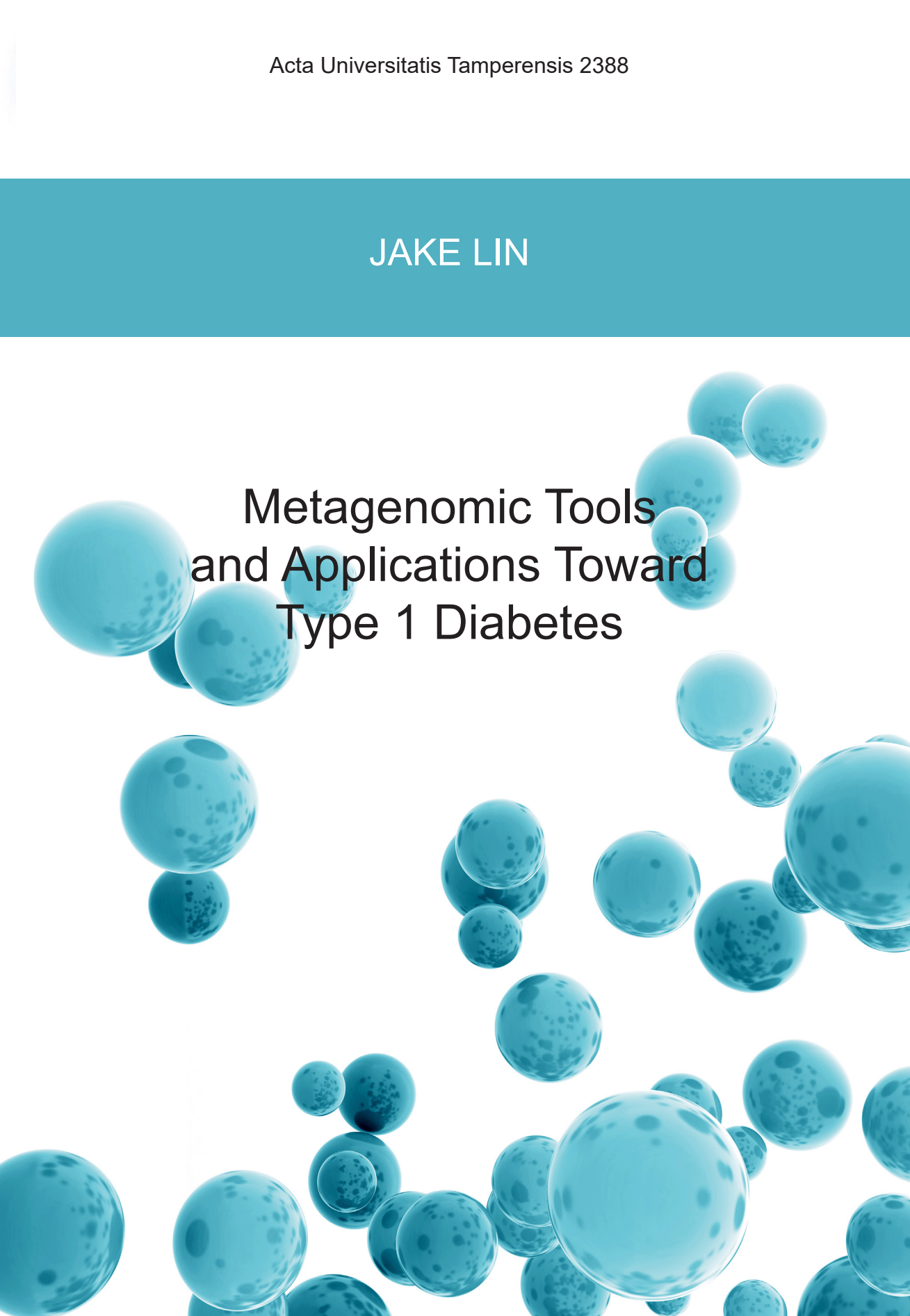


JAKE LIN

Metagenomic Tools and Applications Toward Type 1 Diabetes

The background of the cover features a collection of blue, semi-transparent spheres of various sizes. These spheres are scattered across the white background, with some appearing larger and more prominent than others. The spheres have a subtle texture and a slight gradient, giving them a three-dimensional appearance. They are arranged in a way that suggests movement or a dynamic field, with some overlapping and others floating independently.



JAKE LIN

Metagenomic Tools
and Applications Toward
Type 1 Diabetes



ACADEMIC DISSERTATION

To be presented, with the permission of
the Faculty Council of the Faculty of Medicine and Life Sciences
of the University of Tampere,
for public discussion in the auditorium F114 of the Arvo building,
Arvo Ylpön katu 34, Tampere,
on 25 June 2018, at 12 o'clock.

UNIVERSITY OF TAMPERE

JAKE LIN

Metagenomic Tools
and Applications Toward
Type 1 Diabetes

Acta Universitatis Tamperensis 2388
Tampere University Press
Tampere 2018



University of Tampere, Faculty of Medicine and Life Sciences
Finland
Charles University
University Hospital Motol
Prague, Czech Republic

Supervised by

Professor Matti Nykter
University of Tampere
Finland
D.Sc. (Tech.) Reija Autio
University of Tampere
Finland
Professor Heikki Hyöty
University of Tampere
Finland

Reviewed by

Docent Christophe Roos
University of Helsinki
Finland
Docent Tarja Sironen
University of Helsinki
Finland

The originality of this thesis has been checked using the Turnitin OriginalityCheck service in accordance with the quality management system of the University of Tampere.

Copyright ©2018 Tampere University Press and the author

Cover design by
Mikko Reinikka

Layout by
Sirpa Randell

Acta Universitatis Tamperensis 2388
ISBN 978-952-03-0768-4 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 1897
ISBN 978-952-03-0769-1 (pdf)
ISSN 1456-954X
<http://tampub.uta.fi>

Suomen Yliopistopaino Oy – Juvenes Print
Tampere 2018



ABSTRACT

Large consortiums are using next generation sequencing (NGS) to study the roles of microbial and virus exposures and population dynamics on human autoimmune diseases. These international collaborations are producing large collections of high throughput data in the form of metagenomic sequences. This thesis focuses on bioinformatics tools to support data visualization, high throughput virus extraction pipeline and subsequent application of these tools in an investigation concerning microbiome risk towards development of Type 1 diabetes (T1D).

Impacting mostly children and juveniles, T1D has no known cause and the disorder is characterized by insulin deficiency due to host immune system destruction of insulin producing pancreatic cells. T1D has a known genetic risk marker though a majority, exceeding 80% of new cases do not have a T1D diabetic first degree relative. Multiple viruses and more recently, gut bacterial development have been associated with increase T1D risk. I will discuss historical viral associations and especially Enterovirus as it has been found in pancreatic tissue. Metagenomic sequencing data assayed from regular collection of stool samples from longitudinal studies enrolling high risk children are optimal designs to investigate microbial and viral factors, especially using matched subject designs to control for locality and age. Different microbial exposure patterns and autoimmune milestones, such as autoimmunity can be model statistically.

The primary aim of the tools development is to allow researchers without deep computation backgrounds to perform virus population profiling directly from high throughput sequence samples while allowing custom parameter updates and provide usable and insightful visualizations. Performance was greatly enhanced using parallel design and implementation of centralized database. Another omics tool was developed to allow for plotting of dense network across multiple organisms, such as human, mouse, bacteria and yeast, while supporting circular and other layouts. These tools are web based, open sourced and installation free. The analysis results are derived from Finland based Diabetes Prediction and Prevention study, using 96 samples from 18 matched children and three time points. Bacterial profiling and diversity are analyzed using well cited open sourced methods and statistical packages in R. We also attempted mining bacterial and virus relationships via construction of time based correlation networks together with web based filtering.

This thesis contributes innovative web based tools for metagenomic virus extraction profiling and visualizing genomic networks. Novel microbial T1D autoimmunity associations were reported in the analysis results and bacteriome phage correlation networks were constructed and analyzed in the context of autoimmunity. The developed tools are open sourced and actively used by many research institutes.

FINNISH ABSTRACT

Nykyaikaisten sekvensointiteknologioiden kehityksen myötä on mahdollista tuottaa laajoja ja metagenomiikka-aineistoja. Näiden aineistojen avulla on mahdollista selvittää erilaisten mikrobiaalisten ja viruspohjaisten altistusten sekä populaatiodynamiikan vaikutusta ihmisen autoimmuunisairauksiin. Tässä työssä kehitetään laskennallisia menetelmiä, joiden avulla näitä suuria mittausaineistoja voidaan käsitellä, analysoida ja visualisoida. Lisäksi kehitettyjä työkaluja sovelletaan ykköstyypin diabeteksen tutkimukseen. Tavoitteena on kartoittaa mikrobiaalisia riskitekijöitä, jotka voivat johtaa sairauden puhkeamiseen.

Ykköstyypin diabetes todetaan tyypillisesti jo nuorena ja sairastuneella henkilöllä immuunijärjestelmä tuhoaa insuliinia tuottavia haimasoluja, mikä aiheuttaa haitallisen matalan insuliinitason. Sairauden puhkeamiseen johtavia syitä ei tunneta tarkkaan, mutta useat virukset kuten enterovirus ja viimeisimmän tutkimustiedon mukaan myös poikkeava suoliston bakteerikannan kehittyminen on yhdistetty kohonneeseen sairastumisriskiin. Näitä viruspohjaisia ja mikrobiaalisia tekijöitä voidaan tutkia lasten ja nuorten ulostenäytteiden metagenomiikkaa tarkastelemalla ja erityisesti saatujen aineistojen tilastollisella ja laskennallisella analyysillä.

Työn tavoitteena on ollut kehittää laskennallisia menetelmiä, joiden soveltaminen ei vaadi syvällistä laskennallisten tieteiden hallintaa. Toisaalta laskennalliset työkalut on suunniteltu yleishyödyllisiksi niin, että niitä voidaan mukauttaa erilaisiin sovelluksiin sopiviksi esimerkiksi parametreja säätämällä. Menetelmien toteutuksessa on hyödynnetty rinnakkaistettua laskentaa ja keskitettyjä tietokantaratkaisuja, jotka mahdollistavat tehokkaan analyysin. Kehitetyt työkalut on suunniteltu käytettäväksi web-pohjaisesti eikä käyttäjän siten tarvitse erikseen asentaa ohjelmistoja. Työn soveltavassa osiossa lähtökohtana on ollut laajan diabeteksen ennustamis- ja ehkäisy tutkimuksen yhteydessä kerätty aineisto, joka koostuu 96 näytteestä, jotka on kerätty 18 lapselta kolmena eri ajankohtana.

