Improving read accuracy for potential ctDNA diagnostics

Chen, A.[1,2*], Hannah S. J.[1,2*], Zou, D. H.[1,2], Guilford, P.[1,2], Black, M. A.[1,2], Day, R. C.[1,2]

[1] Department of Biochemistry, University of Otago, Dunedin, NZ, [2] Centre for Translational Cancer Research, University of Otago, Dunedin, NZ. *Joint first authors.

Background

The rising incidence of cancer is driving the urgent need for innovative solutions to address the absence of effective early detection in the global diagnostic market. **Here we outline experiments looking to to improve the state of the art for molecular testing by overcoming major technical barriers.**

We are hoping to develop a transformative, sensitive and flexible pan-cancer diagnostic test to detect cancer onset and monitor progression. Studies have established the reliability of circulating tumour DNA (ctDNA) for monitoring treatment response and identifying drug resistance mechanisms in patients with advanced cancer Zou *et.al.* 2020.

Most cell-free DNA in the blood is derived from healthy tissues. This presents significant hurdles that need to be overcome for successful deployment of liquid biopsies for **reliable early-stage cancer detection**.

- The rarity of mutated tumour DNA in cell-free samples approaches the inherent error rate of common diagnostic methods;
- The number of mutated tumour DNA molecules are likely too low to reliably prove the presence or absence of disease;
- Mutations alone do not inform the tissue of origin of the cancer cells.

There is also an urgent need to expand diagnostic accessibility and capability to rural communities. Early detection of cancer is a critical factor in cancer survival rates and the increasing global burden is driving an urgent need for effective early-stage detection methods. Imaging-based diagnostics are low-throughput, expensive and generally require travel from rural areas. Circulating DNA is a revolutionary sample type that is proving to be a rich source of biomarkers for pan-cancer screening and is a step forward for accessibility since it is minimally invasive (a simple blood draw by a community nurse) and, if handled correctly, has prolonged stability at room temperature Parackal *et.al.* 2019. Our work looks to put liquid biopsies and ctDNA at the forefront of equitable cancer detection in NZ by establishing our pan-cancer testing on the highly accessible MinION platform.

Selected Milestones for cfDNA Research

1948: Mandel and Metais first discovered the presence of cell-free DNA (cfDNA) in the plasma from both healthy and diseased people. Significance not fully realized and was prior to confirmation of the final structure of DNA.

1968: Tan et. al., found patients with systemic lupus erythematosus had increased cfDNA compared to healthy subjects. Confirmed by Koffler (1973) but they also observed increases in cfDNA with other diseases including rheumatoid arthritis and malignant tumours.

1977: Leon et. al., showed increases cfDNA in patients with a range of cancers but also that levels were higher if metastasis had occurred and importantly the levels were significantly reduced in most patients post treatment. This generally coincided with improved clinical condition.

2010: Lo et al., Fetal genome mapped from sequencing of cfDNA from mothers' blood

2016: Snyder et al., Cell-free DNA comprises an *In vivo* nucleosome footprint that Informs Its tissues-of-origin.

DISCLAIMER: Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.

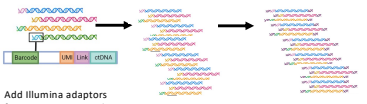
Overview of Universal Molecular Identifiers and Rolling Circle Amplification: Core technologies for improved read accuracy

Here we highlight two approaches to improve base calling accuracy on Nanopore platforms. This reduces erroneous base calls and removes significant amounts of background noise whilst leaving the biological signal intact.

Both the UMI and RCA approaches rely on the generation of multiple copies of individual cfDNA fragments during the library preparation process. The ability to observe multiple copies of the starting molecules in the sequencing data is then also required. Copies are identified as coming from the same original starting molecule and compared to form a consensus sequence, essentially proofreading the data. For us there are two key factors here 1) the starting amount of cfDNA i.e., how many unique molecules are in the starting population, and 2) how many reads we can generate using a single MinION flowcell.

