

# How Did a Deadly Bacterium Evolve *in vivo* during an Epidemic?

Systems biologists embrace whole-genome sequencing and use Illumina<sup>®</sup> technology to uncover patterns of adaptive evolution that bolster bacterial pathogenesis.

#### Introduction

Genetic mutations can undo an antibiotic's effectiveness against bacteria, resulting in the emergence of antibiotic-resistant superbugs that make it difficult to eradicate an infection. In the laboratory, researchers use genotyping arrays to monitor genomic changes as the bacteria evolve, but these studies only reveal a fraction of the bacteria's genome and don't completely replicate what is occurring *in vivo*. Wanting a more comprehensive look, Tami Lieberman and Jean Baptiste Michel, both graduate students at the time in Dr. Roy Kishony's laboratory in the Department of Systems Biology at Harvard Medical School, took advantage of Illumina next-generation sequencing (NGS) technology to view parallel bacterial evolution on a genome-wide scale<sup>1</sup>.

Their first challenge was finding an appropriate sample set to analyze. They began by asking hospitals in the Boston area for urinary tract infection samples, feeling that if they had blood and urine samples from the same person, at specific points in time, that they would be able to track the infection as it moved into the bloodstream. During their search, they spoke with Alexander McAdam M.D., Ph.D., Director of the Infectious Diseases Diagnostic laboratory at Children's Hospital Boston and Gregory Priebe, M.D., Pediatric Critical Care and Infectious Disease Specialist at Children's and a microbiologist at the Channing Laboratory of Brigham and Women's Hospital, who offered an alternative. They had a collection of bacterial isolates from a rare pathogen outbreak in the community's cystic fibrosis (CF) patient population. Taken over a 16-year period, the samples could provide a window into the bacteria's evolution.

iCommunity spoke with Ms. Lieberman to learn how NGS and this remarkable data set enabled them to gain a more comprehensive understanding of bacterial genetic adaptation during pathogenesis.

### Q: Why hasn't there been more research work investigating within-patient bacterial evolution of pathogens?

Tami Lieberman (TL): Until now, it has been extremely challenging to observe within-patient evolution. Investigating it requires finding the changes that occur among millions of letters of genetic code. But classical ways of comparing bacterial isolates relied on comparing only a few genes, and those methods found few or no changes. You need to look at the whole genome to see the changes that occur, and NGS technologies like Illumina's have only recently enabled this.

### Q: Was this the first time that whole-genome sequencing was used to investigate within-patient bacterial evolution?

TL: No, the first such study to my knowledge was in 2006, using the laborious Sanger sequencing method<sup>2</sup>. Since then, a few studies have been performed using Sanger or next-generation sequencing. In these studies, bacterial evolution was detected by simply taking single isolates from the same person at different time points, sequencing the isolates at the whole-genome level, and picking out the few differences that happened in the millions of base pair genomes.

A large challenge faced in these studies was that it was hard to tell which mutations were adaptive, meaning responsible for driving evolution. With the cost of NGS dropping, we realized we could use Illumina whole-genome sequencing technology to investigate within-patient evolution in many closely related infections. We reasoned that if certain mutations are adaptive, they might occur in parallel in multiple independent patients.



Researchers in Roy Kishony's lab at Harvard Medical School used Illumina sequencing technology to view bacterial evolution on a genome-wide scale. From left to right: Tami Lieberman, Graduate Student; Jean Baptiste Michel, Ph.D., Foundational Questions in Evolutionary Biology Postdoctoral Fellow; Roy Kishony, Ph.D., Professor of Systems Biology.



Bacterial isolate samples from the community's CF patient population were donated by Alexander McAdam, M.D., Director of the Infectious Diseases Diagnostic Laboratory, Children's Hospital Boston (left) and Gregory Priebe, M.D., Pediatric Critical Care and Infectious Disease Specialist, Children's Hospital Boston (right).

#### Q: The bacterial isolate collection was from an infectious outbreak of Burkholderia dolosa among Boston-area CF patients. Can you tell us about the outbreak?

TL: The outbreak started in the early 1990s. *B. dolosa* is more deadly than the more common CF infections, such as *Pseudomonas aeruginosa*, and it can be spread from person to person. It is part of a group of related species called the *Burkholderia cepacia* complex. Initially, it wasn't appreciated that *B. dolosa* was a distinct species, and that being infected with a different *Burkholderia* strain didn't mean you couldn't be infected with the more harmful *B. dolosa*. As a result, the infections spread among Boston's CF patient population. Now patients with *B. dolosa* are isolated from other CF patients, and the outbreak has been completely controlled. This isolation can be tough because CF patients and their families are part of a community that supports each other.

"You need to look at the whole genome to see the changes that occur, and NGS technologies like Illumina's have only recently enabled this."