Työssä on kehitetty innovatiivisia web-pohjaisia työkaluja metagenomiikka-aineistojen analysointiin ja visualisointiin. Työn soveltavassa osiossa on analysoitu laajoja ykköstyypin diabetes -aineistoja ja raportoitu uusia sairauden etenemiseen vaikuttavia vuorovaikutussuhteita. Kehitetyt menetelmät perustuvat avoimeen lähdekoodiin ja menetelmät ovat tällä hetkellä aktiivisessa käytössä useissa tutkimusryhmissä.

LIST OF ORIGINAL PUBLICATIONS

This dissertation is composed of the following original publications, which are referred to in this thesis by Roman numerals I, II and III. They are reproduced here by permission.

- I Lin J, Kreisberg R, Kallio A, Dudley AM, Nykter M, Shmulevich I, May P, Autio R (2013) POMO: Plotting omics analysis results for Multiple Organisms. *BMC Genomics* 14:918. doi:10.1186/1471-2164-14-918
- II Cinek O, Kramna L, Lin J, Oikarinen S, Kolarova K, Ilonen J, Simell O, Veijola R, Autio R, Hyöty H (2016) Imbalance of bacteriome profiles within the Finnish DIPP study: parallel use of 16S profiling and virome sequencing in stool samples from children with islet autoimmunity and matched controls. *Pediatric Diabetes* doi:10.1111/pedi.12468
- III Lin J, Kramna L, Autio R, Hyöty H, Nykter M, Cinek O (2017) Vipie: web pipeline for parallel characterization of viral populations from multiple NGS samples. *BMC Genomics* 18:378 doi:10.1186/s12864-017-3721-7

ABBREVIATIONS

AIDS	Acquired Immunodeficiency Syndrome
CVA	Coxsackievirus A
CVB	Coxsackievirus B
DNA	Deoxyribonucleic Acid
DIPP	Diabetes Prevention and Prediction
EV	Enterovirus
FN	False Negative
FP	False Positive
GEXP	Gene Expression
GRN	Gene Regulatory Network
HAdV	Human Adenovirus
HCMV	Human Cytomegalovirus
HMP	Human Microbiome Project
HPC	High Performance Computing
KEGG	Kyoto Encyclopedia for Genes and Genomes
LPS	Lipopolysaccharide
NGS	Next Generation Sequencing
NPOD	Network for Pancreatic Organ Donors with Diabetes
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
POMO	Plotting Omics Networks for Multiple Organisms
PPI	Protein Protein Interaction
PPV	Positive Predictive Value
QC	Quality Control
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
TEDDY	The Environmental Determinants of Diabetes in the Young
TN	True Negative
TP	True Positive
TPR	True Positive Rate
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes

TABLE OF CONTENTS

ABSTRACT	3
FINNISH ABSTRACT	4
LIST OF ORIGINAL PUBLICATIONS	5
ABBREVIATIONS	6
1 INTRODUCTION	9
2 REVIEW OF THE LITERATURE	13
2.1 Human Microbiome Project	13
2.1.1 T1D Population Studies and Cohorts	14
2.2 T1D natural history	16
2.3 Virus T1D associations	17
2.4 Bacterial T1D Associations	19
2.5 Metagenomics analysis	20
2.5.1 Bacteriome processing	21
2.5.2 Virome processing	22
2.5.3 Pipeline development	22
2.6 Network visualization	25
2.6.1 Systems biology	25
2.6.2 Biological networks and interaction types	25
2.6.3 Cytoscape	26
2.6.4 Circos	26
2.6.5 Limitations	27
3 AIMS OF THE STUDY	29
4 BIOINFORMATICS ARCHITECTURE AND METAGENOMICS MATERIALS AND DESIGN	30
4.1 Open sourced and web programming	30
4.1.1 Interface design and workflow	31
4.2 Reference Databases	33
4.2.1 Viral databases	33

4.3	Diabetes Prediction and Prevention and gut microbiome association to autoimmunity	34
4.3.1	Subjects	35
4.3.2	Samples	35
4.3.3	Data availability	35
4.3.4	Hypothesis and motivations	36
4.3.5	Rarefaction and subsampling	36
4.3.5.1	OTU construction and model application	37
4.4	Simulation data and validation samples	37
4.4.1	POMO networks	37
4.4.2	Vipie microbiome samples	38
4.4.3	Simulation data	39
4.4.4	Blacklisted chimeric vectors	39
4.5	Statistical learning	40
4.5.1	Random forest	41
4.5.2	Regression and classification	41
4.5.3	Clustering	42
4.5.4	Sensitivity and precision	42
4.5.5	Shannon diversity	42
5	RESULTS	44
5.1	Plotting omics networks for multiple organisms (Publication I)	44
5.1.1	Nodes, edges and styling	44
5.1.2	Model organisms	45
5.1.3	Brain cancer genomic rearrangements	45
5.1.4	POMO Edge filtering and bundling	47
5.1.5	POMO example metagenomic application	47
5.2	Investigating microbiome associations in early islet autoimmunity (Publication II)	48
5.2.1	Viruses found in stool virome	48
	Gut bacteriome	48
5.2.2	Integrating virome with bacteriome	49
5.3	Vipie web based virus profiling pipeline for multiple samples (Publication III)	51
5.3.1	Mock data precision and recall results	51
5.3.2	Interface design, parameters and status communications	52
5.3.3	Web results and interactive features	56
5.4	Performance and results management	56
6	DISCUSSION	57
7	ACKNOWLEDGEMENTS	63
8	REFERENCES	65
9	ORIGINAL ARTICLES	75

1 INTRODUCTION

Microbiome, as collectively defined and ordered from smallest genome to largest, consists of viruses, bacteria, and fungi at a given site. Metagenomics therefore defines the field to study and recover microbiome genetic material, the microbiota using next generation sequencing (NGS) directly on samples taken from any host tissues or environment sites [Eisen 2017]. Demonstrated by the Human Microbiome Project (HMP) [Human Microbiome Project Consortium, 2012], many millions of microbiome reads can be recovered from NGS computer systems from 300 healthy individuals and sampled from nasal cavity, mouth, skin, gastrointestinal tract and vagina. These five sites are known to contain high amounts of bacterial and viral population, though microbes have been found in other human samples, including from inner ear channel and blood. In essence and similar to Human Genome Project, HMP seeks to define a golden and healthy microbiome reference profile. This reference is in many ways far more complex than the human reference genome as bacterial cells far outnumber human cells in population count. Contributing to the complexity, gut bacterial diversity fluctuates based on subject age, health, diet, geography as well as viral phage abundance. While HMP also investigates virus population within the same sites, the challenge to establish a standard viral reference is highly problematic due to difficulties with identification due to their small genome size, fast mutation rate and lack of a known universal marker. Viruses have always been an integral part of vaccine research as the common but deadly influenza is responsible for an average of 500,000 deaths a year. The infamous 1918 Spanish flu pandemic outbreak killing estimates of 50 million people [Patterson & Pyle 1991], about 2.5% of the global population. Peter Medawar, winner of a Nobel in Medicine in 1960 for his work on acquired immune tolerance, described a virus philosophically as ‘a piece of nucleic acid surrounded by bad news’. As such, some bad news bacterial strains include the deadly tuberculosis and anthrax strains as well as *Salmonella* strains responsible for food poisoning.

There is increasing evidence for gut bacterial imbalances playing pivotal roles in chronic immune disorders such as Type 1 diabetes (T1D), Celiac Disease and IBD [Iweala & Nagler 2006, Cohn et al. 2014, Lernmark 2016, Vatanen et al. 2016]. With a rising disease incident rate exceeding population growth and onset discordance greater than 50% in monozygotic twins, T1D is one of the most common childhood immune disorders. T1D, an irrecoverable and lifetime affliction, is an autoimmune disease where the host immune system targets and selectively destroys the pancreatic beta cells resulting in insufficient amounts of insulin necessary for maintaining safe blood glucose levels. Insufficient amounts of insulin results

in ketoacidosis, reliance of liver ketones as an energy resource and concurrent symptoms include nausea, frequent urination, thirst, mood swings and overall weakness and if untreated eventual organ damages leading to possible death. T1D is accurately predicted and preceded by the confirmation of pancreatic islet cell autoantibodies (ICA) [Bottazzo et al. 1974, Atkinson 2012]. The asymptomatic period between ICA seroconversion and T1D varies between months to years, ICA is a clear signal of an ongoing immune attack on the insulin producing beta cells. While the exact cause or trigger for the initiation of autoimmune intolerance is unknown, viruses have long been suspected and particularly coxsackievirus B (CVB) serotypes in the enterovirus (EV) genus. Even though some strong associations have been published [Hyöty et al. 1995, Filippi et al. 2008], causality has not been proven [Coppieters et al. 2012, Atkinson 2012]. Notably, EV sequences have been found in donated fresh pancreatic tissue from patients recently diagnosed with T1D [Krogvold et al. 2015]. The highest genetic susceptibility is associated with the Human Leukocyte Antigen (HLA) class II marker and allele DR3/DR4 genotype [Redondo et al. 2001], contributing up to 60% and odds ratio of approximately 6.8 [Concannon et al. 2009] according to Gene wide association studies (GWAS), more than 3 times the risk associated with the next highest ranked loci of *INS* and *PTPN22* [Concannon et al. 2009] markers.

The potential research benefits of metagenomics application within T1D and other diseases with suspected environmental causes can not be disputed as metagenomics technology allows for unbiased and systematic detection of previously unknown and uncultured bacteria and human viruses. Labs and cohorts are investigating potential risks and as well as beneficial associations [Vaarala 2012], particularly involving early childhood exposures during innate immune system development [Lehuen et al. 2010]. The timing and beneficial role of childhood non-pathogenic exposures are the principles of hygiene hypothesis [Strahan 1989] first proposed by Strahan towards the large incidence increases with hay fever allergy.

Although mass virus matter detections have been enabled with metagenomics and associated preparation and enzymatic advances, the identification and profiling of the virus population are still challenging due to lack of universal genome marker and also relative fast evolution pace, estimated to be 6 magnitudes faster than human [Duffy et al. 2008]. Using reverse transcription, ribonucleic acid (RNA) in addition to de-ribonucleic acid (DNA) viruses can be capably detected. These important advances are enabling international cohorts to systematically study the role of environment in autoimmune diseases with no known causes. Study designs that allows for longitudinal sample collection is particularly applicable in T1D due to motivation to investigate viral infection history and gut bacterial dynamics prior to autoimmunity. It has been suggested that up to 15% of Type 2 Diabetes are masquerading by T1D and exhibits autoimmunity [Palmer et al. 2005]. As a result of insulin resistance, T2D impacts adults and is primarily driven by obesity from combinations of modern diets and lack of physical activity [Tattersall 2009].

