Searching for Cancer Driver Gene Expression Clues in Cell-Free DNA

Whole-genome sequencing with the MiSeq® and NextSeq® Systems enables researchers to examine nucleosome patterns and infer the gene expression status of cancer driver genes.

Introduction

The nucleosome is a tiny, elegant piece of DNA packaging. Like beads on a long genetic necklace, each nucleosome contains a small turning section of DNA, wrapped around eight specialized proteins called histones. As it turns out, the nucleosome might contain much more than that. It could harbor clues about which genes are driving metastatic cancers.

In 2016, a team from the University of Washington analyzed maps of nucleosome occupancy in cell-free DNA (cfDNA) of healthy individuals.[1] These *in vivo* occupancy maps correlated highly with the architecture and organization of expression seen in single cells. This suggested there could be an epigenetic footprint present in the nucleosome from the tissue of origin. Study authors argued that such a footprint could be used to determine the different cell types contributing to disease when there were no other obvious genotypic differences. Studying the nucleosomes present in cfDNA might also provide new ways to detect cancer using noninvasive screening, such as liquid biopsies.

At the same time, a team in Professor Michael Speicher's laboratory at the Institute of Human Genetics at the Medical University of Graz was also investigating what biologically relevant information could be found in cfDNA. Because cfDNA is nucleosome-protected DNA, they focused on whether functional data could be found in how the cfDNA was packaged. They were right, demonstrating that patterns of nucleosome occupation at transcription start sites provide information to infer the expression of cancer driver genes.[2]

iCommunity spoke with Professor Speicher's colleagues, Ellen Heitzer, PhD, Head of the Liquid Biopsies for Personalized Medicine in Cancer Research Unit, and Peter Ulz, MSc, a bioinformatician. They discussed their nucleosome study and how liquid biopsy data analyzed using their algorithm might provide researchers with important information about cancer progression and treatment response.

Q: What is the disease focus of your research unit?
Peter Ulz (PU): Our institute has 70 employees and is headed by Professor Michael Speicher. Most of the staff perform routine diagnostic assays for hereditary diseases, but we also conduct research into various liquid biopsy applications.

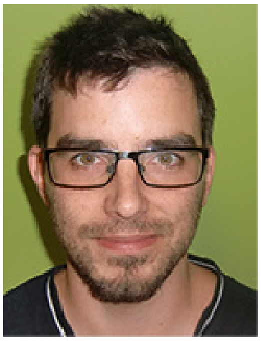
Ellen Heitzer (EH): My main interest is cancer research. I am involved in both routine diagnostics of familial cancer syndromes and research focusing on genetic analysis methods for liquid biopsies. For more than six years, we´ve worked on methods that enable a noninvasive, longitudinal monitoring of cancer genomes from plasma.

In the past, we studied single, circulating tumor cells. We now favor the analysis of ctDNA. It's a better marker, covering the tumor more comprehensively than a single circulating tumor cell.

Q: What is the value of liquid biopsies?
EH: A traditional tissue biopsy provides a snapshot in time of the tumor. However, it's difficult to perform a tissue biopsy every four weeks to see how the tumor has changed. In contrast, we can perform a liquid biopsy easily every few weeks. As a result, liquid biopsies enable us to follow a subject through the entire course of a disease, from diagnosis, to treatment response, to cancer recurrence.

The liquid biopsy research field is dynamic and we're making rapid progress.

Q: From a bioinformatics standpoint, what are the challenges of analyzing DNA sequencing data from liquid biopsy samples?
PU: Many people find the sheer amount of data from a sequencing run challenging to analyze. That's not the most difficult issue in our lab because we are performing small-scale sequencing. Our challenge is the very low signal that we obtain from ctDNA sequencing. Depending on the tumor stage and various other parameters, cfDNA from normal cells can outnumber the tumor-derived DNA significantly. It is difficult to obtain good information from the tiny amount of DNA in general and the even tinier tumor fraction, while eliminating the biases that happen along the way.



Ellen Heitzer, PhD is Head of the Liquid Biopsies for Personalized Medicine in Cancer Research Unit, and Peter Ulz, MSc is a bioinformatician in her group at the Institute of Human Genetics at the Medical University of Graz.

Q: What sparked the idea to study patterns of nucleosome occupancy at transcription factor binding sites in cfDNA?
PU: Because cfDNA is nucleosome-protected DNA, we initially had the vision that there must be some more information within liquid biopsy samples hidden in the way cfDNA is packaged. Upon reading several publications about chromatin remodeling and MNase-Seq,[3] we envisioned this to be applicable to cfDNA. We had data from low-coverage whole-genome sequencing (WGS) and looked to see whether there was a signal type that could indicate whether a gene was expressed or not. We then reviewed sequencing data of more than 100 controls to see if there were similar signals. Those results were encouraging and we used them to build this study. The University of Washington study, which exploited the same signal to deduce tissue-of-origin, further added to our excitement.

