Class Projects for CSE280A:Algorithms for Genetics (**Draft**)

Vineet Bafna

January 23, 2019

Logistics

You can work on these projects in groups of 1 or 2, but no more. All participants will have an oral presentation towards the end of the class. As some of the projects are open-ended, please schedule meetings with the instructor throughout the quarter to discuss progress and converge on specific objectives. A small 5 page report will also be due by the end of the exam week. CSE MS students working on these projects, can choose these projects for their qualifying MS exam option, but must submit a short paragraph describing the specific contribution of each person. Also, they should schedule a meeting with the instructor by week 8 to discuss progress and finalize the objectives.

1 Clarifying amplicons with long (PacBio) reads

AmpliconArchitect (AA) is a tool used to reconstruct the architecture of complex amplicons[2]. Given whole genome short-read data, and a seed region (typically a genomic region with copy number amplification), AA reconstructs an amplicon graph encoding rearranged structures. Nodes in the amplicon graph correspond to genomic segments, and edges correspond to pairs of non-contiguous genomic segments connected by discordant short-reads. The structures often correspond to cyclic extrachromosomal DNA[10], but may not be fully disambiguated because of short-reads. The goal of this project is to disambiguate these structures using longer (PacBio) reads which can often be in excess of 10-15kbp.

Input: An AA reconstructed graph, and PacBio WGS reads.

Output: All paths and cycles in the graph that are supported by long-reads.

Project Steps:

1. Learn about the AA graph structure, and isolate a collection of genomic fragments $\mathcal{F}$ that participate in an amplicon.
2. Use existing software, or write your own to map fragments to PacBio reads.
3. Use graph traversal algorithms to identify paths and cycles that are strongly supported.
4. Discuss extensions with the instructor.

2 Clarifying amplicons with linked-(10X) reads

The project is identical in scope to Project 1. However, the readers must read about 10X linked-read technology[3,4] and discuss with the instructor about changing the approach to handle the data.

3 Aligning Optical Mapping Reads to sequence data

Optical mapping technology doesn't produce sequence. Instead, it outputs all positions of specific k-mer (corresponding to a DNAse restriction site). Therefore, an OM read generated from a genomic fragment f is a sorted collection of numbers, each referring to the position of the k-mer in the fragment. While genomic reads are small, and rarely more than 15kbp, OM reads could be 50-100kbp. A local company, BioNano technologies, is generating such OM reads.

The reference human genome is constantly improving. New versions of the assembly can be downloaded from public databases such as Genbank and others. A 'mapping' of an OM read to the genome assembly refers to an identification of a genomic substring that has a highly similar ordering of restriction sites (k-mers). Because of their lenths, OM reads have been used to identify SVs[7,8,11]. The goal of this project is to do some initial analysis towards identification of structural variants.

Input: A collection of OM reads, and a reference assembly version.

Output: A listing of all split-mappings. By split-mappings, we mean OM reads mapping to more than one contiguous regions of the reference assembly, suggesting a structural variation.

Project Steps:

1. Learn about reference assemblies, and download GRCh38.
2. Write a script that simulates optical reads from the reference sequence, along with errors.
3. Devise a dynamic programming algorithm to align the OM read to the genome. Devise methods to speed up the search (Discuss with the instructor).
4. Design and implement algorithms for detecting Structural Variation as described in lectures.

4 Identifying Viral Integration sites in tumor cell genomes

Viruses are implicated in many human cancers. For example, HPV16 is important for cervical and head-and-neck cancers. While many individuals carry the virus, a much smaller subset get cancer. It is suspected that the integration of the virus into the genome plays an important role in this transition[5]. The goal of this project is to develop a tool for fast detection of viral integration in human sequence.

Input: Given a collection of genomic reads from a donor (in bam format, so that they have been already mapped to a human reference), and sequences from a *target* oncovirus, such as hpv16.

Output: A subset of *hybrid* reads , that are part human, part virus. Such hybrid reads are likely to be part of the unmapped reads.

Project Steps:

1. Note that there are software tools, such as ViFi[9] that can identify hybrid viral-human reads. However, these are slow, as the viral sequence is hyper-variable, and ViFi uses HMMs for sensitive detection. Download ViFi and learn how to use it.
2. Use a tool (recommended: Art[6]) to simulate hybrid reads.
3. Develop a fast filtering strategy to quickly eliminate a reads that is unlikely to be hybrid, while retaining all of the hybrid ones. Talk to the instructor for unpublished ideas.
4. Implement your filtering strategy to get a faster method for hybrid read detection capable of detecting human viral integration.
5. Make a plan to validate your results on simulations using cross-validation etc..
6. Produce plots describing performance depending on variability of viral sequence. Talk to the instructor for research extensions.