Metagenomics together with high throughput NGS sequencing motivates and also exposes the need for programs and tools that aim at data management, quality control, population extraction, diversity profiles and genome functions. Furthermore, intuitive and easy to use tools are needed to support visualization aiming at finding patterns and clusters within dense networks. Within systems biology, machine learning and data mining tools are routinely producing complex multidimensional genomic networks that are incomprehensible unless displayed in appropriate layout and context [Gehlenborg et al. 2010]. Together with genomic network plotting, this thesis will focus on virus extraction web based pipeline using NGS metagenomic files, their applications for microbiome research, and an analysis involving bacteria and virus interactions towards T1D autoimmunity and subsequent disease onset.

This thesis is based on published articles and provides novel web based tools for genomic network visualization and metagenomic virus profiling on multiple samples and subsequent application of virome profiling in T1D. The tools are also implemented with open sourced architecture and the chapter 2 will introduce the existing tools and technological background of visualization tools and also pipelines to profile bacteria and virus populations and diversities. These bacterial microbial computational pipelines have the primary aims of supporting taxonomy identification and estimating protein production.

At the same time, a brief historical summary of T1D genetic disposition and associated environmental factors together with perspective of known viruses and historical implications will also be discussed. Chapter 3 covers materials and methods where the methods will include programmatic tools for open sourced web programming languages and model organism reference database components needed for POMO web app, presented in Publication I. Materials listed include matched case and control samples from early T1D autoimmunity study in Diabetes Prediction and Prevention (DIPP) [Ilonen et al. 1996, Haller & Schatz 2016] project used in Publication II and also several HMP samples used for Publication III. I will also describe bacterial 16S amplicon and metagenomic whole genome shotgun (WGS) analysis programs and terminologies and also statistical methodologies applied. Chapter 4 begins with describing the background of plotting omics graphs and shows example results from human and mouse. The results will include graphs from a metagenomic project drawn using POMO. Publication II investigates role of microbiome imbalance in early age T1D. The study focuses on bacteriome diversity and population dynamic via correlation networks over time integrated with phage findings. Publication III presents Vipie, a web based pipeline capable of sensitively processing multiple viromes with custom assembly and mapping parameters. Results are securely accessible and can be viewed as interactive charts, maps and searchable tabular results. Publication II is the first study to integrate bacterial and viral NGS findings towards T1D. The investigation is novel as an attempt to systematically identify the bacterial host of the gut virus CrAssPhage using interactive and filterable NGS correlation networks. The results chapter includes a scenario of applying POMO visualization for optimized bacterial community tuning. Chapter 5

concludes with discussion on practical usability of the bioinformatics tools and important applications in different projects and cohorts.

Along with all tools involving big data analysis, there are some practical challenges and design restricted dependencies along with statistical modelling and variable interpretation. Particularly with metagenomic data analysis involving population composition changes over time, sequence analyses are confronted with multiple challenges due to wide sample variances in depth and quality that potentially shroud signal strengths. The large percentage of unknown viral mapped reads, so called dark matter sequences stemming from large viral diversity exasperated by lack of common universal reference and relatively limited number of known complete viral genomes. We theoretically discuss current limitations of viral research within T1D and potential viability of construction of a comprehensive virome roadmap.

2 REVIEW OF THE LITERATURE

This chapter introduces important microbiome projects applying microbiome towards human health. T1D prevalence and primary genetic risks are briefly presented along with historical microbial associations, particularly viral. Microbiome NGS processing, technologies and applications are discussed. In addition, I will go over leading network visualization software in support of systems biology interaction networks.

2.1 Human Microbiome Project

Metagenomics has advanced the field of microbiology by enabling the detection and estimating the diversity profile of all organisms, including viruses from any environmental sample via integration of next generation sequencing and modern molecular preparation methods. Traditional microbiology relied upon cultivated samples and extreme biases against organisms that cannot be cultured. Particularly neglected were microbes inclusive of archaea, bacteria and viruses. Bacteria and viruses have had long historical human health associations, and they have been implicated in autoimmune and infectious disorders. While less obvious, there are also frequent mutual beneficial symbiosis relationships. The creation of the Human Microbiome Project (HMP) [Human Microbiome Project Consortium, 2012] is an integrated attempt at cataloguing and detailing microbial profiles across different human sites. Its success and landmark findings of large microbial sequences, bacteria and virus particularly in the human gut and vagina, were revolutionary and profound. Within the human body, the bacterial cells are estimated to outnumbered human cells by ten fold and using open reading frames (ORF) analysis, potentially many more gene products [Savage 1977, Abubucker et al. 2012]. At the same time, it was revealed that there are different bacterial colonies and patterns across nasal, ear and skin sites. These important findings together with the HMP resources enable investigators to address hypothesis concerning microbiome development and population with human health, in turn providing microbial composite references for healthy subjects over different body sites and developmental ages.

2.1.1 T1D Population Studies and Cohorts

T1D is well known to be an autoimmune disorder with a complex environmental role as identical twins have only 27% onset concordance and majority of new subjects do not have Type 1 diabetic first degree relatives. Motivated by strong historical suspicions and epidemic associations implicating gut bacteria and viruses as potential triggers (further discuss in 2.3) for T1D [Filippi & von Herrath 2008, Atkinson 2014], large international T1D cohorts have followed the steps of HMP to apply metagenomics technologies to investigate gut bacteria dynamics and virus exposures found in stool, plasma and other human samples. At the forefront of these cohorts is TEDDY (The Environment Determinants of Diabetes in the Young) [TEDDY Study Group 2008], it has enrolled almost ten thousand genetically at risk children living in US, Germany, Finland and Sweden to adjust for geography and race as they are known to be statistically significant.

In northern Europe spanning Scandinavia and Finland (57.6 per 100,000), risk of T1D are folds higher than other locations, as shown in Figure 1, nearly 100 times higher than some countries such as China (0.6 per 100,000) and Venezuela (0.1 per 100,000) [Atkinson 2008]. The issue of diet likely plays a role, as China and Venezuela have rice and corn based staple diets and among the lowest in incidence. TEDDY has impressively collected more than 30,000 microbiome samples from stool, nasal and plasma from case and control children, covering T1D autoimmunity and onset. Interestingly, celiac disease (CD) concordance with T1D is around 6%, much higher than general population of 1% risk and statistically significant [Cohn et al. 2013]. Together with gene expression, proteomics and metabolomics integrated with the comprehensive metadata such as health and diet history, this collection will form the central materials and design for many important T1D and autoimmune related investigations and findings for years to come. Within Discussions chapter, we covered the experience of applying Vipe pipeline introduced in Publication III, to profile virus populations in thousands of TEDDY stool microbiome samples.

Adding to the mystery is that Russian Karelia bordering Finland and with many of its population having similar genetic and cultural heritage, has 6 times less risk of T1D although exposed to more microbes [Kondrashova et al. 2008]. DIABIMMUNE was created expressly to study gut microbiome concerns toward T1D in Finland, Estonia and Russia [Vatanen et al. 2016]. Interestingly and related to the hygiene hypothesis, Vatanen and colleagues concluded that the complexity of the gut bacteriome due to the birth condition differences between the countries. Other large cohorts include US based SEARCH for Diabetes in Youth Consortium within the United States (Dabelea et al. 2011), aiming to investigate difference between T1D and T2D by identifying all cases in American children under 20. Finland based Diabetes Prediction and Prevention (DIPP) study recently celebrated its 25th anniversary and the study followed at risk children, based on HLA screening, from birth until T1D onset or 15 years old.

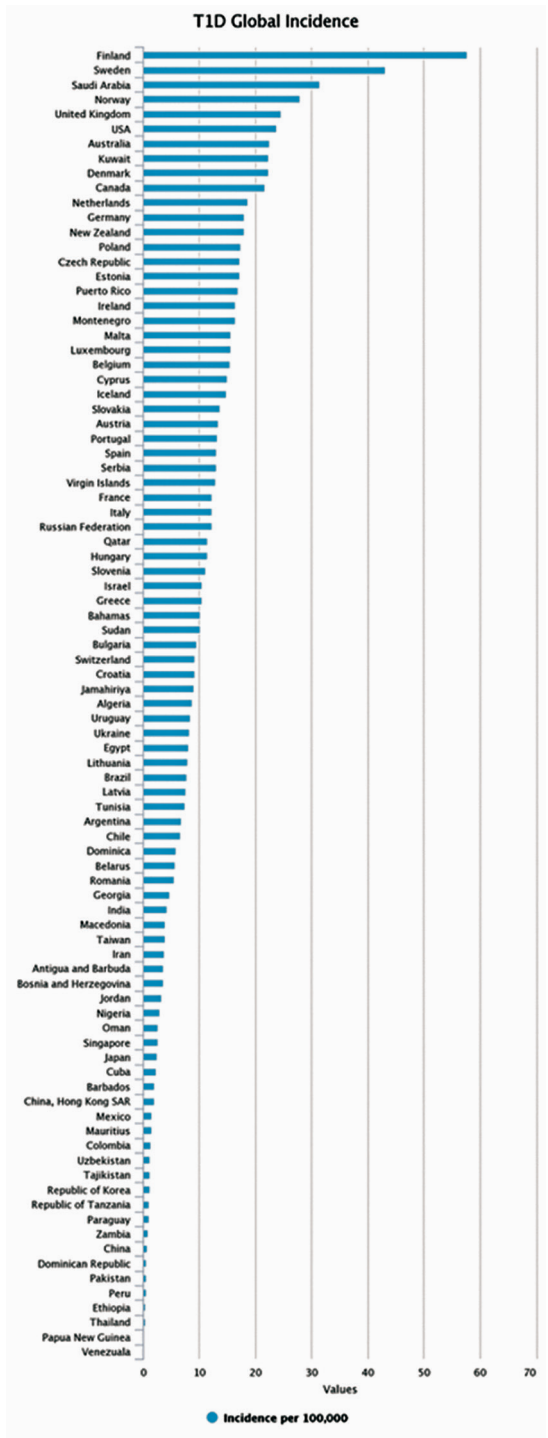


Figure 1. Global T1D incidence – Plotted from Diabetes UK (2013) statistics.

