doi:10.1017/S1431927613003036
Microsc. Microanal. 19 (Suppl 2), 2013

First Nucleotide Sequence Data from an Electron Microscopy Based DNA Sequencer

C.S. Own[1], A.L. Bleloch[2], W. Lehrach[2], C. Bowell[2], M. Hamalainen[2], J. Herschleb[2], C. Melville[2], J. Stark[2], M. Andregg[2], and W. Andregg[2]

1. Voxa, Seattle WA USA
2. Halcyon Molecular, Palo Alto CA USA

Whole-genome sequencing, due to its steadily decreasing cost and improved accuracy, is enabling profound advances in medicine and information sciences [1]. While powerful, the current generation of technologies following the first scalable PCR and Sanger methods of the 1990's are limited by relatively short read lengths (<500 bp) which necessitate expensive computation to reconstruct the genome [2]. Further, they generally cannot detect subtle structural changes nor characterize repeat sequences easily, meaning that genomes are typically incomplete and error-prone [2]. New approaches that can circumvent the inherent error and high computational cost are highly desirable, for example Nanopore-based sequencing which is promising in this regard though it has not yet demonstrated long reads or low enough error rate to launch commercially. An alternative novel approach, DNA sequencing by electron microscopy (EM), is capable of similarly long read lengths (>>1 kbp) with potentially reduced error rates, enabling transformative advantages in speed and quality over existing sequencing methods [3-5]. Since it utilizes information-rich direct imaging of whole molecules, a wealth of genetic data previously unavailable or difficult to obtain could be unlocked, such as haplotypes, structural changes in DNA (i.e., methylation of bases), copy number variants (CNVs), and large structural variations (LSVs). We report here on progress in nucleotide sequence data extraction using a novel EM sequencer developed at Halcyon molecular. This novel sequencer incorporates four main technology components:

1. Nucleotide tagging with EM-dense contrast agents
2. DNA linearization and stretching onto thin substrates
3. Automated high-resolution imaging
4. Base position extraction and bioinformatics

Direct imaging of the DNA helix has recently been demonstrated using atomic-resolution phase contrast EM giving promise for single-nucleotide ID [6], however, the high speed needed to acquire full genomes necessitates orders-of-magnitude greater signal and good correlation of that signal with individual nucleotides. We accomplish this by biochemical attachment of monolithic high-Z tags to nucleotides, which simultaneously boosts detectability and mitigates beam damage to the polymer backbone. Multiple independent labelings for each nucleotide type (A,C,T,G) produce monochromatic ssDNA, which after sequencing are recombined bioinformatically. We have experimented with several dozen labeling chemistries—single-atom through multi-atom (cluster) molecules—and have found >5atom molecular clusters to be advantageous due to high detectability, stability, and suitable size, yet reduce resolution requirements by 5x-10x and increase effective sequencing speed by 25x-100x.

After labeling, the DNA is extracted from solution, linearized, and placed onto a thin oxide or nitride substrate using a novel patented polymer nanopositioning technology called molecular threading, which is compatible with ultra-thin (in this case <5 nm) membranes and delivers linearized DNA tens of μm long at least an order of magnitude straighter and more uniform than commonly-used bulk linearization techniques such as molecular combing [7] (see **Fig. 1**). The linearized ssDNA is then imaged at 60 kV in a Nion UltraSTEM100 aberration-corrected STEM that we customized with new control software for automating repetitive tasks and enabling high-speed image acquisition, producing

images of threaded ssDNA strands two orders of magnitude faster and without user intervention. The acquired images are automatically post-processed on a distributed computing platform which integrates segments and identifies peak label positions adjacent to an estimated backbone using machine learning algorithms. The extracted base sequences are then merged using a bioinformatic reconstruction algorithm we developed. **Fig. 2** is an example of a merged 30 bp test sequence using a novel 6-Re-atom molecular cluster label we developed, showing nucleotide positions and demonstrating characteristics of the data and errors that manifest in this technique. Correct densities along the backbone at the mesoscale (10+ nm regime) have been measured and indicate single-strandedness. We theorize that the errors arise mainly from 1) local clumping of labels on the 1-5 nm scale and 2) potential label "flop" due to a finite label- backbone ligand length of a few nm (19 C-C bonds) which reduces positional accuracy between stretched bases (~0.6 nm). Further reduction of the ligand length is expected to significantly increase accuracy of the sequencer, however, increased steric interaction likely comes as a tradeoff.

