Tim Farrell, tmf@bu.edu
BF MS Internship Project Report
March 21, 2016

Overview

Next-generation sequencing (NGS) technologies present a new set of opportunities for the practice of medicine. With the unprecedented precision and increasing affordability of NGS, coupled with the growth of electronic health record systems, genomic data seems an inevitable, and potentially powerful, addition to the patient electronic medical record. The large degree of genomic variation across populations, however, present obstacles to clinically-actionable interpretations of this datatype. In particular, genomic variation at structural resolution—variants larger than single-nucleotides but smaller than chromosomes—have shown to be highly heterogenous, difficult to resolve and yet most likely to contribute to phenotypic diversity and disease susceptibility.

Complex genomic regions are difficult to resolve because NGS technologies are currently limited to reads with true lengths of 500 base-pairs (bp) or less. This results in reads mapping to multiple locations (chimeric or split reads) for regions with high repetition and ambiguous haplotypes (which allele sections pair with which). Emerging long-read-capable sequencing, or third-generation sequencing (3GS), technology presents an opportunity to overcome these issues; however, their high susceptibility to error preclude their clinical application. In the following project, we sought to develop a long-read sequencing assay, with an associated computational workflow, that could effectively supplement and improve the performance of NGS in the resolution of clinically-relevant haplotypes for phenotype prediction.

We aimed to develop these methods in application to the human Rh blood group system, a set of polymorphic genes that exhibit considerable structural variation and phenotypic diversity. The Rh system presents a model system to develop clinical-grade structural-variant phenotype prediction techniques as it is sufficiently well-characterized and of realistic scope. Additionally, any developments have direct application to transfusion medicine, screening for hemolytic disease of the newborn and, more generally, genomics-enabled antigenic prediction.

Methods/ Results

Experimentally, we utilized the Illumina MiSeq system and the Oxford Nanopore Technologies (ONT) MinION sequencer, a long-read-capable device, to generate targeted genomic data of Rh genes for 93 patients. A custom primer set was used, along with a long-range PCR amplification protocol. Library preparation was carried out in accordance with Illumina and ONT protocols. We utilized serological data available for those patients as antigenic phenotype data.

Due to the underdeveloped nature of the ONT MinION, we were only able to generate sufficient data to test our hypothesis that long reads can act as an effective supplement to short reads for structurally-variant phenotype prediction. Despite this difficulty, we still believe this use case is fitted even for the current state of 3GS technology.

Computationally, we set out to construct a genomics-based antigenic classifier, using a supervised learning approach, that could predict antigen phenotype from genomic data. The initial challenge was to effectively design feature sets that could most accurately represent the full genomic sequence, including structural features, and ultimately the associated phenotype. We chose the most performant of 16 feature type sets (mean coverage at differential genotype positions; 0.81 +/- 0.01 composite antigen prediction accuracy) after training a decision tree classifier on selected feature data from the available short read data and testing antigen prediction performance over 10 iterations of 10-fold cross-validation. Results from the engineering of these features can be found here.

The additional challenge was to develop methods for combining long and short read data such that the presence of long reads led to better-informed structural features, effectively requiring the implementation of simplified split read and haplotype resolution techniques. Due to the difficulties encountered with the ONT MinION, we were unable to generate enough data to fully implement and test this component of the computational pipeline. A full implementation, however, would have demonstrated a "hybrid" short and long read alignment technique to resolve split reads and simple haplotype estimation as an additional feature in the selected data.

Discussion

Overall

Our goal was to use a computational prediction workflow—a classifier capable of predicting antigens from a set of genomic features—to demonstrate whether long-read sequencing data could be used to increase the predictive ability of short-read sequencing data for structurally-variant phenotypes, in a clinically-relevant context. Although we did not succeed in this demonstration, we still believe 3GS technology will inevitably be needed to improve applications in this domain. This work presents one possible approach.

Work Accomplished/ Student Contribution

This work was completed under the supervision of Peter Tonellato, PhD (Laboratory of Personalized Medicine, Dept. of Biomedical Informatics, HMS) and in collaboration with William Lane, PhD MD (Dept. of Pathology, BWH).