2.2 T1D natural history

T1D with an unknown etiology, is a chronic disease requiring daily external insulin. The disease, rising at about 3% per year in most developed countries, has no cure and mostly impacts young people. T1D is classified as an autoimmune disorder as the pancreatic islet beta cells, responsible for producing insulin, are killed selectively and subsequently depleted over time by the host immune system. T1D is known to have environmental factors as disease concordance rates for identical twins are about 30% and recent trends points to increase new incidence without first degree relatives with the disease [Atkinson 2012]. The classification of T1D as an autoimmunity disorder is further solidified with the discovery of islet cell autoantibodies (ICA) [Bottazzo et al. 1974], subsequently antibodies to insulin (IAA), glutamic acid decarboxylase (GADA), tyrosine phosphatase-related islet antigen (IA2A) and zinc transporter (Znt8A) [Pihoker et al. 2005]. These autoantibodies confirmed that beta cells were destroyed from breakdown of self tolerance. ICA prediction interestingly is tightly intertwined with increasing number of autoantibodies, with two or more autoantibodies, close to 85% have been reported as likelihood for T1D onset and more than 90% of T1D have minimum 1 autoantibody [Pihoker et al. 2005] though the time period between diabetic onset and ICA seroconversion varies greatly, between months to many years. Shown in Figure 2 below, though the age of autoimmunity is uncertain, it has been shown that IAA tends to peak around two years old and GADA incidence peaking much later [Taplin & Barker 2008]; as autoimmunity is highly predictive of T1D and irreversible, a revised and key research question falls on environmental factors such as viral and bacterial infections prior to the trigger of autoimmunity. The time period between autoimmunity and disease onset has high variability and T1D symptoms are marked by hunger, thirst, frequent urination, sudden bed wetting and fatigue with mood changes. Organ damage and death results from untreated hyperglycemia, the estimated lifespans of T1D patients are 10 years less than relative population averages. According to American Association of Clinical Endocrinologists (<http://outpatient.aace.com/type1-diabetes/the-burden-of-type-1-diabetes>), T1D burden on worldwide society is 14.4 billion annual US dollars and anticipated to triple by year 2050.

Alongside seasonal frequencies, suspected environmental triggers are EVs, gluten introduction, north to south gradient, lack of vitamin D and a hygiene theory [Strachan 1989] suggesting that young children are not sufficiently exposed to non-pathogenic microbes during the development of immune system within the first years of life. Although it is known that genetics play important roles, due to first degree relative discordance rates, it is also known that there are environmental factors. The dramatic increase in T1D also far exceeds general population growth. HLA-DR3/DR4 genotype combination marker in chromosome 6 has the highest risk, accounting for 90% of the disease [Atkinson 2012] and certain alleles post odds ratio of 6. As such, the Finnish population has the highest average T1D, approaching 60 per 100,000. It is interesting and as reported [Kondrashova et al. 2008] that neighboring Russian Karelia, with similar genetic HLA genotypes, have folds

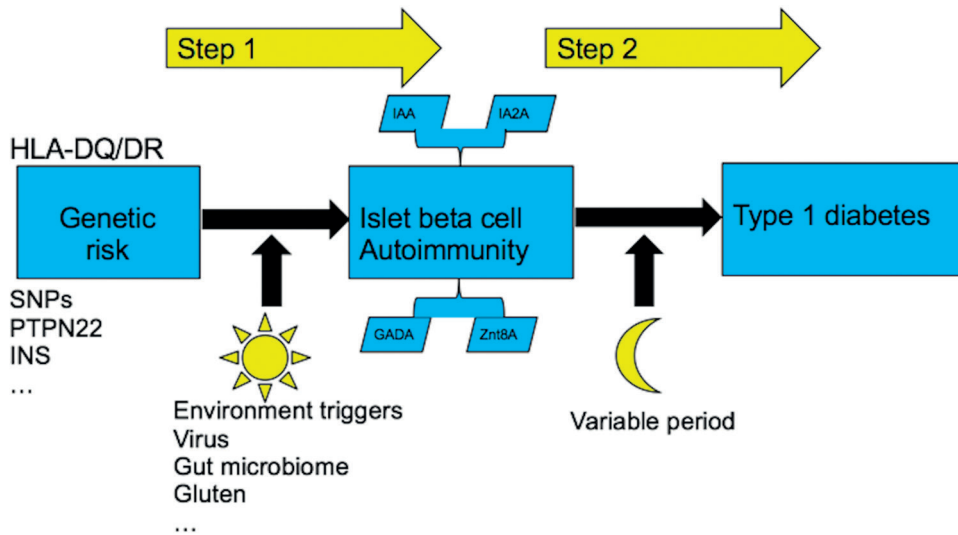


Figure 2. T1D autoimmunity precedes disease. Islet cell autoantibodies are predictive of disease while selected HLA genotypes represent the highest genetic risk.

less incidence. Swedish population also has relative high risk for T1D and has the highest rate of 3% regarding celiac disease [Ascher et al. 1991], double the global population average. Current investigated environmental risks are diet and viral factors. DIABIMMUNE is a recent study cohort involving Finland, Estonia and Russia [Yassour et al. 2016, Vatanen et al. 2016] children, investigating diet and microbial diversity differences. TEDDY, a NIH/JDRF funded prospective from birth cohort [Lee et al. 2015] involving 8500 children from Sweden, Finland, Germany and United States, are performing NGS extraction from regular stool, nasal, plasma along with diet, health history, medication and other information. Together with metabolites and gene expression from plasma, this large and valuable repository will allow informatics and biostatisticians to apply system biology methodologies in addition to nested case control models. DIPP, based in Finland, began in 1994 and the details of its NGS pilot design used for Publication II is further described in chapter 3.

2.3 Virus T1D associations

Viruses and bacteria play important roles in human health. Influenza and polio are two well known diseases caused by virus infections. Viral diseases tend to have seasonality patterns and the infections can be acute or persistently dormant. For example, herpes infections can stay dormant in the human body for years before an acute outbreak. Norwegian scientist Gundersen proposed an infectious virus theory in 1927 [Gundersen 1927] after a mumps epidemic to explain an unexpected rise in T1D in a Norwegian

village. In the almost hundred years since that report, Coxsackie B virus (CVB) is the most implicated virus as causative agent. Pappenheimer in 1951 showed that CVB can cause pancreatic damage in mouse. Using mice, Coleman, Gamble and Taylor found excessive CVB4 antibodies within 3 months of disease onset [Coleman et al. 1973]. From that subliminal work, they went on to report coxsackievirus can induce diabetes in mice, and that importantly, there is a latency period between infection and glycaemia [Atkinson 2012]. With improving biomedical technologies and pancreatic organ collection such as the JDRF funded Network Pancreas Organ Donors (NPOD) [Pugliese et al. 2014], enterovirus sequences have been found in pancreatic tissues and often statistically cited as risk factor associated with beta cell autoimmunity and T1D [Hyöty et al. 1995, Filippi & von Herrath 2008, Laitinen et al. 2014] but also selective serotypes as protective [Ghazarian et al. 2013] exposure due to early innate immunity development. The DIPP study was the first to systematically and comprehensively screen for EV antibodies and to report CVB1 exposures significantly associated in children upon diabetes onset [Laitinen et al. 2014].

The timing of the development of immune system is also at the heart of the hygiene hypothesis [Strachan 1989] where the lack of microbial exposures in early childhood is being investigated for autoimmune disorders such as asthma, celiac disease and T1D [Stene & Nafstad 2001]. Circoviruses have been cited recently as protective, demonstrated by an increase in the non-autoimmune control population [Zhao et al. 2017] compared to children who developed pancreatic islet autoimmunity, though the finding was statistically significant, it was only 11 case control pairs. Very recently, strong signs of interferon, protein for signaling adaptive immunity activation, were found in enterovirus infected longitudinal blood samples from high risked T1D compared to paired children [Lietzen et al. 2018]. Notably investigators in DiViD, a Norwegian study, were the first to find enteroviral capsid protein, VP1, in fresh pancreas of living patients within weeks after diagnosed with T1D [Krogvold et al. 2015]. Though the findings were from six adults [24–35 years] recently diagnosed, the enterovirus VP1 protein pancreatic presence is supportive of the hypothesis that T1D can be triggered by a low grade and chronically persistent enterovirus infection whereas prior it had been assumed to be caused by an acute viral infection. Within a nationwide survey in Taiwan, enterovirus infections were found to be significantly associated with T1D, 5.73 per 10,000 compared to 3.89 per 10,000 enterovirus free children [Lin et al. 2015].

Table 1. Viruses associated with T1D from PubMed. Coxsackievirus B, in the enterovirus genus, has by far the highest co-occurring citation with T1D.

Species/Serotype	Taxonomy/Genus	PubMed T1D Citations	Other associated human ailments (Center of Disease Control)
Coxsackievirus B	ssRNA, Enterovirus	116	Encephalitis, Myocarditis, Grippe
Mumps rubulavirus	ssDNA, Rubulavirus	40	Mumps
Human Cytomegalovirus (HCMV)	dsRNA, Cytomegalovirus	24	Immunocompromised Hepatitis
Human Adenovirus	dsDNA, Mastadenovirus	21	Bronchitis, Myocarditis
Echovirus	ssRNA, Enterovirus	20	Cold, Meningitis
Rhinovirus	ssRNA, Enterovirus	5	Cold
Rotavirus	dsRNA, Reoviridae	5	Cold, Gastroenteritis
Norovirus	ssRNA, Norovirus	4	Gastroenteritis
Rubella virus	ssRNA, Rubivirus	4	Rubella
Coxsackievirus A	ssRNA, Enterovirus genus	2	Hand, foot and mouth disease; herpangina
Circovirus	ssDNA, Circovirus	1	None
crAssphage	ssDNA(unconfirmed)	1	None

2.4 Bacterial T1D Associations

As multiple pancreatic autoantibodies usually precede disease onset, it is clear that T1D is an immune T-cell driven disease, and it has been proven in animal models that a non-functional immune system prevents T1D [Buschard 2011]. T helper cells are activated by inflammation and the theory of a leaky gut and autoimmunity have been proposed [Bosi et al. 2006] and forms the main basis of bacteria-related T1D trigger. The HMP project revealed that there are magnitudes more bacterial cells in human gut than human cells. Importantly, it was shown that certain gut bacteria species are producers of short chain fatty acid (SCFA) butyrate metabolites. These butyrate metabolites, as an energy source, are vital to the health of intestinal epithelial cells [Donohoe et al. 2011] and a non-leaky gut. SCFA, known to be associated with high fiber diets, has also been shown to have beneficial roles in colon cancer [Bergman 1990, Lupton et al. 2004] via direct promotion of colorectal carcinoma cell apoptosis [Lazarova et al. 2004].

Cardwell and colleagues recently reported that children from caesarean (C-section) birth method have a 20% higher chance of T1D [Cardwell et al. 2008]. This finding lends supports to the aberrant gut bacteriome hypothesis as babies delivered from caesarean are initially exposed to bacteria composites from the hospital doctors and hands. It follows that in a natural birth, babies are exposed to the maternal birth channel where it is richly

populated with *Lactobacillus* and *Bifidobacterium* [Muller et al. 2015]; know to be important phyla for early gut microbiome development.