[1] Kilpinen H & Barret JC. Trends in Genetics, 2013 **29**(1): 23-30.
[2] Metzker, M. Nature Reviews Genetics, 2010. **11**(1): 31-46.
[3] Schadt, E, *et al*. Human Molecular Genetics, 2010. **19**(2): 227-240.
[4] Bleloch, *et al*. Microscopy & Microanalysis, 2011. **17**(S2): 1274-1275.
[5] Bell DC, *et al*. Microscopy & Microanalysis, 2012. **18**(5): 1049-1053.
[6] Gentile F, *et al*. Nano Lett, 2012. **12**(12): 6453-6458.
[7] Payne A, *et al*. (submitted).
[8] We thank the U.S. Department of Energy Grant #DE-FG02-02ER63445 and NIH Grant #RC2 HG005592 for financial support.

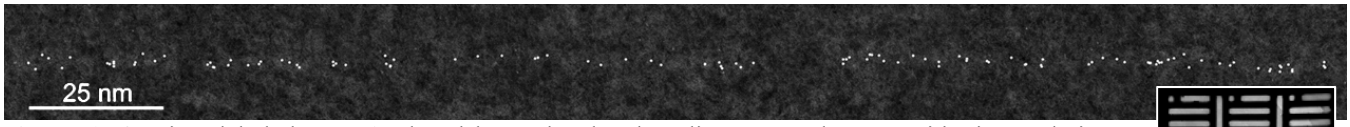

Figure 1. Osmium-labeled ssDNA placed by molecular threading, a novel nanopositioning technique, imaged in the STEM (about 500 bases). This is part of a massive composite image tens of kilobases long comprised of thousands of images. Increased contrast around the molecule indicates ssDNA backbone position, and the fine peaks are atomic positions of labels. Inset: Different types of ssDNA were placed in each 1 um x 8 um longitudinal well of this microfabricated nanowell assay substrate (section shown) and imaged sequentially.

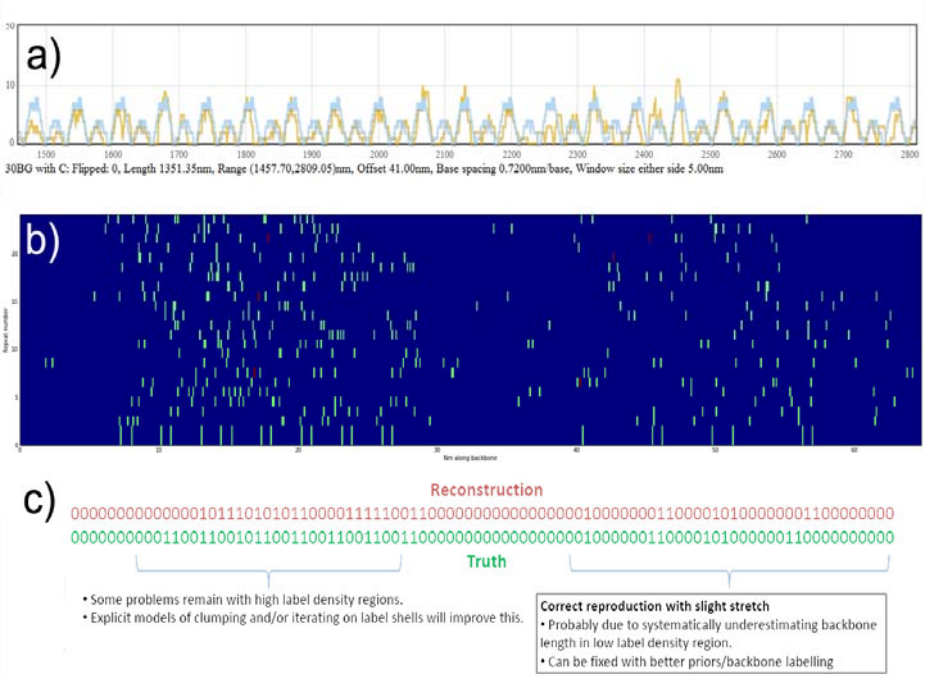

Figure 2. Nucleotide sequence data reconstruction from EM images. a) Measured label density (y-axis) along the backbone (x-axis) for 21 repeats of a 30 bp sequence. Blue = reference, yellow = experimental. b) Label positions for the base sequence discretized and overlaid, showing recovered nucleotide positions (20x coverage). An HMM model is used to infer parameters and reconstruct the sequence shown at bottom. c) Comparison of reconstruction with truth, showing correct density but errors in position; mesoscale ordering is consistent, but the fine scale ordering in the sub-nm regime is not, as would be expected from label-backbone ligands of a few nm.

https://doi.org/10.1017/S1431927613003036 Published online by Cambridge University Press