Scientists are utilizing long-read PacBio® sequencing to provide uniquely comprehensive views of complex plant and animal genomes. These efforts are uncovering novel biological mechanisms, enabling progress in crop development, and much more.

To date, scientists have published over 1000 papers with Single Molecule, Real-Time (SMRT®) Sequencing, many covering breakthroughs in the plant and animal sciences. In this case study, we look at examples in model organisms *Drosophila* and *C. elegans* and non-model organisms coffee, *Oropetium*, danshen, and sugarbeet, where SMRT Sequencing has contributed to a more accurate understanding of biology. These efforts underscore the broad applicability of long-read sequencing in the plant and animal genomics community.

With its multi-kilobase reads and uniform coverage, SMRT Sequencing addresses many challenges, like handling polyploidy and resolving long repeat regions, which cannot be solved with short-read sequence data. The platform also gathers epigenetic data as it sequences, delivering a new layer of information. The increasing affordability of SMRT Sequencing has also enabled generation of reference genomes for non-model species or multiple reference genomes across a single species.

Scientists around the world can use these tools to improve the health of crops and livestock, find alternative ways to produce medicinal components, and achieve biologically relevant results from functional and comparative genomic studies.

**Oropetium thomaeum**

Scientists recently published a high-quality, nearly complete genome of *Oropetium thomaeum*, a grass with an estimated genome size of 245 Mb. Known as a resurrection plant, *Oropetium* can survive prolonged periods of extreme drought and regrow when water becomes available. A team of scientists led by Robert VanBuren, Doug Bryant, and Todd Mockler at the Donald Danforth Plant Science Center sequenced the plant in the hopes of characterizing the biological mechanism responsible for its remarkable tolerance for drought. Eventually, this information could be used to produce crop varieties that are able to withstand severe drought and stress.

In the *Nature* paper “Single-molecule sequencing of the desiccation tolerant grass *Oropetium thomaeum*,” the scientists report a project using about 72x coverage of the *Oropetium* genome generated by SMRT Sequencing. According to the paper, that’s “equivalent to <1 week of sequencing time and <$10k in reagents.” The team used the Hierarchical Genome Assembly Process (HGAP) and Quiver algorithms for data assembly and polishing. The final assembly covered 99% of the genome in 625 contigs, with an accuracy of >99.999% and a contig N50 length of 2.4 Mb.

PacBio sequencing allowed the scientists to assemble even very challenging regions of the genome. Like many plants, *Oropetium* has extensive repetitive regions that cannot be accurately represented by short-read sequencing technologies. Indeed, this limitation is why many draft assemblies of plant genomes are highly fragmented, often splintered into tens of thousands of small contigs. Such assemblies “are missing biologically meaningful sequences including entire genes, regulatory regions, transposable elements (TEs), centromeres, telomeres and haplotype-specific structural variations,” VanBuren et al. write in the *Nature* publication.

With SMRT Sequencing, the team was able to elucidate those elements in the *Oropetium* genome, with its predicted 28,446 protein-coding genes and a significant proportion of repeat regions. For example, one large tandem array featured five identical multi-kilobase repeats and one partial repeat across a 51 kb region – a section that would have been intractable with short-read technologies but was accurately resolved with long-read PacBio sequencing.

The *Oropetium* assembly includes telomere and centromere sequence, long terminal-repeat retrotransposons, tandem duplicated genes, and other difficult-to-access genomic elements. Scientists were also able to produce the full chloroplast genome in a single contig, including “~25 kb of inverted repeat regions which typically...
collapse into a single copy during assembly,” they report.

“The Oropetium genome showcases the utility of SMRT Sequencing for assembling high-quality plant and other eukaryotic genomes,” VanBuren et al. note, “and serves as a valuable resource for the plant comparative genomics community.”

**Drosophila melanogaster**

Scientists used SMRT Sequencing data to uncover a first-of-its-kind duplication event in *Drosophila melanogaster* in which an autosomal genetic region was copied to the Y chromosome, where it acquired a new function. Based at institutes in Brazil, Austria, and the United States, the scientists published their findings in *Proceedings of the National Academy of Sciences*, detailing a study in which they used the *Drosophila* data release from PacBio to characterize a region of the Y chromosome that had never before been accessible. In the publication, “Birth of a new gene on the Y chromosome of *Drosophila melanogaster*,” lead author Antonio Bernardo Carvalho, senior author Andrew Clark, and collaborators discuss their novel findings.