A clear symptom that is predictive of immune activity is inflammation and investigators have suggested bacterial dysbiosis, or gut microbial imbalance as a triggering mechanism for excessive inflammation. Dysbiosis represents an increasing instability and lowered bacterial diversity and thereby drives the likelihood of a leaky gut allowing gram negative bacteria to escape via the gut lumen and into nearby organs including the pancreas. Gram negative bacteria membranes consist of lipopolysaccharide (LPS) and capably releases pro-inflammatory cytokines [Babbas 2006] and can potentially increase havoc on the host immune system and help bring upon autoimmunity. In parallel with lowered alpha (intra-sample) diversity and microbial imbalance, the hygiene hypothesis [Strahan 1989] points to insufficient amounts of exposure to non-pathogenic microbes during early innate immunity development, particularly within modern and industrialized households. It has been shown that gut microbial is particularly robust during initial development and tends to become more stable and resembling adult systems after 2–3 years of life. Both short term [Jakobsson et al. 2010] and long term antibiotics usage have been shown to promote bacterial resistance, lower diversity and shift microbiota stability negatively [Jakobsson et al. 2010].

A recent work using DIPP samples found that the *Bacteroides* phylum population, measured as relative abundance, was significantly higher in case children prior to autoimmunity relative to control children. Moreover, the investigators were able to pinpoint the species to be *Bacteroides dorei* [Davis-Richardson et al. 2015]. The authors had 947 samples from 29 children with 47 controls though the children were drawn from the same Turku hospital, located in the southwest of Finland. As other reports have not yet confirmed this result [Vatanen et al. 2016], it underscores that the development and dynamics of gut microbial population is an active research area with important ramifications in the field of autoimmune disorders impacting children. Within Publication II, our design is similar though smaller with 18 case control matched groups and includes children from other Finnish cities Tampere and Oulu, we also found that imbalance within *Bacteroides* phylum is significant but that the species *Bacteroides vulgatus* were higher in controls compared to cases prior to autoimmunity, additional details are provided in results and discussions chapters.

2.5 Metagenomics analysis

Metagenomics design involve direct sequencing from environmental samples. The amounts of sequenced reads produced varies greatly and depend on the amount of genetic richness and sequencer capacity. Reflected in HMP and our experiments, samples typically average millions of reads resulting in gigabytes per sample file. The two central aims for bacteria and viral metagenomic experiments are identifying and profiling the population and also

answering the functional, or protein products of this population. Bacterial profiling most commonly uses targeted amplicon 16S rRNA gene to identify bacteria taxonomy. While cost effective, one limitation is that novel species cannot be identified. Leading tools and common processing steps to process 16S reads are listed below. Sequences from whole metagenome shotgun experiments are needed to address bacteria functionality. While not applicable to this thesis, MEGAN [Huson et al. 2011] and HUMAnN [Abubucker et al. 2012] programs were used and cited for functional analysis within the HMP. Virus profiling demands shotgun experiments as it lacks a universal marker gene. First, a quality control (QC) step focuses on quality control (QC) of read quality, coverage and length requirements. Second, using the QC outputs, the reads are passed into assembly methods to generate an intra-sample representative set of reads or contigs. This important step is commonly called de novo assembly and some virus profiling pipelines can choose to bypass this step and proceed directly to database mapping. The features and aims of recent published virus profiling pipelines are listed in Table 2, including Publication III.

2.5.1 Bacteriome processing

For bacteriome processing, with hundreds of citations, the two most popular and extraction programs are Mothur [Schloss et al. 2009] and QIIME [Caporaso et al. 2010]. Both programs are comprehensive and offer QC, assembly, qualitative and quantification functions. Recently, the creator of Mothur wrote an informative online article comparing the two tools (<http://blog.mothur.org/2016/01/12/mothur-and-qiime/>) and notes that both relied on similar reference databases, produced consistently good and comparable results. Mothur is a self contained program while QIIME in essence is a wrapper calling other published algorithms, hence easier to pass in custom parameters.

Our work in Publication II follows standard 16S sequence processing and in part based on standard of procedure recommended by Mothur and QIIME. Operational taxonomic units (OTU) are found from clustering of highly similar reads and taxonomy levels are assigned using SILVA [Quast et al. 2013] database. The processing pipeline details, including parameter assigned are further described in the next chapter. While QIIME and Mothur have functions for plotting and diversity analysis, we convert the QIIME format OTU table to text based using Biological Observation Matrix (BIOM) [McDonald et al. 2012]. The converted table and taxonomy distance matrix files subsequently act as inputs for R packages phyloseq [McMurdie & Holmes 2013], DESeq2 [Love et al. 2014] for diversity profiling, plotting and comprehensive statistical modelling and comparisons.

2.5.2 Virome processing

Compared to bacteria identification via 16S sequencing, virus identification with massive parallel sequencing samples is more complicated as viruses do not have a universal genome marker. As such for Publication II, after quality control filtering, we performed de-novo assembly using Velvet [Zerbino et al. 2008] to retrieve intra-sample contigs. The non-overlapping contigs are fed into local BLAST (Altschul et al. 1990) against all known viral genome database for identification. For population estimation, we randomly select 100,000 reads and then apply aligner BWA [Li and Durbin 2009] against virus list from BLAST. For Publication II, all identified human viruses are then confirmed using PCR and statistical testing on the extracted virus data matrix was done in R using conditional logistics testing.

2.5.3 Pipeline development

From the experience of publication II and having evaluating current virome pipelines (listed in Table 2 and adapted from Publication II), we reasoned that a web based tool capable of profiling multiple virome samples would benefit ongoing and future virus NGS studies.

Table 2. Comparing available NGS virus extraction pipelines – Vipie, presented in Publication II, is the first web based tool to allow multiple samples while integrating different *de novo* assembly methods.

Pipeline	Vipie [Lin et al. 2017]	ViromeScan [Rampelli et al. 2016]	VirusTAP [Yamashita et al. 2016]	Virome [Wommack et al. 2012]	MetaShot [Fosso et al. 2017]
Primary goal	Parallel analysis of multiple viral metagenomes from web and suited for molecular epidemiology studies.	To profile viromes using databases of existing eukaryotic viruses without assembly.	Identification of viruses in a sample, after a thorough elimination of known non-viral sequences.	Classification of all putative ORF found in a viral metagenome, characterization of viral communities.	Highly accurate and comprehensive workflow for host-associate microbiome classification on multiple samples.
Web based	Yes.	No.	Yes.	Yes.	No.
Outputs	Interactive table, plots, clustered heatmaps and raw downloads.	Static population pie charts. Sample based clustered heatmaps.	Contig based hits and seamless web BLAST interface.	Rich collection of sample source virome ORF and sequence categories.	A Krona graph and Interactive Taxonomy HTML table and csv file.
Source data	Paired-end reads; <i>fastq</i> format.	Single-end or paired-end reads; <i>fastq</i> format.	Paired-end reads. Accepts also single-end reads; <i>fastq</i> format.	<i>sff</i> , or <i>fastq</i> ; intended for the 454-generated metagenomes.	Paired-end reads in <i>fastq</i> format.

Pipeline	Vipie [Lin et al. 2017]	ViromeScan [Rampelli et al. 2016]	VirusTAP [Yamashita et al. 2016]	Virome [Wommack et al. 2012]	MetaShot [Fosso et al. 2017]
Trimming and filtering	YES, as the first step.	YES, after selection of viral reads, at the level of a bam file.	YES, as the first step.	YES: quality based; duplicate filtering; contamination	YES, as the first step.
De-novo assembly	YES, a choice of assemblers.	No.	YES, a choice of assemblers; done after subtraction steps.	No.	No.
Subtraction of human ref. and bacterial ribosomal sequences	Optional, only for the output of dark matter sequences.	YES, using Human Best Match Tagger. No for ribosomal.	YES, also other host databases available (mouse etc).	Not specified for human. Ribosome is removed using BLAST against rDNA db.	Yes, reports identification of human host reads and bacterial mappings.
Means of virus identification	(a) BLAST against a pan-viral database. (b) Remapping of original reads to the identified candidates.	Mapping to the members of the virus database using <i>bowtie2</i> [Langmead & Salzberg 2012]	BLAST search against the NCBI nt database.	Protein BLASTP upon two databases. Several tiers of classification of the ORFs.	Custom similarity workflow with hamming distance.
Virus database for identification	A custom database containing 20759 human, animal, plant and bacterial viruses.	Eukaryotic viruses only. Four custom databases available for download.	Specificity is maintained by the subtraction steps prior to assembly and BLAST search.	UniRef 100 peptide database, five annotated protein databases, MetaGenomes On-line.	TANGO [Alonso-Alemany et al. 2014] and NCBI Taxonomy.
Action when a read maps to different viruses	Score is split among the hit reference sequences.	Not specified.	Not specified.	Not specified.	Parsed for human endogenous retrovirus otherwise classify as ambiguous and discarded.

De novo assembly is pivotal as the step efficiently reduced the number of reads while optimizing for quality and maximizing relevant continuous read lengths. Listed in Table 2, VirusTap [Yamashita et al. 2016] is the only pipeline allowing *de novo* assembly though only one sample is allowed. Another limitation is that the *de novo* assembly is done after filtering and removal of human and bacterial reads, thereby eliminating all possible viral reads

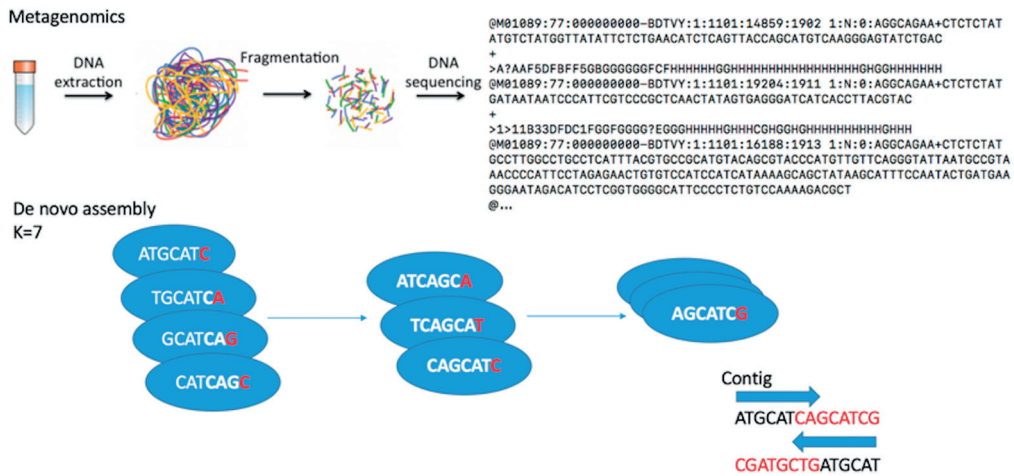


Figure 3. Simplified short read assembly flow using de Bruijn graphs. Using example kmer length of 7, unique sequences with 6 continuous overlapping bases are stored as unique nodes and contigs are extracted from graphs constructed from the nodes. Kmer size selection decides the number of nodes created and implies a tradeoff between specificity and sensitivity.

similar, or projects investigating viral and host molecular mimicry. Actually, molecular mimicry has been touted as a possible trigger for T1D [Atkinson 2001] and it has been noted that Coxsackie virus B and GAD auto-antigen that is expressed in islet beta cells have a similar sequence of amino acids [Atkinson 1997]. In addition, casein protein found in cow milk have been shown to have cross reactivity with human insulin [Adler et al. 2011] thereby potential trigger for the pancreatic insulin autoantibody. For publication III Vipie pipeline, we have integrated five published assemblers, Velvet [Zerbino et al. 2008], MetaVelvet, IDBA, MEGAHIT (formally known as DE) [Li et al. 2015], and ABySS. These well cited assemblers, capable of *de novo*, were integrated based on our own experiences, collaborator practices and review recommendations [Narzisi & Mishra 2011].