We choose to benchmark the performance of the methods by using a custom 21 gene QiaSeq panel already used in our laboratory for targeted sequencing of gastric cancer samples using Illumina MiSeq sequencing. This panel consisted of approximately 2,800 amplicons tiled across the targeted loci and incorporated Universal Molecular identifiers within the amplicon tails.

A: UMI approach



Add Illumina adaptors from a 21-gene panel with sample barcoding and UMIs.

Amplify using PCR

Add ONT adaptors

Sequence using MinION flowcells on the GridION platform.

B: RCA approach

cfDNA is splinted and circularized

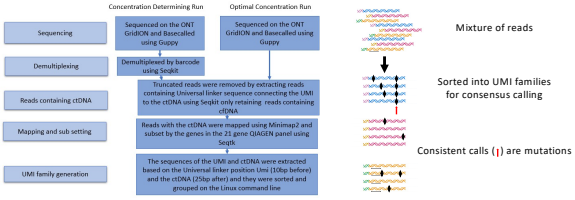
The circular DNA is copied to generate long reads containing repeated copies of the original

Add ONT adaptors

Sequence using MinION flowcells on the GridION platform.

Details of the Bioinformatics workflows used for UMI and RCA

C: Workflow for UMI analysis

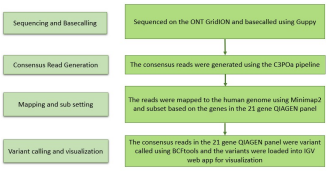


Mixture of reads

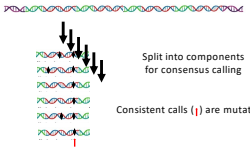
Sorted into UMI families for consensus calling

Consistent calls (|) are mutations

D: Workflow for RCA analysis



Long reads containing repeats



Split into components for consensus calling

Consistent calls (|) are mutations

Universal Molecular Identifiers vrs Rolling Circle Amplification

Here we show selected data from our UMI and RCA sequencing runs on MinION flowcells from the same starting material. Figure E is a table showing the results obtained when we use differing amounts of input material for the UMI approach. It is more efficient to find duplicated reads when lower starting inputs are used. However, for early cancer detection we want to maximise the input of cfDNA to improve our chances of finding the rare tumor-derived fragments in the background population of DNA from non-cancerous cells. Even when a whole MinION run was done with the 10ng input library we were only able to see two or more copies of 25% of the sequenced fragments.

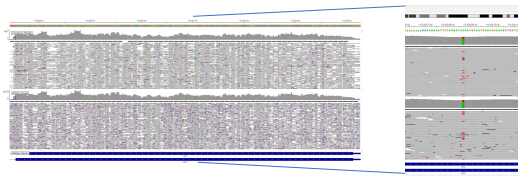
E: Tables of run metrics for UMI experiments

Sample	Total UMI-ctDNA families	UMI-ctDNA families with 1 starting molecule	UMI-ctDNA families with 2 starting molecules	UMI-ctDNA families with 3 or more starting molecules
Run 1 5ng	1,388,605	1,061,367 (76%)	195,200 (14%)	132,038 (10%)
Run 1 10ng	2,378,539	1,871,407 (79%)	354,581 (15%)	152,551 (6%)
Run 1 20ng	2,246,556	1,943,453 (87%)	256,966 (11%)	46,137 (2%)
Run 1 40ng	2,565,227	2,217,838 (87%)	291,679 (11%)	55,710 (2%)

Sample	Total UMI-ctDNA families	UMI-ctDNA families with 1 starting molecule	UMI-ctDNA families with 2 starting molecules	UMI-ctDNA families with 3 or more starting molecules
Run2 10ng	4,073,960	3,060,190 (75%)	503,398 (12%)	529,909 (13%)

F: Example of consensus sequence vrs raw component reads using the RCA method

IGV plots of APC coverage. The coloured line and scale at the top show the genome coordinates and reference sequence, the top depth plot and read alignment shows the coverage in the BAM file from the consensus generated reads from the rolling circle experiment and the second depth plot and read alignment shows the coverage of individual cfDNA fragments prior to consensus calling. The lower raw read alignment, has more sequencing errors. Zoom in shown to the right.