“We emphasize the utility of PacBio technology in dealing with difficult genomic regions,” the scientists note. “PacBio produced a seemingly error-free assembly of the *FDY* region, something that has eluded us for years of hard work.” The 55 kb region, which includes several pseudogenes as well as the newly discovered functional *FDY* gene, has proven challenging to sequence and assemble since it exists only on the Y chromosome and is full of highly repetitive sequence. Some 75% of its length, the scientists report, is made up of transposable elements.

The paper chronicles the team’s efforts to characterize the *FDY* region using RT-PCR, clonal sequencing, and publicly available genome assemblies. Most existing assemblies did not fully cover the region. The scientists turned to the PacBio data and an assembly produced with the MinHap Alignment Process (MHAP), which provided their first view of the region’s full sequence. “Fortunately … the PacBio [MHAP] assemblies covered not only *FDY*, but also substantial flanking regions,” the scientists write. By comparing it to Sanger and Illumina sequence data, they concluded that the PacBio assembly is complete and accurate.

Unlike mammalian Y chromosomes, which are thought to evolve primarily by gene loss, the *Drosophila* Y chromosome appears to be the result of millions of years of gene gains. The team demonstrates that the new gene they detected, named *FDY* for *flagrante delicto Y*, was formed about 2 million years ago in a single duplication event of the gene *vig2* and its flanking sequence from chromosome 3R. The flanking sequence originally included four other genes, “but they became pseudogenes through the accumulation of deletions and transposable element insertions, whereas *FDY* remained functional, acquired testis-specific expression, and now accounts for ~20% of the *vig2*-like mRNA in testis,” the scientists write. Today, *FDY* shares 98% sequence identity with its *vig2* parent.

“Hence a female-biased gene (t) gave rise to a testis-biased gene (FDY),” the authors conclude. “This seems to be a case of gene duplication followed by neofunctionalization, the first reported, to our knowledge, for the *Drosophila* Y.”
RESEARCH BRIEF

Scientists used SMRT Sequencing to perform epigenetic analysis of *C. elegans*, discovering N6 methylation (6mA) along with the corresponding demethylase and putative methyltransferase responsible for trans-generational epigenetic signaling. The finding raises the exciting possibility of a molecular system for epigenetic inheritance in *C. elegans*, making it a useful research model for heritable epigenetic transmissions in eukaryotic systems.

Learn more in the *Cell* paper: DNA methylation on N6-adenine in *C. elegans*.5

**Coffea arabica**

At Cornell University, scientist Marcela Yepes is working with the global coffee research community to ensure the long-term health and viability of the world’s most valuable plant. While there are about 100 species in the *Coffea* genus, the particular strains cultivated to produce coffee – a market valued at $90 billion – have very little genetic diversity. There is a pressing need to add to the coffee germplasm now, Yepes says, to address disease and climate pressures.

As part of the International Coffee Genome Network, Yepes and her lab focus on contributing useful resources to the rest of the consortium and other scientists aiming to improve the hardiness of the coffee plant. For this project, SMRT Sequencing data were generated for the genomes of two coffee strains – the 1.3 Gb tetraploid *Coffea arabica* and its 660 Mb diploid parent, *Coffea eugenioides* – to better understand existing genetic variation and to establish high-quality reference genomes that would serve the community for years to come.6 The results were “outstanding,” says Yepes, noting that she was particularly impressed by the way SMRT Sequencing helped resolve polyploidy in assembly. For *C. arabica*, the longest read was more than 65 kb, and the genome assembly had a contig N50 of more than 700 kb, a 2.5-fold improvement over the previous assembly. More than half of the sizable genome was covered in fewer than 440 contigs. The C. eugenioides assembly had a maximum contig length of 6.1 Mb, compared to an earlier assembly with the longest contig at 187 kb. These genome assemblies will be instrumental in Yepes’ efforts to understand the functional genomics of common coffee strains. With these resources, Yepes and other coffee researchers will be able to perform targeted sequencing or genotyping to find traits of interest, helping to accelerate the positive impact of genomics on coffee plants around the world.

**References**


The International Coffee Genome Network are generating genome resources to support research improving the hardiness of coffee, ensuring the long-term health and viability of the world’s most valuable plant.