BEYOND FOUR BASES: EPIGENETIC MODIFICATIONS PROVE CRITICAL TO UNDERSTANDING *E. COLI* OUTBREAK

Studies of the *E. coli* outbreak in Germany demonstrate the fundamental need for long-read sequencing and DNA base modification data. Using SMRT® sequencing technology, scientists were for the first time able to reveal some of the complex mechanisms underlying gene regulation processes in the organism.

A newly published paper adds to insights generated in a 2011 study investigating last year's deadly *E. coli* outbreak in Germany; together, they offer a fascinating new view of the mechanisms of gene regulation in a microbe.

Dr. Eric Schadt, Chair of the Department of Genetics and Genomics Sciences and Director of the Institute for Genomics and Multiscale Biology at Mount Sinai School of Medicine, helped lead these efforts to elucidate the strain of *E. coli* responsible for the outbreak. "All these bugs living in us and around us are affecting us on a much deeper level than we've appreciated," he says. "Even in a microorganism there are complex networks at play, and being able to study these on a genome-wide scale for the first time is really causing a revolution within the microbiology community."

The PacBio *RS* sequencing platform "has the throughput to completely finish these small microbial genomes in under a day," Schadt says.

The papers are remarkable for different reasons. "Origins of the *E. coli* Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany," published in July 2011 in the *New England Journal of Medicine*, provided a sharper understanding of the particular microbial strain involved in the outbreak and cleared up lingering questions about the strain's phylogeny. "Genome-wide map of methylated adenine residues using single-molecule real-time sequencing in pathogenic *Escherichia coli*," published in November 2012 in *Nature Biotechnology*, offers a whole-genome interrogation of base modifications, elucidating that strain's "methylome." That information gives valuable clues about the gene regulation processes of the *E. coli* strain in the German outbreak.

The research projects are also a glimpse into the relatively nascent field of generating full-genome data sets for multiple dimensions — such as DNA, RNA, and proteins — and then pulling those together into a multiscale view of the organism's biology. "Living systems are composed of lots of pieces interacting in very complex ways," Schadt says. "The way to understand such systems is to take into account more of the information on a global level, not just a single protein level. That is what drives systems biology."

Dr. Eric Schadt, Director of the Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine

The Outbreak

In May and June of 2011, a highly virulent strain of *E. coli* broke out in northern Germany, affecting thousands of people. Fifty died, and nearly 1,000 suffered kidney failure from hemolytic-uremic syndrome. The severity of this particular microbe launched investigations in labs around the world to try to understand why this bug, which produced Shiga toxin, was so deadly. The outbreak strain serotype was O104:H4, which previously had not been associated with such pathogenicity.

Scientists from a number of organizations sequenced outbreak samples on next-generation DNA sequencers and contributed their data to public repositories. But with short-read sequencing platforms, the genome assemblies that became publicly available were fragmented into more than 300 pieces, known as contigs, making it no trivial matter to identify the phylogenetic basis of the microbe. Add to that a starting data bias — several sequenced genomes were available for one serotype of *E. coli*, but just a few from the O104:H4 serotype — and it's no wonder that early attempts to determine the origins of the outbreak strain wound up being incorrect.

Eric Schadt, then Chief Scientific Officer at Pacific Biosciences, and his colleagues saw an opportunity to help the community make inroads into understanding the outbreak strain by adding long-read sequence data to the short-read data already available. Teaming up with scientists at medical schools and the World Health Organization, they launched their investigation, using the PacBio® *RS* platform to sequence not just the German strain but also several other *E. coli* strains from the underrepresented serotype to pinpoint the outbreak strain's origins. Their *NEJM* paper reports 13 genome sequences, including an assembly of the outbreak strain that used the long PacBio reads to get down to just 33 contigs. Combining PacBio data with existing short-read data, the team was able to represent the full *E. coli* genome in a single contig.

The PacBio *RS* sequencing platform "has the throughput to completely finish these small microbial genomes in under a day," Schadt says, noting that this project would not have been possible on any other sequencer currently available. That speed was critical to the team's efforts and points to real-time utility of the sequencer for outbreaks and other situations where information is needed urgently.

A key finding from the team's investigation was that the widely reported origin of the strain — that it was an enterohemorrhagic strain which had acquired enteroaggregative properties through horizontal gene transfer — was not quite right. "What our work showed is that it was exactly the opposite — it was definitely an enteroaggregative strain, and the way we could tell that was that we sequenced nine other strains of the same serotype. They matched perfectly to the outbreak strain," Schadt says. "It was a pretty fundamental difference, but these bugs are tricky."

The team reported that the outbreak strain had acquired enterohemorrhagic traits, including the relatively recent insertion of a Shiga-toxin–encoding lambda-like prophage element. The strain had also developed a devious form of antibiotic resistance: the standard treatment of ciprofloxacin actually boosted the expression of the bug's Shiga toxin gene, leading to much higher levels of the toxin, which was partly responsible for the onset of hemolytic-uremic syndrome in patients. In a real-time setting, this information could be critical to selecting the best possible treatment for a particular microbial strain.