Shown and simplified in Figure 3, genomic material from different organisms are amplified within a sample and *de novo* assembly attempts to construct long reads, known as contigs without using a reference genome. As metagenomic analysis tends to lack or a composite of many reference genomes, *de novo* assembly is often a key step. The choice of kmer size is the most important as larger kmer implies longer overlaps, so higher specificity but larger kmers also result in lower sensitivity, since lower kmer increase the chances of overlaps and can lead to more contigs. In Figure 3, an example shows kmer length of 7 where sequences are split into that length, and sequences with a K-1 similarity are grouped, provided by Velvet initial function velvetg designed to build group of nodes and also meeting the coverage cutoff. The node also includes its reverse complement as a twin identifier. Velvet, the second step optimally connects the nodes, and the path becomes the

contig. The practical choice of K essentially is a tradeoff between specificity and sensitivity, since longer K has less noisy overlaps but also lower sensitivity due to less number of nodes.

Viruses, due to its small genome size and large divergence, are under represented across bioinformatics databases. For example, currently within the EBI (<https://www.ebi.ac.uk/uniprot/TrEMBLstats>) resource, only 3% of sequences submitted are classified as viral and only 7,512 complete genomes are available within NCBI (<https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi>). Strikingly, it has been estimated that there are at minimum over 300,000 mammalian viruses [Anthony et al. 2013]. For Publication II, we found that 74% were likely viral reads mapping to unknown viruses. The viral findings were only improved after including for RNA viruses, made possible after applying reverse transcription in sample preparation. Partially listed in Table 1, viruses play important roles in many facets of human health, and the need for a virome roadmap [Delwart 2013], to systematically identify novel human and animal calls for improved standards and tools.

2.6 Network visualization

In this section different types of systems biology interaction networks together with prominent visualization tools are introduced.

2.6.1 Systems biology

Systems biology is based on integrating computer science and mathematics to model genomic interaction networks of biological systems. Systems biology advocates that to understand biological functions fully, the network relationships and structure of the relevant genomic component states must be mapped and modelled. While it is important to know the complete list of the genes and their protein products, the characterizing will remain a description and less insightful in investigating how phenotypes, the complex functions and traits of genomic system interact. This paradigm was first introduced by von Bertalanffy where described dynamic interactions as central problem in modern science [von Bertalanffy 1950]. Still today, the main motivation of biomedical and cancer research is to detect signs of malfunction within the system, equivalent to finding genomic patterns or aberrant network interactions.

2.6.2 Biological networks and interaction types

One prominent type of systems biology network is protein-protein interaction (PPI) networks. As proteins seldom act alone, co-activations are approximated by correlation scores and then clustering. Most proteins are classified as stabled or transient. Stabled

interactions tend to be permanent with high structural affinity where transient proteins are temporal, consist of multiple states and associated with signaling. Gene regulatory networks (GRN), also called protein-DNA interaction, represent proteins such as transcription factors and other chromatin-associated proteins control gene expression. Gene regulatory elements can act as promoters, elevating gene expression, or reduce expression, repressors. Metabolic and enzyme networks are important tools to investigate metabolism states and relevant interactions. As there are well defined dependent chemical reactions prior to each enzyme activations, metabolic networks are time series and include helper molecules, also called co-factors. Certain biochemical and metabolic pathways are well conserved while some genetic pathways are unique to humans. Kyoto Encyclopedia of Genes and Genomes [Kanehisa & Goto 2000] and WikiPathways [Slenter et al. 2018] are good and active resource for PPI, GRN and metabolic pathways stratify by organisms and domains. Chromosomal structural events, such as copy number, large deletions and insertions and gene fusions have obvious important ramifications in development and particularly in cancer they are somatic, meaning that the structural aberrations are specific to cell types or even subcellular clones. Copy number, involving whole chromosomes, are usually detected directly from sequencing platforms. Large aberrations can be detected with existing bioinformatics tools [Chen et al. 2010, Bressler et al. 2012] though gene fusion events are less straightforward.

2.6.3 Cytoscape

Cytoscape [Shannon et al. 2003] was the first comprehensive open sourced software to support molecular networks while offering a framework and settings for defining layouts and graph element styling. The styling attributes, using color and size, essentially are annotation descriptors and defined as mappers. The software also had support for multiple layouts, such as hierarchical, tree and circular. A layout controls the placement order and spacing of the graphical elements and is especially important for larger and denser networks.

2.6.4 Circos

Circos, introduced in 2004 by Martin Krzywinski, has advanced the circular layout form to be the aesthetic view of choice for most genetic and protein interaction views, particularly involving chromosomal structures. Because all genes and proteins can be resolve to a position, then an outer circular ring, the perimeter, segments can represent different chromosomes and other rings can represent attributes and annotations of those chromosomes. Genetic relationships are depicted as inner curved edges between two outer ring positions. Other advantages of circular layouts include more optimal spatial usage and particularly for cancer networks, intuitive patterns on trans chromosomal structural patterns and somatic aberrations. A grand highlight of circos was when circos graph was

featured in a Nature article concerning challenges of cancer [Ledford 2010] and since then, circos graphs have appeared in multiple prominent journals.

2.6.5 Limitations

Cytoscape and circos required installations and contain multiple dependencies. For Publication I, we seek to integrate and implement a web based software that is installation free and allows plotting genetic position networks for multiple model organisms, including human, mouse, fly, yeast and plant based organisms. Web based filtering is supported, and multiple annotation rings can be added using simple text files. In the Discussions, we present a scenario involving a metagenomic project using POMO to integrate different E. coli strain community tuning and usage of energy resources. POMO is open sourced software and can be used on modern browsers without registration.

POMO is a self-contained and lightweight web program following LAMPS (Linux Apache PHP/Python SQLite) architectural design, depicted in Figure 4.

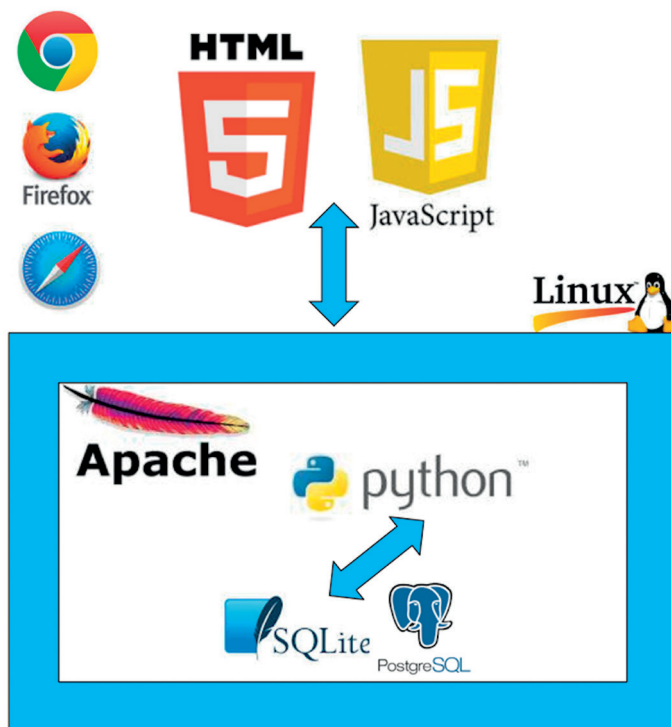


Figure 4. Components for open source web programming – Linux based tools, commonly known as LAMPS architecture, are the main components and libraries for POMO and Vipie, presented in Publication I and III.

The production and increased importance of biological networks also motivates the need to design frameworks to properly and consistently visualize the interactions and structures, particularly for cross chromosomal events. Meaningful visualization within large collaborations is essential for big data analysis and aids with effective communication and collaboration.

3 AIMS OF THE STUDY

The central aim of this study focuses on analysis of metagenomics data, particularly virome profiling and analysis of gut bacterial and viral interactions within T1D. In addition, the study presents the development of a web tool to plot genomics networks for multiple organisms, including bacterial and virus as custom defined references. In detail the aims are:

1. To perform data analysis on the role of gut bacteria and bacteriophages in T1D autoimmunity.
2. To develop and automate a web based pipeline, derived from aim 1, for profiling virus populations in metagenomics samples.
3. To develop an easy to use tool for plotting of genomic networks for multiple model organisms
4. To evaluate and apply these resources within metagenomic studies and T1D cohorts.

4 BIOINFORMATICS ARCHITECTURE AND METAGENOMICS MATERIALS AND DESIGN

In this chapter I will introduce virome samples, example networks, open sourced tools and reference databases used in Publication I and III. I will go over some design features and benefits particularly within open sourced paradigm and web programming. In addition, the motivation and design of Finland based Diabetes Prediction and Prevention (DIPP) study's pilot cohort focusing on gut microbiome profile association with early age T1D autoimmunity addressed in Publication II will be provided.

4.1 Open sourced and web programming

Open sourced programming and practices have the primary advantage of code visibility and hence ideal environment for promoting collaborations. As the paradigm allows for free non-profit usage and modification, it is ideal for meeting biomedical academic research goal of reproducibility. All code and statistical scripts within this thesis, including POMO web plotting and Vipie pipeline are open sourced. User guides and example samples and networks, discussed further in samples section, are available in sourceforge repositories. Moreover, all key dependent technologies (when relevant, referenced in chapter 2), are listed in Table 3, and the core primary languages JavaScript, Python and R are also open sourced. The architectures and web flows for POMO are captured in Figure 5A and Vipie, Figure 5B.

Table 3. POMO and Vipie, built on open sourced technologies and dependencies, are web based applications available for non-profit research.