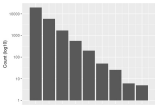
RCA is far more efficient at enabling consensus basecalling since it greatly reduces the sequencing burden required to enable proofreading. When we sequence a cfDNA fragment from an RCA library we automatically pull-out multiple copies facilitating consensus basecalling (See figure F above). The UMI approach requires us to pull out a sister copy at random from a pool of millions/billions of library molecules.

Summary and Ongoing Work

We are combining the portability and relatively inexpensive set up costs of the MinION platform with a simple blood draw to develop an equitable NGS based early cancer diagnostic capable of querying and reporting on multiple aspects of circulating tumor DNA. The approach aims to amplify the disease signal from limited amounts of ctDNA whilst also helping to determine the tissue of origin.

We are currently combining consensus basecalling and increased library length with other projects in the laboratory that attempt to improve basic platform basecalling accuracy (using methods additive to improved pore design and basecalling algorithms) and methods to maintain methylation patterns during DNA amplification.

Technical Note: Several attempts to sequence ctDNA directly on the R9 and R10 flowcells consistently exhibited, very rapid blocking of the pores. This has remained an issue for us. However, we have observed that increasing the length of the library appears to mitigate this problem. For example, direct sequencing of ctDNA only generated Mbs of data from MinION flowcells whereas sequencing of the targeted Illumina library i.e. cfDNA extended by sequencing linkers to approximately 350 bp gave much better performance (5-10 Gb). Use of the much longer RCA DNA generated data yields similar to a standard MinION run (20-30 Gb). To capitalize on this phenomenon, we are now exploring concatenation of the cfDNA fragments into strings. The figure to the right shows the frequency of unique fragments in individual reads in library preparations we sequenced recently. Number of fragments per string are along the bottom axis.



References

Mandel, P., & Metais, P. (1948). Les acides nucléiques du plasma sanguin chez l'homme [Nucleic acids in human blood plasma]. Comptes rendus des séances de la Société de biologie et de ses filiales, 142(3-4), 241-243.

Tan, E. M., Schur, P. H., & Carr, R. I. (1966). Kappa and lambda light chains in human gammaglobulins. The Journal of Clinical Investigation, 45(2), 317-328.

Koffler, D., Agnello, V., Winchester, R., & Kunkel, H. G. (1973). The occurrence of single-stranded DNA in the serum of patients with systemic lupus erythematosus and other diseases. The Journal of Clinical Investigation, 52(1), 198-204.

Leon, S. A., Shapiro, B., Sklaroff, D. M., & Yaros, M. J. (1977). Free DNA in the serum of cancer patients and the effect of therapy. Cancer Research, 37(3), 646-650.

Lo, Y. M., Corbetta, N., Chamberlain, P. F., Rai, V., Sargent, I. L., Redman, C. W., & Wainscoat, J. S. (1997). Presence of fetal DNA in maternal plasma and serum. The Lancet, 350(9076), 485-487.

Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., & Shendure, J. (2016). Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell, 164(1-2), 57-68.

Parackal, S., Zou, D., Day, R., Black, M., & Guilford, P. (2019). Comparison of Roche Cell-Free DNA collection Tubes® to Streck Cell-Free DNA BCT®s for sample stability using healthy volunteers. Practical Laboratory Medicine, 16, e00125. doi: 10.1016/j.plabm.2019.e00125

Zou, D., Day, R., Corallo, J. A., Parackal, S., Mitchell, W., Black, M. A., ... Jackson, C., & Guilford, P. (2020). Circulating tumor DNA is a sensitive marker for routine monitoring of treatment response in advanced colorectal cancer. Carcinogenesis, 41(11), 1507-1517. doi: 10.1093/carcin/bgaa102

Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R. J., Green, R. E., & Vollmers, C. (2018). Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. Proceedings of the National Academy of Sciences, 115(39), 9726-9731. doi: 10.1073/pnas.1805447115