5 Clarifying the amino-acid sequence of coding VNTRs

Variable Number Tandem Repeats are repeating units (RUs) of length 6-100. Approximately 2,000 VNTRs are found in coding sequence, where variation in the number of RUs (RU-count) can change the protein sequence, including shortening of the sequence due to premature stop-codons. Recent methods[1] can genotype VNTRs from short-read whole genome sequencing. The goal of this project is to take the a collection of reads for from a coding VNTR, where the RU-count is known, and reconstruct the most likely amino-acid sequence. You should focus on VNTRs that can be spanned completely by a short-read.

Input: A collection of Illumina reads from a coding VNTR, the RU-count supported by the reads and a reference protein where the VNTR occurs.

Output: A multiple alignment of the reads, with consensus DNA sequence, and a predicted protein sequence.

1. Use published tools, or your own code to align the reads, accounting for any sequencing errors. You can use tools for computing multiple-alignments, or *de novo* assembly, or others.
2. Generate a consensus DNA sequence.
3. Use a conceptual translation of the consensus and the reference protein sequence to generate the predicted amino-acid changes due to the VNTR.
4. Contact the instructor for data and to discuss research extensions of the problem.

References

[1] M. Bakhtiari, S. Shleizer-Burko, M. Gymrek, V. Bansal, and V. Bafna. Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.*, 28(11):1709–1719, Nov 2018. Oral presentation at RECOMB2018.

[2] Deshpande, V. and Luebeck, J. and Nguyen, N-P.D. and Bakhtiari, M. and Turner, K.M. and Schwab, R. and Carter, H. and Mischel, P.M. and Bafna, V. Reconstructing and characterizing focal amplifications in cancer using Amplicon Architect. *Nature Communications (In Press); early version on bioRxiv*, 2018.

[3] Rebecca Elyanow, Hsin-Ta Wu, and Benjamin J Raphael. Identifying structural variants using linked-read sequencing data. *Bioinformatics*, 34(2):353–360, jan 2018.

[4] Stephanie U. Greer, Lincoln D. Nadauld, Billy T. Lau, Jiamin Chen, Christina Wood-Bouwens, James M. Ford, Calvin J. Kuo, and Hanlee P. Ji. Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome Medicine*, 9(1):57, dec 2017.

[5] Ian J. Groves and Nicholas Coleman. Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us? *Journal of Pathology*, 245(1):9–18, 2018.

[6] Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. ART: A next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 2012.

[7] Le Li, Alden King-Yung Leung, Tsz-Piu Kwok, Yvonne Y. Y. Lai, Iris K. Pang, Grace Tin-Yun Chung, Angel C. Y. Mak, Annie Poon, Catherine Chu, Menglu Li, Jacob J. K. Wu, Ernest T. Lam, Han Cao, Chin Lin, Justin Sibert, Siu-Ming Yiu, Ming Xiao, Kwok-Wai Lo, Pui-Yan Kwok, Ting-Fung Chan, and Kevin Y. Yip. OMSV enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. *Genome Biology*, 18(1):230, dec 2017.

[8] Lee M. Mendelowitz, David C. Schwartz, and Mihai Pop. Maligner: a fast ordered restriction map aligner. *Bioinformatics*, 32(7):1016–1022, apr 2016.

[9] N. D. Nguyen, V. Deshpande, J. Luebeck, P. S. Mischel, and V. Bafna. ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. *Nucleic Acids Res.*, 46(7):3309–3325, Apr 2018.

[10] K. M. Turner, V. Deshpande, D. Beyter, T. Koga, J. Rusert, C. Lee, B. Li, K. Arden, B. Ren, D. A. Nathanson, H. I. Kornblum, M. D. Taylor, S. Kaushal, W. K. Cavenee, R. Wechsler-Reya, F. B. Furnari, S. R. Vandenberg, P. N. Rao, G. M. Wahl, V. Bafna, and P. S. Mischel. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature*, 543(7643):122–125, 03 2017.

[11] Anton Valouev, Lei Li, Yu-Chi Liu, David C Schwartz, Yi Yang, Yu Zhang, and Michael S Waterman. Alignment of Optical Maps. *JOURNAL OF COMPUTATIONAL BIOLOGY*, 13(2):442–462, 2006.