> "We learned that the nucleosomes at the transcription start sites show different coverage patterns for genes that are expressed and those that are not."

Q: Why did you choose to use WGS instead of whole-exome sequencing (WES) for this novel method?
PU: The nucleosome occupancy signals are located at the very start of the transcription site. Most exome sequencing kits enrich the coding parts of the genome and not the untranslated region, which is where we obtain the highest signal. Even if the exome enrichment analyses would enrich the untranslated region, the part shortly before the untranslated region would be lost. Unfortunately, this is the region where we find the expression information.

Q: What were the results of your study?
PU: We used WGS to analyze the cfDNA found in the plasma of cancer subjects. During gene expression, a nucleosome is removed at the transcription start site. As cfDNA is released into the bloodstream when cells undergo apoptosis, this nucleosome depleted region (NDR) is preferentially digested by nuclease because it is not protected by histones. This means that DNA fragments of inactive genes, where no nucleosome is removed, are more abundant in the circulation than active genes. We showed that nucleosomes at the transcription start sites show different coverage patterns for genes that are expressed and those that are not. This pattern can be established from sequencing data. Using a machine learning algorithm, we can predict what genes might be driving the disease with high accuracy.

The most interesting aspect of the study is that the method works in healthy individuals and cancer subjects. We can infer quite a bit about what genes are expressed. So far, we have sequenced two ctDNA samples from cancer subjects at high coverage. The data looks promising and we are working to see how applicable the method is in a broader range of samples, and whether we can use it for detailed analyses of different tumor entities.

Q: How did you confirm the expression results you inferred from WGS for the cancer samples?
PU: To confirm our expression prediction, we compared our data to the RNA-Seq data from the corresponding primary tumor. We focused on regions with high copy number gains, because these regions are relatively enriched compared to balanced regions, and can reach a very high tumor fraction in the circulation. We extracted the RNA-Seq data from those regions and compared it to what our algorithm predicted for gene expression. Those predictions were confirmed. Currently, we are refining our methods to be more sensitive, so we can make these predictions more generally with the ctDNA that we find in plasma.

> "To confirm our expression prediction, we compared our data to the RNA-Seq data from the corresponding primary tumor."

Q: What software do you use to analyze the data?
PU: We use the Burrow-Wheeler Aligner (BWA) for DNA alignment and have some scripts that use SAMTools to look at the coverage around the transcription start sites. For gene expression prediction, we use a machine learning approach based on support vector machines that we developed.

Q: What challenges did you face in conducting this study?
PU: Sample size and the cost of sequencing were significant challenges. Although we have hundreds of samples and a corresponding data set from low-coverage WGS, these data could not be used due to lack of sequencing depth at the regions we were interested in analyzing. We could not afford to sequence them all at high coverage to compensate for the low signal, so we merged the data to overcome this issue.

Q: What enhancements will be necessary to support the transition of your method into a clinical setting?
PU: We've found that there are issues that can obscure the results that we obtain using our method. For example, there are differences in the nucleosome organization of genes that are highly expressed and those that are not. Genes that are poised for high gene expression have very strict nucleosome organization, and our method is successful in predicting the expression of those genes. Genes whose nucleosomes are not strictly organized are more difficult to assess and our algorithm often predicts that they are unexpressed, whether they are or not. As a result, our algorithm can't predict gene expression for

every gene in a sample currently. We're working on correcting this.

We will also need a higher tumor fraction of liquid biopsy to apply this method in a clinical setting, because the signal is "contaminated" with normal DNA from the usual apoptotic processes. The two cancer samples in our study had high tumor fractions. We might not find that in all subject samples. Even for samples with high tumor fractions, we still must focus on the regions that are overrepresented and that have copy number alterations and gains. Currently, we are limited to a few regions where we can predict whether genes are expressed. We hope to find more.

> "For gene expression prediction, we use a machine learning approach based on support vector machines that we developed."

Q: How could cfDNA gene expression assist our understanding of disease in the future?
PU: Scientists have been looking for genetic variations and mutations in tumors. This has led to tremendous advances in molecular oncology and the identification of treatment responses. However, some information is missing. We believe our analysis of gene expression in the tumor will allow us to follow what's going on in that tumor during disease and treatment. We hope this will give us a whole new spectrum of information that can be used to identify drug targets or monitor treatment, and catch the first signs of therapy resistance when they appear.

