BIOF 520: Culture-independent pathogen detection for public health and patient care

Diana Lin[1,2]

[1]Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada

[2]Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada

CLINICIAN WORKFLOW

The clinician will collect the biological specimen from the patient, depending on what symptoms are being exhibited (1). For example, if the patient is exhibiting gastrointestinal symptoms, a gut or fecal sample may be taken, whereas with respiratory symptoms, sputum or buccal sample may be taken. Along with the sample, metadata about the patient (e.g. age, sex, location, etc.) is also recorded for public health purposes.

LABORATORY WORKFLOW

With each different type of biological specimen, the appropriate corresponding sample preparation and processing protocol will be used. Following standard library construction protocols, a syndromic PCR panel (2) will be conducted to test for common pathogens such as *E. coli,* salmonella, etc. If the test results are positive, then the pathogen has been identified, and the workflow can proceed to public health surveillance. If the results are negative, metagenomic sequencing will be completed. To enrich prokaryotic DNA, the protocol Depletion of Abundant Sequences by Hybridization (DASH) (3) will be used to deplete the eukaryotic human DNA from the microbiome and pathogen DNA. In this protocol, human DNA is targeted for cleavage through the use of the recombinant Cas9 protein with a library of guide RNAs, where the cleaved DNA sequences no longer have the required adapters, consequently not amplified and sequenced. DASH is an ideal method as it can be adapted for any sample type while increasing yield without additional cost (4).

BIOINFORMATICS CLINICAL CARE WORKFLOW

For the clinical care workflow (rapid turn-around time of 1 day), long read sequencing with the MinION Mk1B sequencer, chosen for its price ($1000), speed, ease of use, and portability, will be used. To ensure assembly of only pathogens and not microbiome, the long reads will be filtered using BioBloomTools (BBT) (5), a state-of-the-art tool used to filter out contaminants from sequencing data. The bloom filters (BF), a fast, accurate and memory-efficient probabilistic data structure, will be constructed from sequences obtained by the NIH Human Microbiome Project (6). The reads that match those in the BF will be *filtered out*. If time-sensitive, a multi-index BF can be used instead (7). In this case, the majority of typical microbiome sequences will be filtered out, but some may remain so a metagenomic-capable assembler should be employed. The remaining long reads will be assembled using Flye (8) if time-sensitive, otherwise using Canu (9), based on the results of a comparative study of long read metagenomic assemblers (10).