#### Q: Why was this outbreak so devastating?

**TL:** Lung infections are difficult to treat in CF patients, in part because you can't get high enough concentrations of antibiotics into the lungs. In addition, *B. dolosa* is highly resistant to antibiotics. In CF patients, *B. dolosa* infection can cause life-threatening "cepacia syndrome"—a rapid health decline characterized by fevers, necrotizing pneumonia, and bacteremia. Ultimately, 39 patients in the Boston CF community were infected with the bacteria, and 16 of them died.

### Q: What made this sample collection ideal for your whole-genome research?

**TL:** We were looking for samples that had been taken from the same patients over a period of time. This collection existed because initially no one knew what was going on with this infection. When the outbreak began, *B. dolosa* wasn't recognized as a distinct species. Thanks to the foresight of the doctors at Children's Hospital, more than 16 years of bacterial isolates from these patients were frozen and stored. Because these isolates were taken from an outbreak, we knew that the genomes would be very closely related, simplifying our analysis. It was the perfect model system for us to analyze.

#### Q: Which patient samples did you include in your research?

**TL:** We chose to study 14 de-identified patients, including patient zero of the epidemic. We had one out-group sample from a patient who was not from the epidemic that allowed us to infer the genotype of the ancestor of the epidemic and enabled us to root the phylogenetic tree. Seven patients were included on the basis of availability of isolates from the blood. The remaining six were chosen at random among infected patients who did not have bacteremia.

#### Q: The sequencing was performed on a Genome Analyzer™ system at the Partners Center for Personalized Genomic Medicine. How difficult was it to prepare all of those samples?

TL: We made a library copy of each of the bacterial samples and performed the sample prep here. Using the Nextera® DNA Sample Prep Kit, Jean Baptiste and I were able to extract DNA and prepare over 100 samples in just two weeks. The Nextera kit made sample prep quite easy. It would have taken much longer any other way.

### Q: What were the results of the whole-genome sequencing and SNP calling?

TL: Reads were aligned using the *B. dolosa* genome belonging to an isolate recovered from patient zero, previously sequenced by the Broad Institute. Alignment was performed using Illumina's CASAVA software, with SAMtools used to identify SNPs between isolates. On average, genomes were read to a depth of 37×, providing high-quality reads for 93% of the genome. We identified 492 polymorphic loci that differed between at least two isolates of the epidemic. These SNPs informed us about how the outbreak spread from person to person, how these infections progressed from organ to organ, and identified parallel evolution, mutations that these bacteria commonly acquire in independent infections.

# Q: Was the bacteremia seen in some of the patients facilitated by a specific gene mutation?

**TL:** We thought we might find one or more mutations that enabled bacteria to escape into the bloodstream. But rather than finding blood isolates generated by a single clone in the lung, we found diverse blood isolates that were related to different lung isolates. That suggests there's no 'super bacteria' responsible for bacteremia. Instead, it's caused by the lung's failure to prevent *B. dolosa* from crossing the alveolar–capillary barrier. We're not sure if this transmission occurs through a single event at one lesion or over time, perhaps at multiple locations.

### "The Nextera kit made sample prep quite easy. It would have taken much longer any other way."

#### Q: Does this finding have implications in clinical decision making?

**TL:** Yes. Lung transplants are often performed to extend the lives of CF patients. For those infected with *B. dolosa*, replacing the lungs might solve the potential leaking of bacteria through lung lesions, but it would be less effective in stemming the proliferation of bacteria that have become well adapted to the bloodstream.

# Q: Can the mutations be tracked to show the transmission path of the bacteria?

**TL:** We found that, in most cases, all the bacteria from a patient share a few mutations in common, creating a genetic fingerprint. We can track that fingerprint through a number of patients, enabling us to recreate the transmission network. In the future, this type of genetic fingerprinting could help epidemiologists determine where an infection started, whether it was acquired in the hospital or from the environment, and who infected whom. This can be done today, but at much lower resolution. With whole-genome sequencing, we can determine the transmission network with more confidence.

#### Q; When you assayed the B. dolosa samples for resistance to ciprofloxin (a fluoroquinolone), you found 112 isolates that varied over two orders of magnitude in their resistance profiles. Did you expect this?

TL: Variation in resistance profiles over time has been seen many times before. Recently it's been noted in *P. aeruginosa* that isolates taken from the same person at the same time can have very different antibiotic susceptibilities. Fluoroquinolones exert a strong selective pressure and there is a lot of genetic variation provided by the large bacterial populations in the CF lung, which means that the probability of finding resistance is quite high. What surprised us is that so much of the antibiotic resistance could be accounted for by mutations in only one gene, a homolog of *gyrA*. Additionally, we were surprised that these mutations arose *de novo* within each patient—we even found evidence in two cases for ciprofloxacin resistance evolving multiple times within the same person.

#### Q: You also looked at mutations in O-antigen presentation. What is the O-antigen responsible for and what is its role in pathogenesis?