Application	POMO ¹	Vipie ²
Web	HTML5, Javascript	HTML5, Javascript
Language	Python2.7	Python2.7
Clustering	None	R
Database	SQLITE 3	Postgres 9
Webserver	Apache 2	Apache 2
JavaScript libraries	jQuery, Cytoscape, D3, QuickVis	jQuery, Highcharts, DataTables
Server side	None	FastQC, Velvet, BLAST, BWA
Denovo assembly	None	Velvet, MetaVelvet, ABySS, SOAPDenovo, IAA
Supported Browsers	All, IE 10+	All, IE 10+

¹ <https://sourceforge.net/projects/finnpomo>

² <https://sourceforge.net/projects/vipie>

4.1.1 Interface design and workflow

Web browser based programs, compared to local desktop programs, have the advantages of installation free, easy version updates and consistent browser interfaces. Some negative features are dependency on network speed and security breach concerns. As the programs are designed for university and academic research centers connected to high speed internet, network connection speed is less of a concern. HTTPS protocol, designed to encrypt data exchange, is enforced for security. Vipie also requires username and password registration for security, status communication and process management reasons. Input files, such as protein interactions and annotations, are uploaded using HTML5 dialogs. Network plotting on the browser is very fast; thousands of edges can be plotted in seconds and filtering is supported without storing uploaded files. There is an upper limit of 20 gigabytes archived file size per job to lessen chance of server overload. Archived files are compressed and therefore this upper limit meets most metagenomic Illumina BaseSpace project downloads. On large jobs that cannot be split or reduce, we recommend trying one of the offline tools listed in Table 2, including Vipie local instance. Studies involving large number of samples are more suitable for high performance and parallel computing (HPC) architectures. Further described in discussions chapter, HPC Vipie instance has been used to process thousands of TEDDY stool virome samples integrating Linux native HPC paradigm.

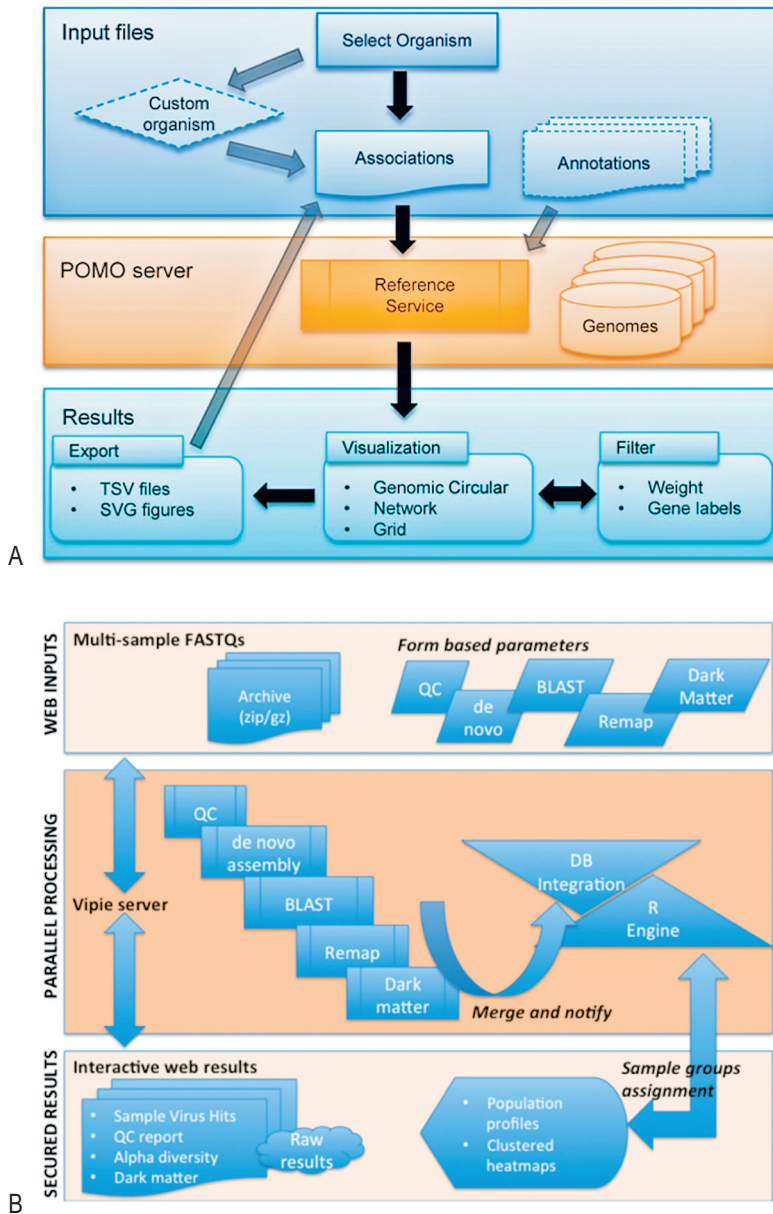


Figure 5. POMO and Vipie architectures – (A) POMO graphs are plotted from omic interactions and annotations files, uploaded from modern browsers. (B) Multiple metagenomics virome samples can be archived and uploaded in Vipie. Virus profile results are presented as clustered heat maps and interactive table.

4.2 Reference Databases

Within all disciplines of biomedical research and particularly systems biology, the importance of model organisms as reference databases and comparative resources cannot be overstated. It was only in 1997 that Baker's Yeast became the first complete eukaryote genome release [Goffeau et al. 1996] and it showed essential metabolism related gene homologs and also surprisingly human oncogenes [Botstein et al. 1997] in the yeast genome. Obviously besides human, for comparative genomics and biomedicine, annotation reference genomes of higher organisms such as worm, fly and mouse are required. POMO, listed in Table 4, has integrated those organisms and in addition to human reference, also plant based model organisms and bacteria *E. coli*. Rat and assorted *E. coli* strains have been added after publication due to user demand. POMO also has a feature to allow users define custom genomes and plot any nodes and edges based on positional index features.

Table 4. Model organism annotations – POMO supports multiple organisms and can translate their gene labels into chromosome coordinates via integration of their genome reference annotation sources.

Organism	Build	Source	URL
Human	GRCh37.p11	ENSEMBL	www.ensembl.org/Homo_sapiern/Info/Index
Fly	BDGP5	Fly base	Flybase.org
Mouse	GRCm38.p1	MGI	www.informatics.jax.org
Rat	RGSC3.4	Rat Genome Sequencing Consortium	http://rgd.mcw.edu/sequences/rgp_info.shtml
Worm	WBcel235	Worm Base	Wormbase.org
Yeast	EF4	SGD	www.yeastgenome.org
Zebra fish	Zv9	ZFIN	www.zfin.org
Arabidopsis	TAIR10	TAIR	www.arabidopsis.org
Rice	MSU6	MSU	Rice.plantbiology.msu.edu
Tomato	SL2.40	SolGenomics	Solgenomics.net
<i>E. coli</i>	MG1655	Ecocyc	Ecocyc.org

4.2.1 Viral databases

Vipie has integrated a local BLAST [Altschul et al. 1990] instance towards NCBI refseq database, reduced using CD-HIT [Li & Godzik 2006] consisting of all viral references. In addition, full viral genome sequences from Genbank are also stored and indexed. A PostgreSQL instance manages pairwise similarities between all relevant accessions. During population remapping, the pairwise scores are applied in scenarios where fragment reads are mapped to multiple genomes, thereby the reads are proportionally assigned to the highly similar strain accessions. Currently, 3 421 viral accessions have been processed and stored.

In cases of novel viral accessions, Vipie contains an automatic function to retrieve from NCBI web service the full genome sequence. Following the download, similarity scores between other genomes and indexes are computed.

Vipie workflow interface includes sample based web reporting and optional incremental flagging of potential human genome and non-viral microbial ribosomal mapped reads prior to viral population remapping. The database server has integrated the complete human reference [The Genome Sequencing Consortium 2001] update 19 and NCBI bacterial ribosomal 16S and fungal via 23S releases (<ftp.ncbi.nlm.nih.gov/genomes/TARGET>). Furthermore, using the high quality 5SRNAdb reference [Szymanski et al. 2016], the workflow also checks for ribosomal RNA and associated gene resources containing more than 7,000 organisms covering three separate taxonomic domains: Archaea, Bacteria and Eukaryota.

4.3 Diabetes Prediction and Prevention and gut microbiome association to autoimmunity

The Finnish based study Type 1 Diabetes Prediction and Prevention (DIPP) is committed to assess and innovate methods to predict, delay and ultimately prevent the disease with goal of advancing understanding of the mechanisms leading to T1D autoimmunity and pathogenesis. DIPP recently celebrated its 20th anniversary of discovery and innovation [Haller & Schatz 2016], the program was launched in 1994 and is one of the first from-birth prospective, large scale T1D focus projects. Until now, 200,000 infants have been screened for high risk genetic alleles and close to 17,000 candidates have been invited and enrolled in the study. Once enrolled, the children are encouraged to participate in 3–12 month interval appointments until they are age 15. Coincidentally and as reported in multiple countries [Gale 2002, Gardner et al. 1997, Karvonen et al. 2000], diabetes rate and onset have shifted towards younger children. As most new T1D cases fall on children without first degree relatives, DIPP's application of HLA risk screening seems appropriate and comprehensive.

During scheduled visits, blood samples are drawn to check for autoimmunity with the comprehensive islet autoantibody markers ICA, IAA, GADA and IA-2A (ZNT8 is also tested if child is positive for any other markers). Inflicted children and their parents are offered counselling, guidance and opportunity to take part in intervention trials aiming at preventing or delaying the development of T1D. Medical history such as infections, vaccinations, allergies and other background household data, including pets and diet were also collected along with checking for glucose tolerance. Stool samples were collected monthly until the child was 3 years old. DIPP investigators have a history of launching sub-studies to focus on potential important environmental factors.

4.3.1 Subjects

Publication II is based on a Metagenomic pilot project to assess gut microbes in stool samples from selected DIPP children with very-early islet autoimmunity confirmation and subsequent clinical T1D onset. The median age for the onset of autoimmunity was 17.4 months, and all case children confirmed for T1D at median age of 3.2 years (interquartile range 1.4–4.4 years). Healthy children matched for location, HLA, time of birth and age of sampling were selected as controls. In European countries, researchers have reported a steep and troubling increase in diabetes in the very young where 0–4 years old kids showing the highest relative increase, 5.4% and almost double the increase in ages 10–14, 2.9% [Patterson et al. 2009]. We reason that the immune systems of very young children might be more susceptible to virus infections and bacterial gut imbalances. In addition, subject infection and healthy histories were not considered.

4.3.2 Samples

For all DIPP study subjects described in Publication II, three stool samples were analyzed. These samples were collected in 3 month intervals, approximately at 3, 6 and 9 months time points prior to the onset of islet autoimmunity. Stool sample ages, all prior to autoimmunity, ranged from 2.9 months to 17.3 months. Samples were collected at home and mailed overnight at -70 Celsius. A few case-control pairs only contributed one or two samples and all together, a total of 92 samples covering 18 case-control subject pairs were sequenced with Illumina MiSeq.