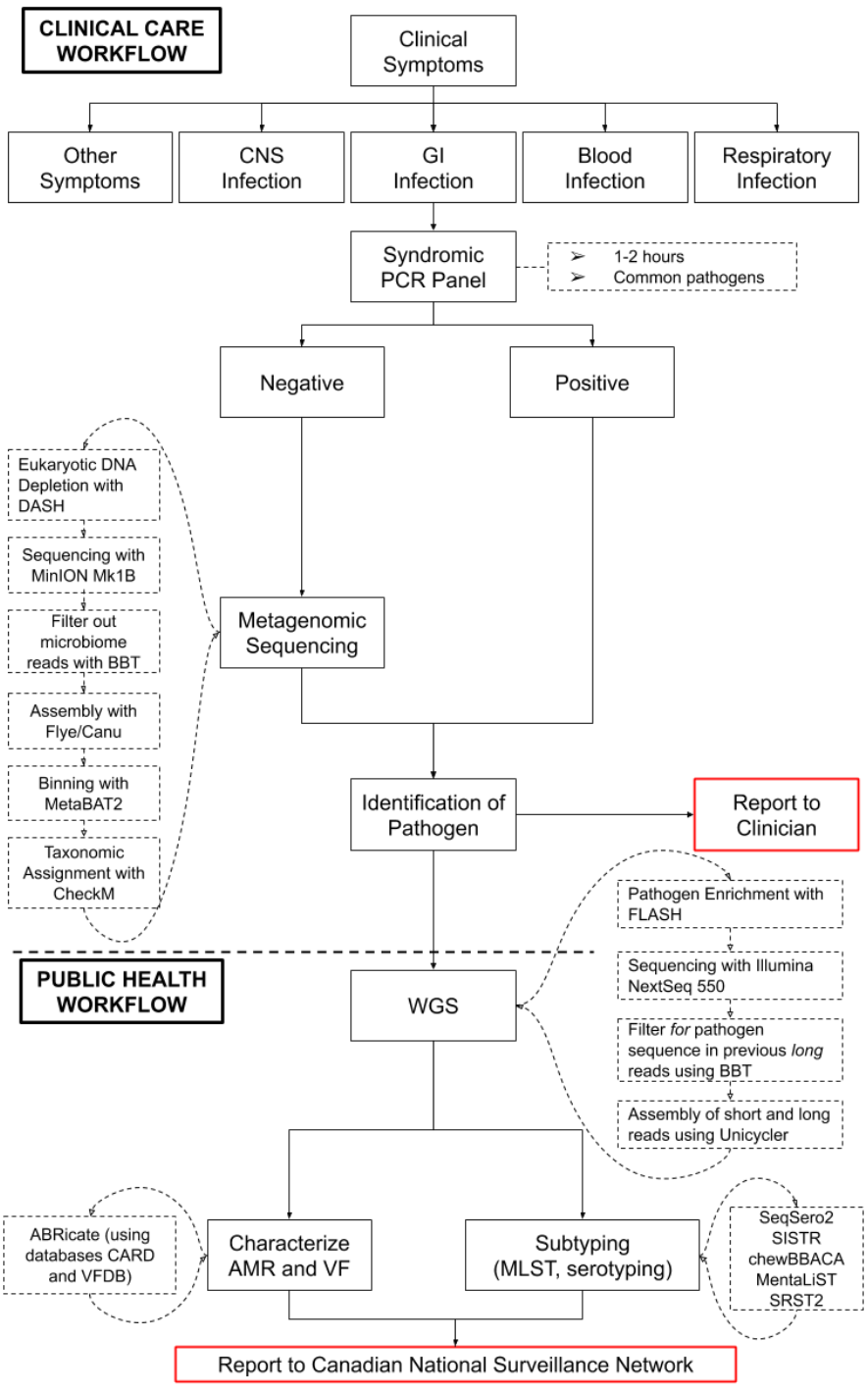
First, the reads will be aligned to the assemblies using BWA (11) as abundance input for binning. Using MetaBAT2 (12), chosen for its adaptive binning algorithm thereby bypassing any parameter fine-tuning, the resulting contigs will be binned based on tetranucleotide frequencies and abundance scores. Then, CheckM (13), chosen for its binning and quality control capabilities, will then be used to bin based on taxonomy and operational taxonomic unit (OTU), while assessing the contamination and completeness of each assembled metagenome. Abundance can be quantified with RPKM by aligning the reads to each bin using BWA. Optionally, for better results, only assemblies with high completeness and low contamination can be used. At this point, since human DNA and microbiome DNA had been filtered out before assembly, the pathogen will be identified. This sequence can be queried against our pathogen database to identify closely related sequences. This ends the workflow for clinical care, and the identified pathogen can be reported to the clinician to facilitate patient treatment. This data can be uploaded to the Canadian Genomics Cloud (which in turn, is integrated with electronic health records).

BIOINFORMATICS PUBLIC HEALTH WORKFLOW

For the public health workflow (rapid turn-around of 3 days), whole genome sequencing (WGS) will be performed. Since the pathogen (and its sequence) has been identified at this point (for designing guide RNAs), the enrichment protocol Finding Low Abundance Sequences by Hybridization (FLASH) (14) can be used to enrich pathogen DNA after DNA extraction. Since FLASH only works for genomic DNA and cDNA, a reverse transcription step must be added beforehand if the pathogen is an RNA virus. FLASH works very similarly to DASH but before library prep,

where the Cas9 targets are cleaved and available for adapter ligation and sequencing. FLASH is an optimal method for pathogen DNA enrichment as it requires low hands-on time and has a low cost per sample (4). In this case, sequencing will be carried out on the Illumina NextSeq 550, a $400,000 flexible benchtop sequencer with a run-time of 12-30 hours, resulting in paired-end 150 base-pair short reads. Next, the raw long reads from the metagenomic sequencing will be filtered with BBT, using a BF with the identified pathogen sequence. Contrary to before, the reads that match those in the BF will be *filtered for* and kept for assembly. Then, a hybrid assembly, combining the high accuracy of short reads with the long range information of long reads, will be executed. The short and long reads will be assembled by Unicycler (15), a hybrid assembler designed for small (bacterial, viral, and organellar) genomes. Finally, the assembled pathogen sequence will be deposited to our existing pathogen database, expanding it further.