TL: The presentation of O-antigen repeats in the lipopolysaccharide (LPS) layer of the bacterial outer membrane is known for its role in virulence in related species. It's been thought that the O-antigen, which is composed of repeating units of complex sugars, confers some sort of protection for bacteria, such as preventing the immune system from recognizing it.

In our study, mutation of a single nucleotide in the glycosyltransferase gene BDAG\_02317 correlated exactly with the presentation of O-antigen repeats. Unlike what we saw for ciprofloxin resistance, there is a reason why O-antigen presentation is controlled by one gene. The ancestor had a stop codon at this locus and in order to restore function of that gene, the bacteria needed to mutate. We found that gain-of-function mutations at the BDAG\_02317 gene had occurred independently in nine subjects.

What was interesting was that some of the *B. dolosa* isolates had O-antigen while others did not. Based on the transmission network we constructed using phylogenetic methods, the evolutionary history suggests that the ancestor of the epidemic, as well as the initial colonizer of each patient, lacked the O-antigen. We hypothesize that the O-antigen's large hydrophilic chain somehow gets in the way when a bacteria is trying to establish infection in a new patient, but confers an advantage once the infection is established. The mutations in the BDAG\_02317 gene restore this sugar coating, enabling long-term survival in the body.

### Q. In addition to the gyrA gene and BDAG\_02317, did you find any other genes with multiple mutations?

**TL:** Yes. We found in total 17 genes that acquired three or more mutations over the course of the epidemic. This pattern of common mutation is called parallel evolution, and our data suggests that parallel evolution can be used to identify adaptive evolution in the body. These 17 genes had 18 times as many non-synonymous mutations as would be expected by neutral drift, suggesting they are under the pressure of natural selection *in vivo*. Genes with one or two mutations per gene didn't have this signature.

### "With whole-genome sequencing, we can determine the transmission network with more confidence."

Numerous mutations found in the *gyrA* gene and several glycosyltransferases in the O-antigen cluster stress the importance of these pathogenic pathways in lung infections. The most mutated gene, a homolog of *fixL*, had 17 non-synonymous mutations and is linked to oxygen-dependent gene regulation. Eleven of the 17 genes encode proteins involved in membrane synthesis, secretion, and antibiotic resistance.

While all of these genes are important to *B. dolosa* survival in the body, we can identify only a few that are involved in antibiotic resistance. We're unaware of the roles that 6 of the 17 play in lung infections, making them potential targets for new therapeutics. I think one of the interesting stories still to come from this research concerns the role that each of these genes plays in the pathogenesis of lung infections.

## Q: What are the potential clinical applications of genome-wide sequencing infectious outbreak patient samples?

**TL:** As the cost and time required for DNA sequencing decrease, variations on this approach can be used to track an outbreak in real time, or to follow evolution of infection in single patients in real time. We may find that such parallel evolution also occurs during outbreaks of infections with shorter courses. In this way, sequencing will enable the quick identification of the weaknesses of particular bacterial strains, and may enable doctors to predict and prepare for mutations before they occur.

One caveat is that our approach required a lot of samples over a period of time—we needed to see the same mutations happening over and over again in order to determine instances of adaptive evolution. Sample collection and storage will be important in conducting these studies in the future. If methods can be designed that can extract more information from fewer samples, the implications will be far-reaching.

## Q: What were the advantages of using the Genome Analyzer for this study?

TL: Besides the cost, one advantage was the availability of standardized tools we could use to analyze the Illumina data. Previously, I was an experimental biologist. While both Jean Bapstiste

and I have quantitative and programming experience, we came into this project with minimal exposure to sequencing and bioinformatics. Yet, we were still able to get some really interesting information out of the data.

### Q: What are the next steps for the research team and your work in particular?

TL: The Kishony lab is using whole-genome sequencing to look at the genomic basis of antibiotic resistance both in the lab and in the clinic. We are actively using genomic sequencing on diverse projects including some sequencing using the MiSeq<sup>®</sup> system.

I am continuing to use NGS technologies to investigate genetic adaptations during pathogenesis, this time applying these techniques to more common infections. We are also assessing the population diversity within a single patient that this evolution generates, in both pathogenic and commensal bacteria.

#### References

- Lieberman TD, Michel JB, Aingaran M, Potter-Bynoe G, Roux D, et al. (2011) Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet 43: 1275–1280.
- Smith EE, Buckley DG, Wu Z, Saenphimmachak C, Hoffman LR, et al. (2006) Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. Proc Natl Acad Sci U.S.A. 103: 8487–8492.

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

#### FOR RESEARCH USE ONLY

© 2012 Illumina, Inc. All rights reserved.

Illumina, illuminaDx, BaseSpace, BeadArray, BeadXpress, cBot, CSPro, DASL, DesignStudio, Eco, GAllx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, Sentrix, SeqMonitor, Solexa, TruSeq, VeraCode, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 070-2012-003 Current as of 11 April 2012

