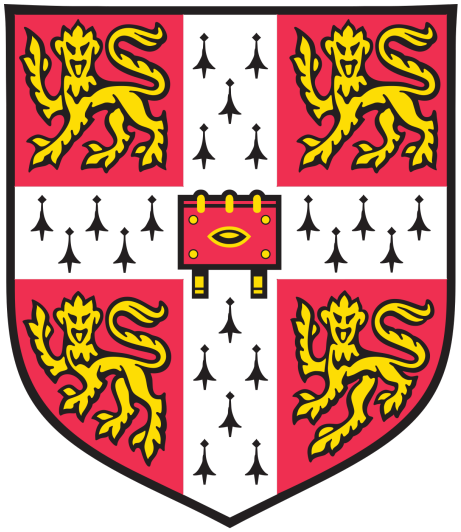
Integrative Analysis of the Human Gut Phageome Using a Metagenomics Approach



Luis Fernando Camarillo Guerrero

Gonville & Caius College
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

August 2020

Dedicated to my parents (Rosa María and Leopoldo) and sister (Marisol) for loving me and always supporting me in an unconditional manner from Day 1 of my life

Dedicada a mis padres (Leopoldo y Rosa María) y hermana (Marisol) por quererme y apoyarme de una manera incondicional desde mi primer día de vida.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee

Luis Fernando Camarillo Guerrero
August 2020

Summary

Integrative Analysis of the Human Gut Phageome Using a Metagenomics Approach

Luis Fernando Camarillo Guerrero

Bacteriophages (or phages; viruses that infect bacteria and archaea) profoundly influence microbial communities. Given the impact of the gut microbiome composition and function on human health, there is a growing focus on phages that inhabit the gut ecosystem. However, the extent of viral diversity, biology, and worldwide epidemiology of gut phages remain largely unknown. In this thesis, I carry out a comprehensive genomic analysis of gut phages by harnessing the biggest collection of phage genomes, gut bacteria isolates, and human gut metagenomes.

I begin by introducing the Gut Phage Database (GPD) which is the largest genomic resource to date of human gut phage genomes and product of mining 28,060 faecal metagenomes and 2898 gut bacteria isolate genomes. I use machine learning to improve the quality of the predictions and investigate ways to organise the viral diversity in order to improve the characterisation of gut phages in downstream analyses.

Afterwards, I describe common functions and auxiliary metabolic genes encoded by human gut phages. I also highlight instances of hypervariable domains which may indicate the presence of phage receptor binding proteins. I then shift the focus to the analysis of two clades of gut phages, namely the Gubaphage and the *Picovirinae* subfamily. The Gubaphage is a novel phage clade uncovered in this work which is highly prevalent across the world. The *Picovirinae* clade was the most common predicted phage taxonomy in GPD. Host assignment allows me to study patterns of phage diversity across bacterial clades of the human gut and investigate their host range.

Finally, I analyse global patterns of the human gut phageome and its association with lifestyle and bacterial composition. I assess the idea of a core virome as well as in what degree my data agrees with this concept.

Acknowledgements

I would like to thank my supervisor Trevor Lawley for trusting me and allowing me to embark on a PhD project of my own. I'm aware of how fortunate I am to have had this freedom. Trevor, I am very grateful for your continuous guidance, and support throughout my PhD journey.

I would like to thank my colleagues from my lab team162 who I have been fortunate to learn a lot from, in particular Yan Shao and Hilary Browne. Yan not only became a great friend, he challenged me throughout my PhD with stimulating discussions about my project. Thanks Yan, I appreciate all your support and you intellectually pushed me to reach new heights in my knowledge about phages. Hilary, I appreciate your accessibility and the time you invested proofreading presentations/reports/thesis/manuscript, you give amazing advice.

I also would like to thank my manuscript collaborator Alexandre Almeida. Alex, your positive outlook was always refreshing. I learned a lot working along with someone of your scientific calibre. Thanks for all your advice and friendship.

I extend my gratitude to the Wellcome Sanger Institute for funding my PhD, Carl Anderson, Christina Hedberg-Delouka, Annabel Smith, and the Pathogen Informatics team.

I would like to acknowledge Darwin College which was the foundation of my social life in Cambridge. To all the people I met there, I appreciate your friendship and have fond memories of you. To the Darwinian Michael Schneider, who actually I met at Imperial College London, thanks for your friendship and for all the nights out at the Cambridge Bops. Honourable mention to the Darwin Salsa society, every week I was looking forward to go dancing on Monday at 8:00 pm. I would also like to thank the Mexicans living in Cambridge, I had so much fun hanging out with you all and really made me feel at home.

I thank Jose Manuel Aguilar Yañez for playing an integral role in my formation as a Scientist and always believing in me.

Last but not least, I thank my family to whom I dedicate this thesis. Your unconditional love and support were vital for me throughout this endeavour – *¡Lo logramos!*

Publications

Camarillo-Guerrero L.F., Almeida A., Rangel-Pineros G., Finn R.D., Lawley T. (2020). Massive expansion of human gut bacteriophage diversity. (In press). *Cell*.

Fung C., Tan S., Nakajima M., Skoog E.C., **Camarillo-Guerrero L.F.**, Klein J.A., Lawley T.D., Solnick J.V., Fukami T., Amieva M.R. "High-resolution mapping reveals that microniches in the gastric glands control Helicobacter pylori colonization of the stomach." PLoS biology 17.5 (2019): e3000231.

Contributions

This thesis is the result of my own work except:

- Metagenome assembly, sequence viral prediction with VirFinder and VirSorter, and dereplication at 95% sequence identity was carried out by Alexandre Almeida.
- Read mapping of GPD predictions to 28,060 metagenomes was carried out by Alexandre Almeida.
- Bacterial taxonomic assignment of gut isolates with the GTDB toolkit was carried out by Alexandre Almeida.
- The tool to assign a taxonomic rank to GPD predictions was developed by Guillermo Rangel Pineros.

Abbreviations

Acr	Anti-CRISPR
Abi	Abortive infection
AMG	Auxiliary Metabolic Gene
ANI	Average Nucleotide Identity
ARG	Antibiotic Resistance Gene
BACON	Bacteroidetes-Associated Carbohydrate-binding Often N-terminal
BAM	Bacteriophage Adherence to Mucus
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTHR	Collagen Triple Helix Repeat
DNA	Deoxyribonucleic acid
FP	False Positive
GPD	Gut Phage Database
GTDB	Genome Taxonomy Database Toolkit
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
IBD	Inflammatory Bowel Disease
ICTV	International Committee on Taxonomy of Viruses
ICE	Integrative and Conjugative Element
Ig	Immunoglobulin
ImmeDB	Intestinal Microbiome Mobile Elements Database
KEGG	Kyoto Encyclopaedia of Genes and Genomes
MCL	Markov Cluster
MGE	Mobile Genetic Element
ML	Machine Learning
NCBI	National Centre for Biotechnology information
OMV	Outer Membrane Vesicle
PCA	Principal Component Analysis
PC	Protein Cluster
PD	Parkinson's Disease
PICI	Phage-Inducible Chromosomal Islands
PtW	Piggyback-the-Winner

QC	Quality Control
RBP	Receptor Binding Protein
RNA	Ribonucleic acid
RT	Reverse Transcriptase
RM	Restriction Modification
SaPI	Staphylococcus Aureus Pathogenicity Islands
VC	Viral Cluster
VLP	Viral-Like Particle
VMR	Virus to Microbe Ratio

Table of Contents