THE AUTHOR FILE

Stein Aerts

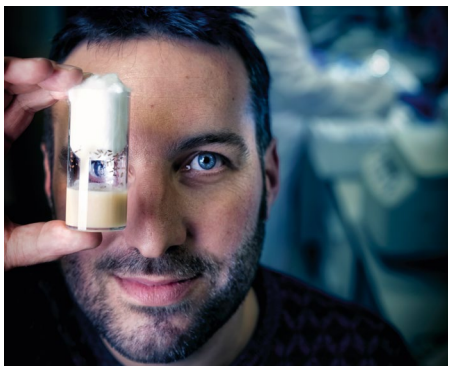
His enhancer-discovery tool cisTopic is both wet-lab and dry-lab. John Coltrane matters, too.

"It's still me," says Stein Aerts about his diverse science chapters to date. He is on the faculty at KU Leuven's department in human genetics, and he's in neuroscience as the head of the VIB Center for Brain & Disease Research. In high school Aerts was a math nerd who taught himself coding, and in college he turned to agricultural engineering, including a stint in Mexico and research on nitrogen fixation in bean plants, then switched gears to join a biotech company as a programmer, completed his PhD in bioinformatics with Yves Moreau at KU Leuven in Belgium, and spent his postdoctoral fellowship studying fly genetics and neuroscience with Bassem Hasan, who was then in Leuven.

Gene regulation is the guiding motif in his work, says Aerts. The latest tool from his lab is cisTopic, which helps labs pull findings out of their single-cell assay for transposase-accessible chromatin with high-throughput sequencing (scATAQ-seq) data. With scATAC-seq, "in every cell you read something different," he says.

Soon after uploading the tool to GitHub and the paper to bioRxiv, Aerts and his team began receiving happy e-mails. People wrote that they could finally discern cell types and determine enhancers and promoters from their fragmentary data, too. On a single-cell level, all open chromatin locations have an equal opportunity to be measured. Each one is either occupied by a transcription factor or not. But scATAQ-seq analysis presents a conundrum: you need to cluster the cells by cell type to build an ATAC profile, but you need the right combinations of peaks from genomic regions to determine cell types.

Current tools use different ways of clustering scATAC-seq data that are not quite optimal because, for example, some tools draw on pre-existing knowledge for the clustering, he says. Together with graduate students Carmen Bravo González-Blas and Liesbeth Minnoye, he worked out how to leverage topic modeling, and borrowed Bayesian inference techniques from natural language processing to do so. Textual analysis is about probability estimations: software can analyze a text corpus, determine topic distribution and the importance of each term to a topic, then use these probabilities to cluster documents. In this vein, says Aerts, a single cell is a bit like a document of many



Stein Aerts. Credit: L. Daelemans

words, and a word can, for example, be a regulatory region that is either observed or not. Mathematical topic modeling unearths underlying topics in the data, and each topic has a number of regulatory regions associated with it, which scientists can use for regulatory motif discovery.

"Beautiful, now we have a topic matrix," he says of the way it felt when they realized the method worked. "We can cluster the cells on the topics and we can cluster the regions on the topics." Enhancers and promoters are a kind of genomic grammar that emerges from studying these data. The team applied cisTopic to study how chromatin accessibility changes when a transcription factor in melanoma cells is perturbed. Separately, they analyzed transcriptomic data from brain tissue and its typical mix of different cell types.

It was Bravo González-Blas who decided to try latent Dirichlet allocation, the successor to latent semantic indexing in text mining, and found it could handle the kind of sparse data scATAC-seq delivered. She added collapsed Gibbs sampling, a technique to optimize the topic models. Minnoye produced the ATAC-seq data, designed the microfluidics, was cisTopic's first user and worked with Bravo Gonzalez-Blas to iron out the method's kinks. They are both wet-lab and dry-lab scientists. The team hopes to keep developing methods, for example, ways to integrate the topics they find with larger datasets about regulatory networks. Aerts fosters collaboration. "Each student and postdoc usually has one collaboration," he says. "Stein is a biologist's bioinformatician," says Casey Bergman of the University of Georgia, who was a virtual postdoc mentor to Aerts. Aerts deeply understands the molecular biology of gene regulation, and his rigorous engineering background lets him develop timely computational tools to solve the field's hard problems. Bergman calls Aerts an archetypal version of an "antedisciplinary scientist," a term Harvard University researcher Sean Eddy coined. Aerts "defies disciplinary boundaries by combining expertise from multiple fields in a single mind, and it's this rare integrative talent that is propelling his group to the forefront of genomics and computational biology."

> "We usually follow our passion and excitement."

Aerts is sometimes teased about having spent his entire career in Leuven. "Scientists have to be shaken up regularly," which, to some, means switching locations for their training. "I think it was cool to show that it's possible in other ways." He has switched fields and techniques plenty. "We usually follow our passion and excitement," he says. To not interrupt his wife's career trajectory and to raise their three sons in one country, Aerts stayed in Leuven. Moving a lot can be hard on children, he says. Together they travel plenty, also on "adventurous holidays" such as to Peru and Iran. In Europe, he is happy to be able to balance life and work and is surprised to not see more Americans apply for academic positions there. Aerts enjoys the jazz of Miles Davis and John Coltrane. He is particularly fond of a Brazilian-Belgian choreography duo's rendition of Coltrane's "A Love Supreme," in which the dancers themselves seem to be the instruments, he says. "This is really spectacular." ❐

Vivien Marx

Published online: 10 April 2019
https://doi.org/10.1038/s41592-019-0397-8

Reference
Bravo González-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods*, https://doi.org/10.1038/s41592-019-0367-1 (2019).