EH: Another benefit of our approach is that we can perform copy number analysis and extract expression data from the same data set. In a recent study, we reported the emergence of novel focal amplification of prostate cancer.[4] Tumor genomes are highly dynamic during tumor progression. In the future, our approach could tell us whether a gene that is located in this focal amplification is expressed in the tumor, and whether it is responsible for driving the progression of the disease or resistance to treatment. We believe that we can apply this approach to any type of cancer, given that enough tumor DNA is present in the cfDNA.

> "The MiSeq and NextSeq workflows are easy and the sequencing speed is fast, providing significant time savings."

Q: If new markers are identified, can you use the raw data from these studies to assess the presence of different gene expression levels to understand disease progression?
PU: We can incorporate newly identified markers if they are based on the expression of genes. For the time being, we'll need to focus on markers in regions with copy number gains. That might restrict the search for other kinds of markers. The theory is, if the signal is high enough and there's a significant amount of tumor fraction, we could check whether there has been some kind of response in the gene. We are working on that now.

Q: How have the MiSeq and NextSeq Systems enabled your research?
EH: The MiSeq and NextSeq workflows are easy and the sequencing speed is fast, providing significant time savings. We can perform low-coverage WGS in 2-3 days from plasma DNA extraction to the final results. Both systems provide fast data access, which is a significant advantage for us.

Q: Why did you choose the TruSeq® Nano DNA Library Prep Kit?
EH: We chose the TruSeq Nano DNA Library Prep Kit because it could support low sample input. We obtained great results. For copy number analysis and WGS, we are happy with this kit and intend to keep using it.

Q: What are the next steps for your research?
PU: We are focused on making our method more sensitive, so that it can work with cfDNA that has a lower fraction of tumor DNA. That would enable us to look at samples from earlier in the disease course. The earlier we can obtain information about a tumor, the better chance we have of informing treatment decisions.

EH: We are also trying to analyze other markers, mutations, and copy number alterations out of the plasma DNA that could be used without prior knowledge of the primary tumor and the genes driving its growth. We are trying to extract other information beyond the usual suspect mutations to discover new markers that can help us monitor and screen cancer patients. It is a significant amount of bioinformatic work.

> "We are focused on making our method more sensitive, so that it can work with cfDNA that has a lower fraction of tumor DNA. That would enable us to look at samples from earlier in the disease course."

Q: Do you plan to use Illumina products in your future studies?
PU: We will continue using the MiSeq and NextSeq Systems, and the TruSeq Nano DNA Library Prep Kit.

We are now testing our approach with less sequencing data, by performing target enrichment, as well as reducing our coverage needs by downscaling the experiments where the signal is low.

For Research Use Only. Not for use in diagnostic procedures.

We are already testing some Illumina DNA enrichment kits for this purpose. By optimizing our approach, we hope to make use of all our existing data sets.

EH: We're also using a combination of the TruSeq Nano DNA and Nextera® Rapid Capture protocols for targeted resequencing. We also have an Illumina Concierge project underway where we are running the TruSeq Cancer Amplicon Kit with unique molecular identifiers. If our studies are ever to transfer to the clinic, we have to focus on known, actionable targets. That will be more efficient than analyzing the whole genome to determine the right treatment or why a patient is resistant to a certain therapy.

Learn more about the Illumina systems and products mentioned in this article:

MiSeq System, www.illumina.com/systems/miseq.html

NextSeq System, www.illumina.com/systems/nextseq-sequencer.html

TruSeq Nano DNA Library Prep Kit, www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-nano-dna.html

Nextera Rapid Capture, www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/nextera-rapid-capture-exome.html

TruSeq Amplicon-Cancer Panel, www.illumina.com/products/by-type/clinical-research-products/truseq-amplicon-cancer-panel.html

Illumina Concierge Custom Design Service, https://science-docs.illumina.com/documents/Services/custom-product-services-data-sheet-html-070-2019-002/Content/Legacy/CP-Services/custom-product-services-data-sheet-070-2019-002/custom-product-services-data-sheet-070-2019-002.html

References

1. Snyder MW, Kircher M, Hill AG, et al. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues of origin. *Cell*. 2016; 164: 57-68.
2. Ulz P, Thallinger GG, Auer M, et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nature Genetics*. 2016; 48: 1273-78.
3. Valouev A, Johnson SM, Boyd SD, et al. Determinants of nucleosome organization in primary human cells. *Nature*. 2011; 474: 516-620.
4. Ulz P, Belic J, Graf R, et al. Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer. *Nat Commun*. 2016; 7: 12008, doi: 10.1038/incomms12008.

For Research Use Only. Not for use in diagnostic procedures.