno@yahoo.com

Collins RA, Cruickshank RH. The seven deadly sins of DNA barcoding. *Mol Ecol Resour.* 2013;13(6):969–975.

Correspondence: Dr. R. A. Collins at rupertcollins@gmail.com

Hebert PDN, Cywinska A, Ball SL, et al. Biological identifications through DNA barcodes. *Proc Biol Sci.* 2003;270(1512):313–321.

Correspondence: Dr. Paul Hebert at phebert@uoguelph.ca

Pennisi E. DNA barcodes jump-start search for new species. *Science.* 2019;364(6444):920–921.

RNA sequence analysis identifies somatic clonal diversity across normal tissues

The accumulation of somatic mutations that transform normal cells into cancer cells is a well-established paradigm in cancer biology. While much has been learned about mutations that function as cancer drivers through the study of tumor samples, biologic understanding of the early steps in the transition of normal cells to precancerous and then to malignant cells remains elusive. DNA sequencing has been the primary tool for clinically assessing cancer, but the popularity of RNA sequencing (RNA-seq) technology has resulted in large repositories of RNA-seq data. The majority of this RNA data has been studied for gene expression, but some researchers are asking if this trove of data can be tapped to assess somatic mutations. Identifying mutations using RNA-seq is difficult because of higher false-positive rates for single nucleotide variants due to several factors, including cell cycle bias, strand bias, alignment complexity in the transcriptome, RNA editing, and random errors introduced during reverse transcription and polymerase chain reaction. In an effort to reliably identify somatic mutations from RNA-seq data, Yizhak, et al., developed RNA-MuTect, an informatics method for filtering RNA-seq data and calling somatic mutations. They initially focused on a training set of 243 tumor samples, representing six tumor types, from The Cancer Genome Atlas, for which DNA and RNA were co-isolated from the same cells. Applying their standard somatic mutation calling pipeline that was developed for DNA, the authors found that the number of mutations in RNA exceeded the number in the corresponding DNA by a factor of five. Moreover, 65 percent of the DNA-based mutations were not detected in the RNA, and 92 percent of the RNA-based mutations were not found in the DNA. RNA-MuTect, which was developed to address the excessive number of mutations detected only in the RNA, is based on several key filtering steps, including removing alignment errors using two different RNA aligners, removing sequencing errors using a site-specific error model built on thousands of normal RNA-seq data, and removing RNA editing sites using databases. When compared to the authors' DNA-developed pipeline, it filtered out 93 percent of called mutations. When accounting for the allele fractions of the DNA mutations and coverage of RNA transcripts, RNA-MuTect detected 82 percent of the sufficiently covered mutations. It retained an overall median sensitivity of 0.7 after filtering, removing as few as 10 percent of mutations that were detected in the DNA. Analysis also identified a yet-unreported mutational signature in the RNA dominated by C > T mutations. Of these mutations, 75 percent were sufficiently covered but not detected in the DNA, which suggests that this signature may reflect a C > U RNA-editing process. To further test their method across a comprehensive collection of normal tissue, the authors turned to the Genotype–Tissue Expression (GTEx) project, a collection of data generated from more than 30 normal primary tissues from hundreds of healthy people. Using RNA-MuTect to evaluate 6,707 RNA-seq samples against their matched-blood DNA, the authors detected 8,870 somatic mutations in 37 percent (2,519) of the samples, representing nearly all of the individuals studied. The skin, lung, and esophagus typically harbored the greatest number of mutations. To determine whether somatic mutations in normal tissue occur in known cancer genes, the authors tested for the frequency of nonsynonymous mutations within Cancer Gene Census (CGC) genes. They found that three percent of the samples and 33 percent of the subjects carried at least one nonsynonymous mutation in a CGC gene, with skin, esophagus, adipose, adrenal gland, and uterus tissues significantly enriched with mutations in CGC genes. The most frequently mutated cancer-associated genes were *TP53* and *NOTCH1*. The authors concluded that RNA analysis can reveal somatic variations after proper filtering and analysis of RNA-seq data. Moreover, RNA-based analysis can identify underlying mutational processes and significantly mutated genes related to cancer transformation.

Sheng Q, Zhao S, Li CI, et al. Practicability of detecting somatic point mutation from RNA high throughput sequencing data. *Genomics.* 2016;107(5):163–169.

Correspondence: Dr. Yu Shyr at yu.shyr@vanderbilt.edu or Dr. Yan Guo at yan.guo@vanderbilt.edu

Yizhak K, Aguet F, Kim J, et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science.* 2019;364. doi:10.1126/science.aaw0726.

Correspondence: Dr. Gad Getz at gadgetz@broadinstitute.org