Sifting gold from the data deluge

November 8 2017

Next-generation DNA sequencing technologies have flooded databases and hard drives worldwide with large data sets, but are researchers getting the most they can out of this deluge of data? In a new study in the October issue of *Applications in Plant Sciences*, Dr. Brent Berger and colleagues propose one way to sift the remaining gold out of large sequence data sets. The authors show that a new data mining technique can be used to glean valuable information from existing data sets, and prove the concept by retrieving sequence from genes influencing the peculiar floral structures seen in the plant family Goodeniaceae.

DNA sequencing has become so cheap that even if a researcher is only really interested in the sequence of a few genes, it is often most practical to just sequence the whole genome. Bioinformatic techniques can pick out the desired gene sequence later, with less hassle than targeting specific genes to sequence. This practice, known as "genome skimming," has become an increasingly popular way to answer questions about relationships between plant species.

The premise of genome skimming is to use low-coverage shotgun sequencing to retrieve DNA sequence from high-copy fractions of the genome. In shotgun sequencing, the genome is broken up into small chunks for sequencing, and then stitched back together computationally using the overlaps between the chunks, a process called assembly. The amount of "coverage" corresponds to how many of those small chunks are sequenced; the higher the coverage, the easier it is to stitch the genome back together, resulting in a more complete genome sequence.

But higher coverage is more expensive, and some questions can be answered with a cheaper, low-coverage sequencing run. "High-copy fractions" of total genomic DNA, such as chloroplast genomes or nuclear ribosomal DNA, are in higher abundance in the sequence pool, and so can be fully sequenced even in cheap, low-coverage runs. Sequence from these high-copy genomic fractions are typically used to resolve evolutionary relationships between different species and groups. But in the process of genome skimming, researchers produce and then discard huge amounts of potentially valuable sequence data. "Many genome-skimming data sets are used for assembling the chloroplast genome, which in our case, only used 3% of the sequenced data," remarked Dr. Dianella Howarth, a co-author on the study.

In this study, the authors took a second look at a genome-skimming data set previously used to resolve evolutionary relationships in the Goodeniaceae, a family of plants commonly called "fan flowers" or "half flowers" due to their intriguing flower shape, which looks like somebody cut the flower in half. The authors wanted to see if this genome-skimming data set could be plumbed for more information on the genetics behind this unique floral structure. They used several software packages to assemble previously unused sequence fragments from the low-copy fraction of the original genome-skimming data set. They then searched the resulting assembly for sequence from a set of genes called *CYCLOIDEA* genes, which are involved in floral structure and symmetry.

The authors were able to retrieve enough portions of the genes, from multiple species, to create full alignments of all four *CYCLOIDEA* genes in the core Goodeniaceae. These data could prove useful for future studies on the evolution of the bizarre floral structure seen in this group. "Comparing sequences from *CYCLOIDEA*-like genes across this clade could provide clues about the precise sequence changes that result in changes in floral morphology," explained Dr. Howarth.

More generally, Dr. Howarth continued, "Pieces of any gene of interest could potentially be mined from genome-skimming data sets that have already been completed." A piece of a gene may not sound like much, but there are a surprising number of uses for these fragments. "These data could provide enough information to determine useful nuclear regions for phylogenetic analyses or pinpoint possible gene duplication events. Additionally, probes for target enrichment sequencing could be generated quickly across a clade to examine candidate genes and their regulatory regions in evo-devo studies."

Data mining approaches like these allow for much fuller use of genome-skimming data sets. This allows for important questions to be answered with existing data, and opens the door to scientists without access to the resources to produce large-scale data sets—for example, scientists at smaller colleges or countries without large grant-making bodies. As DNA sequence data continue to flood in, studies such as this point to ways to make sure we don't let valuable information float by.

More information: Brent A. Berger et al, The Unexpected Depths of Genome-Skimming Data: A Case Study Examining Goodeniaceae Floral Symmetry Genes, *Applications in Plant Sciences* (2017). DOI: 10.3732/apps.1700042

Provided by Botanical Society of America