Studying Adenine Methylation

Despite the remarkable findings from the team's *E. coli* sequencing effort, the DNA sequence alone did not fully explain the unusually high virulence seen in the outbreak. Some months after the initial sequencing work was done, Schadt and his colleagues had an opportunity to reanalyze the data — this time looking at chemical modifications to DNA bases.

This layer of information was of particular interest to Schadt and his research team because of the growing recognition during the last ten years that methylation can play a significant role in microbial virulence and other biological processes.

"If you want to understand why this bug is so much more virulent than these other bugs that have the same serotype, if you don't have all of the axes of variation that can be happening, then you might miss answering that critical question," Schadt says.

Analyzing these epigenetic marks would not have been possible on any other sequencer. Other sequencing platforms use DNA amplification which strips away base modifications prior to the sequencing process. But a unique aspect of single molecule, real-time (SMRT) sequencing is that it collects base modification data simultaneously as it collects DNA sequence data. Scientists at Pacific Biosciences have demonstrated that the polymerase used in their sequencing technology responds differently to modified bases than it does to unmodified bases, leaving a kinetic signature that can be detected in the data. Indeed, the polymerase is sensitive enough that it pauses at measurably different intervals for different types of modifications, leading to distinct kinetic patterns for more than a dozen different types of base modification.

"Some of the base modifications that we're detecting appear to be associated with virulence, but that's virtually unknown because people haven't been able to see it before."

That array of modifications represents a far larger epigenetic universe than could be assessed with earlier tools, Schadt says. "Some of the modifications that we're detecting appear to be associated with virulence, but that's virtually unknown because people haven't been able to see it before."

The computational ability to analyze this data became possible after Schadt and his colleagues had completed their first examination of the *E. coli* outbreak strain, but because the base modification data was an inherent part of the sequence data, they were able to go back after

the fact and reanalyze the same data set to detect base modifications genome-wide.

While some of the software to perform that detection was just being crafted at the time of Schadt's study of *E. coli*, the tools have since been incorporated into the standard software offered with the PacBio instrument. Any user can detect base modifications on new bacteria being sequenced, or reanalyze older sequences gathered by the PacBio *RS*.

The Epigenetic View

By analyzing the outbreak strain sequence for base modifications, in this case N^6-methyladenine residues, Schadt and the team discovered a series of methylase enzymes that appeared to target specific sequence motifs throughout the genome as they made their chemical changes. For example, Dam methyltransferase targeted the A residue in DNA with the sequence motif GATC, while a methyltransferase found in the Shiga toxin region acts on the CTGCAG motif. "We found a whole array of methylase-like enzymes that were making modifications by targeting different motifs," Schadt says. "It was almost like a language."

Many of those enzymes had not been previously characterized, Schadt notes. But there seems to be little question about the importance of the motif-targeting patterns the team found: "They're highly non-random, and the targeting had an effect on the transcription of genes," he adds.

The team followed up the base modification study with RNA-seq to determine how these marks were affecting the transcriptome. "Connecting the RNA sequence to the epigenetic changes at this point isn't quite seamless," Schadt acknowledges. His team relied on age-old biology techniques to examine the functional consequences of base modifications, including knocking out the gene making the changes and then sequencing the RNA of the knockout alongside the RNA of the wild type. In other work, they compared the K12 strain with added methylations to the original K12 to study the effect of those methylations. Data was sifted to show expression signatures, pathway enrichment, and more.

> "Without base modification information, you simply don't have a complete picture of all the variation and the phenotypic variability that we see."

"The dogma in the field for these restriction modification systems is that they're protecting the DNA of the bug from degradation when they release the endonuclease to degrade other DNA," Schadt says. "But what we found was that these modifications were having a very significant impact on the transcription of genes, and that the genes being affected were enriched in a number of different pathways." Notably, they found marked enrichment for pathways linked to horizontal gene transfer in the outbreak strain. Throughout the organism's genome, many pathways were up- or down-regulated by one of the methylases found in a mobile element next to the Shiga toxin gene, which is known to have an impact on virulence. The removal of this methylase in a knockout experiment led to structural genomic changes that may indicate involvement with processes associated with pathogenicity and virulence.

For Schadt, these findings emphasized the importance of adding more layers of data to the picture of any organism's biology. He notes that the microbiology community in particular is seeing a revolution with this kind of information and "turning on whole areas of investigation that weren't considered in the past." Being able to generate DNA sequence and methylomes for a microbe on a genome-wide scale in less than a day is a major step toward filling in the picture of how these organisms function.

"The A's, G's, C's, and T's that you get from sequencing are a big piece of the puzzle of putting a genome together — that's why it is so important to have long reads for a complete genome assembly," Schadt says. "But beyond that, those bases can be chemically modified, changing how proteins interact with that particular sequence and having significant functional consequences. Without base modification information, you simply don't have a complete picture of all the variation and the phenotypic variability that we see."