In terms of my contribution, I helped setup the lab equipment for the ONT MinION sequencing experiments, prepared the amplicons/ libraries and performed the ONT MinION sequencing runs. The Illumina data was available from a previous project of Dr. Lane. The primer set was designed by Dr. Lane. The first sequencing experiment resulted in poor performance due to a faulty flow cell and necessitated updating our workflow. Follow-on sequencing experiments will be run with the arrival of new materials.

On the computational side, I designed and implemented all aspects of the pipeline. I wrote the supervising learning workflow in order to engineer the most performant feature type set and to train and test the classifier on the data generated in the lab. I read the relevant literature for designing and developing the hybrid alignment strategy for combining short and long read data.

Impact on Professional Development
In terms of professional development, this work has allowed me to apply my bioinformatics training in a clinically pertinent context, giving me the chance to work directly alongside clinicians and lab personnel to translate bioinformatics tools and technologies into medically-useful applications. These experiences have undoubtedly had a positive impact on my professional development, and will prove invaluable as I look forward to industry work as a bioinformatics engineer.

References

[1] Jameson JL and Longo DL. 2015. Precision medicine – personalized, promising and problematic. *N Engl J Med.* 372(23): 2229-2234.

[2] Biesecker LG and Green RC. 2014. Diagnostic clinical genome and exome sequencing. *N Engl J Med.* 370:2418-25.

[3] Roberts NJ, Vogelstein JT, Parmigiani G, Kinzler KW, Vogelstein B and Velculescu VE. 2012. The predictive capacity of personal genome sequencing. *Sci Trans Med*. 4, 133ra58.

[4] McLauglin HM, Ceyhan-Birsoy O, Christensen KD, Kohane IS, Krier J, Lane WJ, Lautenbach D, Lebo MS, Machini K, MacRae CA, Azzariti DR, Murray MF, Seidman CE, Vassy JL, Green RC, Rehm HL and for the

MedSeq Project. 2014. A systematic approach to the reporting of medically relevant fndings form whole genome sequencing. *BMC Medical Genetics* 15:134. doi:10.1186/s12881-014-0134-1.

[5] Baker M. 2012. Structural variation: the genome's hidden architecture. *Nat Methods*. 9(2): 133-139.

[6] Ornella L, Perez P, Tapia E, Gonzalez-Camacho JM, Burgueno J, Zhang X, Singh S, Vicente FS, Bonnett D, Dreisigacker S, Singh R, Long N and Crossa J. 2014. Genomic-enabled prediction with classifcation algorithms. *Heredity* **112**, 616–626.

[7] Silvestri GA, Vachani A, Whitney D, Elashoff M, Smith KP, Ferguson JS, Parsons E, Mitra N, Brody J, Lenburg ME, and Spira A. 2015. A bronchial genomic classifer for the diagnostic evaluation of lung cancer. *N Engl J Med* 373;3.

[8] The Cancer Genome Atlas. 2015. Genomic classifcation of cutaneous melanoma. *Cell* 161, 1681–1696.

[9] The Clinical Lung Cancer Genome Project and Network Genomic Medicine. 2013. A genomics-based classifcation of human lung tumors. *Sci Transl. Med* **5**, 209ra153.

[10] Anstee DJ. 2009. Red cell genotyping and the future of pretransfusion testing. *Blood.* 114(2):248- 256. doi:10.1182/blood-2008-11-146860.

[11] McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, et al. 2014. Illumina TruSeq Synthetic Long-Reads Empower *De Novo* Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. *PLoS ONE* 9(9): e106689. doi:10.1371/journal.pone.0106689.

[12] Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC and McCombie WR. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res* 25: 1750-1756.

[13] Abel HJ , Duncavage EJ. 2014. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genetics* 206 (2014) 432e440.

[14] Qiu P, Cai X, Ding W, Zhang Q, Norris ED and Greene JR. 2009. HCV genotyping using statistical classifcation approach. *J of Biomed Sci*, 16:62. doi:10.1186/1423-0127-16-62.

[15] Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W and Salit M. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 30(2): 246- 251.