Corrected: Publisher Correction

THE AUTHOR FILE

Sündüz Keleș

Moving athletically, from Turkey to California to Wisconsin and from engineering to the biostatistics of Hi-C.

"I'm not a biologist, but I love biology," says Sündüz Keleş, who is on the faculty in biostatistics and medical informatics at the University of Wisconsin School of Medicine and Public Health.

At Bilkent University in Turkey, Keleş had studied industrial engineering. She applied to a number of graduate engineering programs but only one biostatistics program, the one at the University of California, Berkeley. "California was just so alluring," she says. At the time, Keleş thought: "If I don't like it, I can switch back to engineering; but of course I loved it," she says. Her mother teases her that she headed to California to be as far away from her as possible.

A statistician at Bilkent University had introduced Sündüz Keleş to the field with a project involving statistical modeling and survival analysis. She continued this type of work during her PhD research with UC Berkeley biostatistician Mark van der Laan. When "the big buzz" began about the human genome project, she ran to buy the issues of *Nature* and *Science* with the papers on the draft human genome. "I remember being so excited about it," she says. "And I said, 'I want to work on genomics.'"

After completing her PhD, she landed her current faculty position. But she put that on hold for a year to be a postdoctoral fellow at UC Berkeley with van der Laan and statistician Sandrine Dudoit. There, she worked on statistical methods and software for analyzing ChIP-chip data, from experiments involving chromatin immunoprecipitation combined with DNA microarray analysis.

Keleş is "clearly an outstanding scholar," says Dudoit, who considers Keleş a collaborator and friend. To those who know her personally, "she is also an extremely vibrant, warm, and funny person," she says. She fondly recalls cooking sessions in Berkeley, explorations together in Istanbul and Keleş taking her through her first steps as a jogger. "She is a bundle of energy and a fitness enthusiast; I remember her teaching me about 'burpees' and 'kettlebells,'" she says.

Sündüz Keleș. Credit: M.F. Keleș

Keleş and her graduate student Ye Zheng have developed FreeHi-C (fragment interactions empirical estimation for fast simulation of Hi-C data), a way to generate simulated Hi-C data. "We can generate data that really looks like real Hi-C data," she says. In high-dimensional genomics, "we don't have real ground truth," says Keleş, which heightens the need to mimic actual data well so as to compare different methods and delve into one's data.

The new method grew out of her interest in gene regulation, which can be due to influences near and far, as chromatin changes its configuration. Many computational methods for analyzing Hi-C data exist, but they are not uniformly benchmarked and evaluated, which encumbers reliable inferences about these regulatory influences.

When labs perform Hi-C, they obtain raw reads, summarized in a matrix of chromosomal interactions. "For most of the methods, just having those matrices is good enough," she says. One data analysis challenge labs face is a dearth of replicates. When they simulate their matrices, their dataset may not be quite realistic.

The FreeHi-C simulated data include reads with random mutations and indels, and the software 'learns' its parameters from the actual data. Free Hi-C emerged when Keleş and Zhang worked with Hi-C reads and sought a way to simulate the data at the read level with all the nuances of actual data. The tool is available from her lab's website and will soon also be in a portal for data analysis tools that she is not ready to name.

She hopes labs will find the method useful. For example, a lab might have pilot data from an experiment with two conditions and only one replicate from each condition and seek to identify how regions of the genome are interacting in these two conditions. But in this instance, estimation can be unreliable and the false discovery rate hard to control.

FreeHi-C can give that lab "free replicates," she says. They might not capture all biological variation but can capture the experiment's technical variation, which increases both the accuracy and statistical power of an analysis. The tool can also help labs determine how deep to sequence. Scientists can "down-simulate or up-simulate" to see the types of chromosomal interactions lost or gained at different sequencing depths.

Hi-C data can be generated from single cells, but it's still challenging. Keleş sees how FreeHi-C could help to capture the heterogeneity in single-cell datasets. "So that's looking a little bit into the future—not too distant."

"I'm not a biologist, but I love biology."

Keleş's projects develop from her interactions with biologists, for whom she might tweak off-the-shelf methods or develop tools where none exist. It's a "very collaborative campus," she says of the University of Wisconsin.

She encourages members her lab to be part of these collaborations. "It's a little bit like matchmaking," she says: matching the right person to the right real-world problem. Statisticians might be tempted to work on theoretical problems, but "by talking to biologists you really see their perspective, see what they think is the real significance," she says.

When she is not in the lab, she spends time with family, which includes three young daughters. As long as it's above zero degrees Fahrenheit, she goes for runs. "Or I'm in the gym," she says. "For this phase of my life, these are the activities I sample through." ❐

Vivien Marx

Published online: 13 December 2019
https://doi.org/10.1038/s41592-019-0696-0

Reference

Zheng , Y. and Keleş, S. FreeHi-C simulates high-fidelity Hi-C data for benchmarking and data augmentation. *Nat. Methods* https://doi.org/10.1038/s41592-019-0624-3 (2019).