The resulting assembly will be screened for antimicrobial resistance genes and virulence factors using ABRicate (https://github.com/tseemann/abricate) and the following databases: the Comprehensive Antimicrobial Resistance Database (CARD) (16) and Virulence Factors Database (VFDB) (17). Next, *in silico* subtyping will be conducted to identify the isolate. If the identified pathogen is salmonella, serotyping can be conducted with SeqSero2 (18) and SISTR (19) to identify surface antigens. Multi-locus Sequence Typing (MLST) for strain identification can be achieved with chewBBACA (20) using the genome assembly, and with MentaLiST (21) and SRST2 (22) using short reads. The combined subtyping, serotyping, MLST profiles can be compiled into a report for the Canadian national surveillance networks. Using the clinical individual data from the Canadian Genomics Cloud and these MLST profiles, health authorities can conduct transmission tracking and outbreak detection using generated transmission maps. This entire workflow is summarized in **Figure 1**.



2019-nCoV

Overall, this workflow should work relatively well for the 2019 novel coronavirus as it is an RNA virus. This is accounted for in the workflow by adding a reverse transcription step before library prep for WGS since most of the protocols selected only work on genomic DNA and cDNA sequences.

Figure 1. The complete workflow for culture-independent pathogen detection for public health and patient care. *Adapted from Jimmy Liu.*

REFERENCES

1. Bachmann NL, Rockett RJ, Timms VJ, Sintchenko V. 2018. Advances in Clinical Sample Preparation for Identification and Characterization of Bacterial Pathogens Using Metagenomics. Front Public Health 6:363.
2. Ramanan P, Bryson AL, Binnicker MJ, Pritt BS, Patel R. 2018. Syndromic Panel-Based Testing in Clinical Microbiology. Clin Microbiol Rev 31.
3. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, DeRisi JL. 2016. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol 17:41.
4. Carleton HA, Besser J, Williams-Newkirk AJ, Huang A, Trees E, Gerner-Smidt P. 2019. Metagenomic Approaches for Public Health Surveillance of Foodborne Infections: Opportunities and Challenges. Foodborne Pathog Dis 16:474–479.
5. Chu J, Sadeghi S, Raymond A, Jackman SD, Nip KM, Mar R, Mohamadi H, Butterfield YS, Robertson AG, Birol I. 2014. BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. Bioinformatics 30:3402–3404.
6. Proctor LM, Chhibba S, McEwen J, Peterson J, Wellington C, Baker C, Giovanni M, McInnes P, Dwayne Lunsford R. 2013. The NIH Human Microbiome Project. The Human Microbiota.
7. Chu J, Mohamadi H, Erhan E, Tse J, Chiu R, Yeo S, Birol I. Improving on hash-based probabilistic sequence classification using multiple spaced seeds and multi-index Bloom filters.
8. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546.
9. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. Genome Res 27:722–736.
10. Latorre-Pérez A, Villalba-Bermell P, Pascual J, Porcar M, Vilanova C. Assembly methods for nanopore-based metagenomic sequencing: a comparative study.
11. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.
12. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7:e7359.
13. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.
14. Quan J, Langelier C, Kuchta A, Batson J, Teyssier N, Lyden A, Caldera S, McGeever A, Dimitrov B, King R, Wilheim J, Murphy M, Ares LP, Travisano KA, Sit R, Amato R, Mumbengegwi DR, Smith JL, Bennett A, Gosling R, Mourani PM, Calfee CS, Neff NF, Chow ED, Kim PS, Greenhouse B, DeRisi JL, Crawford ED. 2019. FLASH: a next-generation CRISPR diagnostic for multiplexed detection of antimicrobial resistance sequences. Nucleic Acids Res 47:e83.
15. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595.
16. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN, Doshi S, Courtot M, Lo R, Williams LE, Frye JG, Elsayegh T, Sardar D, Westman EL, Pawlowski AC, Johnson TA, Brinkman FSL, Wright GD, McArthur AG. 2017. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res 45:D566–D573.
17. Chen L, Zheng D, Liu B, Yang J, Jin Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. Nucleic Acids Res 44:D694–7.
18. Zhang S, den Bakker HC, Li S, Chen J, Dinsmore BA, Lane C, Lauer AC, Fields PI, Deng X. 2019. SeqSero2: Rapid and Improved Serotype Determination Using Whole-Genome Sequencing Data. Appl Environ Microbiol 85.
19. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, Taboada EN. 2016. The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. PLoS One 11:e0147101.
20. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA. 2018. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microb Genom 4.
21. Feijao P, Yao H-T, Fornika D, Gardy J, Hsiao W, Chauve C, Chindelevitch L. 2018. MentaLiST - A fast MLST caller for large MLST schemes. Microb Genom 4.
22. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med 6:90.