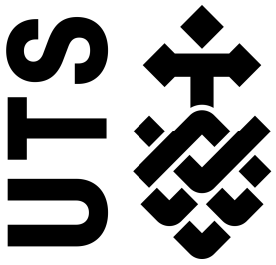
The integration of fine-scale DNA-DNA associations by inclusion of Hi-C DNA cross-linking information into metagenomic community analysis

by

Matthew Zachariah DeMaere

A Thesis Submitted In
Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy



University of Technology Sydney
Sydney, NSW, Australia

February, 2019

Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organisation of the entire tapestry.

Richard Feynman - The Character of Physical Law: Messenger Lectures

Certificate of Original Authorship

I, Matthew Zachariah DeMaere, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Science at the University of Technology Sydney. This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed
prior to publication.

Date: 2019-03-11

Matthew Z. DeMaere

Acknowledgments

I would like to thank my principal supervisor Associate Professor Aaron Darling for his time, advice and support during my candidature. In particular, for supporting the original initiative, a patient ear for long digressions and tolerance for random-walk problem solving. In addition, I would like to thank my co-supervisor Professor Steven Djordjevic for his time, advice and crucial support. I would like to thank the ithree institute for fostering an environment conducive to quality research and their continual pursuit of a positive postgraduate student experience. I would also like to thank Professor Ian Charles for his valuable assistance in initiating and arranging the candidature.

I would like to thank my mother Jennifer, father James and sister Kathleen for their enthusiasm and continued support in what has been a *long* journey.

Lastly, I would especially like to thank my wonderful wife, Klára, for her ceaseless support despite the increasing complication of our lives and her ever expanding contribution to this endeavor.

Matthew Z. DeMaere
Sydney, Australia, 2019.

Thesis Format

A thesis by publication.

List of Publications

Peer reviewed

1. **DeMaere, M. Z.**, & Darling, A. E. (2016). Deconvoluting simulated metagenomes: the performance of hard- and soft- clustering algorithms applied to metagenomic chromosome conformation capture (3C). *PeerJ*, 4, e2676. https://doi.org/10.7717/peerj.2676

2. **DeMaere, M. Z.**, & Darling, A. E. (2018). sim3C: simulation of Hi-C and Meta3C proximity ligation sequencing technologies. *GigaScience*, 7(2). https://doi.org/10.1093/gigascience/gix103

3. **DeMaere, M. Z.**, Darling, A. E. (2019). bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biology*, 20(1), 46. https://doi.org/10.1186/s13059-019-1643-1

Contents

List of Figures

List of Tables

Abstract

Much of our understanding of the microbial world has been obtained using culture-based methodologies, a paradigm that has stood since the 19th century. And yet it has long been known that most of the Earth's microbial species are resistant to laboratory culture. It is reasonable to expect; therefore, that applying equal scrutiny to all microbial life will lead to significant discoveries. Motivated by this, metagenomics eliminates the culturing dependency by directly sampling DNA from an environment; successfully shedding light on the once unseen majority.

The technical limitations of present-day sequencing technologies have meant, however, that in achieving culture-independence, traditional shotgun metagenomic sequencing experiments make a considerable sacrifice. That sacrifice comes in the form of information loss where, in preparing DNA for sequencing, much of the "same-cell" and "same-chromosome" information is destroyed; information which is essential when reconstructing the individual genomes. Purely computational solutions to overcoming this sacrifice have proved insufficient; surpassed instead by strategies which employ changes in the experimental design aimed at reducing the information loss.

A recent strategy is the inclusion of a new form of sequencing data, provided by the Hi-C sequencing technique. Originally conceived to study the three-dimensional structure of chromatin, the Hi-C sequencing technique captures *in vivo* proximity interactions between DNA loci in an all-vs-all manner. When applied to direct metagenomic sampling, the physical structure of the microbial community (chromosome, cell and community) strongly influences the probability of observing proximity interactions between loci, and this pronounced modulation can be exploited to recover the information lost during shotgun sequencing.

This thesis details the effective integration of Hi-C into metagenomic sequencing studies to accurately reconstruct individual genomes, thereby deconvoluting the metagenome. To accomplish this, first an *in silico* investigation of the effectiveness of graph clustering as a

means of metagenome deconvolution was conducted; where Hi-C proximity interactions defined the edges and assembly contigs defined the nodes. A parametric sweep of experimental and community composition parameters was carried out, exploring how the degree of evolutionary divergence (from species to strains) affected the quality of deconvolution. For each iterate in the sweep, a ground-truth was constructed and quality assessed using a novel external validation measure supporting overlapping clusters and variable object weights.

This work led to the design and implementation of the first metagenomic Hi-C read-pair simulator, sim3C, capable of simulating complex community definitions and simple three-dimensional structural elements. While in pursuit of the final objective of metagenome deconvolution, sim3C enabled an externally validated development process.

Lastly, as the outcome of the final objective, bin3C is demonstrated; an open-source solution to Hi-C driven metagenome deconvolution. In an unsupervised manner, bin3C reconstructs individual genomes from metagenomic data. Using external validation of simulated data, bin3C is shown to have high precision and good recall. When a real human microbiome was analysed, bin3C achieved leading performance, resolving 20 more nearly-complete MAGs (57% gain) than its closest competitor.

Dissertation directed by Associate Professor Aaron E. Darling
ithree institute, Faculty of Science
University of Technology Sydney