Listed in detail with exact protocol steps, primers, size selections and included in supplement material [Kramna et al. 2015], Illumina sample preparation and amplifications were performed on stool supernatant, enriched for virus particles using filtering and ultracentrifugation to optimized viral signals. Sample preparations included reverse transcription to allow identification of DNA and RNA single and double stranded viruses. Amplification steps were adapted from prior work [Blinkova et al. 2010].

4.3.3 Data availability

The 16S amplicon bacteriome reads at 250 base pairs are available as compressed NCBI sequence read archives with identifier PRJNA311147. The virome sequence files are available as read archive PRJNA275568 in NCBI. All human virus findings were validated further with PCR.

4.3.4 Hypothesis and motivations

The primary hypothesis aims to explore associations between gut bacteriome profiles with T1D autoimmunity over time. We seek to investigate difference in within sample by assessing visually and quantitatively model alpha diversity and community inter-sample beta diversity, dissimilarity as measure in UniFrac [Lozupone et al. 2007] and Bray-Curtis [Bray & Curtis 1957]. Alpha diversity measures applied include Shannon [Shannon 1948], Chao1 [2003] and Simpson [1949] diversities.

While previous gut microbiome autoimmunity cohort studies [Vatanen et al. 2016, Davis-Richardson et al. 2014] have published bacterial imbalances and selected bacterial strains as increased risks with significance, none of the studies has had perfect agreements with another likely due to project designs and premises. The closest common findings are that changes at phylum Bacteroides is associated with T1D autoimmunity though it is also possible that fluctuation in Bacteroides, as it is one of the dominant phyla in human gut, is a side effect of inflammation from autoimmunity development. Our investigation applies multiple statistical models toward differential abundance and diversity measures between time points and case and control samples. The bacterial and viral phage correlation networks represent a novel attempt at finding potential host of crAssphage and association towards T1D autoimmunity.

4.3.5 Rarefaction and subsampling

It is common and understandable that there are large differences in the number of sequence reads across samples. Subsampling, also called rarefaction, is a much discussed statistical procedure [McMurdie & Holmes 2014]. The support for subsampling is based on that samples with larger read sizes will have greater chance of finding rare microbes compared to samples with lower yields. To have comparable proportions, investigators used subsampling and it typically demands finding the read size from the smallest sample removal and then setting all other samples to this read size by randomly selecting reads from relevant samples. This procedure while addressing read size biases, likely also removes relevant information and introduces biases into diversity metrics and population profiles.

Reduction of valuable sample yields is far from optimal as potential important and real but lowly represented microbial organisms are likely to be removed during subsampling processing. We did not perform subsampling and instead applied proportional differential abundance using DESeq2 [Love et al. 2014] and taking advantage of *phyloseq-to-deseq* tool within the phyloseq package [McMurdie & Holmes 2013]. These packages are available in R and Bioconductor [Huber et al. 2015]. The authors of phyloseq package also have included access functions to data model attributes and plotting functions for abundance and multiple diversity measures with aggregation functions covering all taxonomic levels.

4.3.5.1 OTU construction and model application

Following best practice and standard 16S protocol processing, sample paired reads were merged and screened using Mothur [Schloss et al. 2009] software. The contig results then were further processed using Qiime [Caporaso et al. 2010]. Qiime scripts were applied as parameter customizations was easier to set. Operational taxonomic units (OTU), essentially groups of clustered sequences at 97% similarity were computed and the representative sequence was selected using the most abundant option. OTU typical standard similar thresholds are between 90% to 100% where 100% represent grouping on perfect matches. Chimeric reads defined to be non-prokaryotic were removed via *usearch* [Edgar 2010]. Bacterial taxonomy was assigned using SILVA, version 108 [Quast et al. 2013] database. To account for possible sequencing error, we removed OTUs with less than 5 reads or found in only one sample, out of 92 total prior to abundance profiling and case control testing. Dependent on distribution pattern, multiple models in R were applied and false discovery testing was corrected using Benjamini-Hochburg [Benjamini-Hochburg 1995]. For diversity significance between case and controls, Student's t.test was used. In testing OTU counts between case and control with expected non-normal distribution, Wilcoxon signed-rank test [Wilcoxon 1945]. Differential expressed abundance testing was applied using negative binomial generalized linear model provided in DESeq2 [Love et al. 2014]. Moreover, correlation analysis with Spearman rank [Lehman 2005] was performed between OTUs and bacteriophages across all time points. In addition, investigation of possible time lagged effect, approximately 3 months, of viral phages on bacterial was conducted. Networks were constructed on pairs with threshold higher than 0.3 absolute correlation and p-value less than 0.001.

4.4 Simulation data and validation samples

4.4.1 POMO networks

POMO input genomic networks are algorithm and technology independent. Nodes must map to chromosome positions or for regulatory components, the position defined within the node label. Distal somatic aberrations, including gene fusions and other inter-chromosomal events are all supported. Genomic circular view layouts are ideal for depicting structural aberrations and rearrangements. Within the results, we included scenario of visualizing and filtering somatic structural rearrangements founded in deep sequencing data using BreakDancer [Chen et al. 2009] from The Cancer Genome Atlas (TCGA) glioblastoma (GBM) study [Zheng et al. 2013]. Chromothripsis events, defined to be large localized rearrangements, associated with poor survival are defined in an annotation file that can be uploaded. We also visualized gene expression correlation network results calculated between human embryonic (hESC) and induced pluripotent stem cell (hiPSC) [Närvä et al. 2010]. The high quality yeast protein-protein interaction [Schwikowski et al.

2000] network was one of the first genome wide comprehensive protein studies and also part of the initial Cytoscape publication, can be found in the POMO archive. Example network and annotation files to define custom model organisms and comparative networks are accessible at <https://sourceforge.net/projects/finnpomo/files/source-archive.zip>.

4.4.2 Vipie microbiome samples

To assess Vipie performance, workflow and interface, we used 11 samples from previous published HMP, and Japanese virus extraction projects [Yamashita et al. 2016] and also three unpublished samples from an African viral diversity project [Rodríguez-Díaz et al. 2014, Mangani et al. 2014]. The sample type, source and sequence details are provided in Table 5. The published samples are included to serve as comparisons between Vipie results and known published works. Potential artefact reads were removed against viral blacklist accessions. The blacklist accessions are NCBI submissions containing vector, chimeric and synthetic constructs in their genus descriptions and further discussed in 4.4.4. Reads mapped to bacterial and human references are reported in Figure 11 and further discussed in discussions chapter. The statistical comparisons are further discussed in the results chapter. A compressed file containing all described 11 samples are available here: https://binf.uta.fi/vipie/data/vipie_archive_ssampld.zip.

Table 5. Vipie pipeline results and performance were tested with the following virome samples [Table 3 Publication III].

Accession	Source	Sample Type	Number of Reads
SRS072276	HMP	Blood	438,879
SRS072318	HMP	Blood	753,994
SRS014466	HMP	Vagina	367,077
SRS015072	HMP	Vagina	495,256
SRS072313	HMP	Nasal	320,672
SRS072261	HMP	Nasal	367,384
SRS072366	HMP	Nasal	114,414
S11	Africa	Stool	1,634,821
S12	Africa	Stool	1,191,427
S14	Africa	Stool	1,143,784
DRA004165	Japan	Diarrheal stool	1,108,688

4.4.3 Simulation data

In order to compute exact precision and sensitivity, simulation metagenomic NGS reads were generated from ART sequence simulator [Huang et al. 2012] and systematically assessed. The composite had a total of approximately 20 million reads, of which 94% (19,582,500) were human, 4.8% bacterial (986,114) and 0.7% (146,886) viral. Vipic results compared favorably with among the best known pipelines, including MetaShot [Fosso et al. 2017] where this composite archive originated. Both precision and sensitivity scored above 96% and details are listed in Table 7 in Results chapter.

4.4.4 Blacklisted chimeric vectors

Vipic currently has built in filtering for a list of chimeric viral accessions that have been reported as problematic and potential Illumina sources [Mukherjee et al. 2015]. The filtering can be optionally turned. Potential chimeric reads are an ongoing concern for all sequencing analysis as they are specific to projects, sequencers and sample of origin. For instances fresh water [Fernandez-Cassi et al. 2017] and human stool have different baselines and the geographical location of sequencing might also need to be considered. Unexpected and consistent artefact [Wooley a& nd Ye 2009] viral reads found in relevant control sequence samples are potential signs for sequencer contamination or leak through from previous runs.

Table 6. Vector and synthetic accessions are optionally blacklisted to improve accuracy.

Accession	Description
AF324493.2	HIV-1 vector pNL4-3, complete sequence.
AY656167.1	Chimeric dengue virus vector p4-D3L-ME, complete sequence.
AY656169.1	Dengue virus type 3 vector p3, complete sequence.
AY705791.1	Borna disease virus rescue plasmid pBRT7-HrBDVc, complete sequence.
AY744148.1	Dengue virus type 2 vector p2, complete sequence.
FJ436096.1	Synthetic construct Gallid herpesvirus 2 clone pC12/130-10, complete sequence.
FJ436097.1	Synthetic construct Gallid herpesvirus 2 clone pC12/130-15, complete sequence.
FJ593289.1	Human herpesvirus 1 transgenic strain 17, complete genome.
GU179001.1	Human herpesvirus 5 transgenic strain Merlin, complete genome.
GU474419.1	Synthetic construct modified HIV-1 subtype C backbone, complete sequence.
GU980198.1	Human herpesvirus 5 transgenic strain CINCY+Towne, complete genome.
HQ687214.1	Virus-induced gene silencing vector pCAPE2-PsPDS, complete sequence.
KF022001.1	Autographa californica nucleopolyhedrovirus transgenic, complete sequence.
KF493877.1	Human herpesvirus 5 transgenic isolate Towne-BAC-der, complete genome.
KJ540270.1	Vibrio phage CTX plasmid pCTX-3 Kan, complete sequence.
KP343683.1	Cyprinid herpesvirus 3 isolate FL BAC revertant ORF136 Luc, complete genome.
KR093640.1	Moraxella phage Mcat16, complete genome.
AY376438.1	Dengue virus vector p4(Delta30)
KX576684.1	Zika virus vector pZIKV-ICD, complete sequence.

4.5 Statistical learning

Statistical learning, relevant in all data sciences, is the backbone of systems biology and here essential and relevant vocabulary with high level concepts are introduced [James et al. 2013]. Statistical learning methods can be supervised or unsupervised. Classification and linear regression are supervised algorithms as they involved an output label, or outcome in relationship to the input whereas unsupervised methods do not require output labels and focuses on input variable patterns or distribution, such as clustering. Regression learning implies understanding and investigation of the relationships between variables within a data set. Input variables are predictors, also called independent and the outcome (output), can also be called response or dependent. Modelling essentially attempts to answer, or infer how the outcome of Y is affected as input variables $\{X_1, \dots, X_p\}$ changes. In the case that Y is numeric, the model is regression and when Y is categorical, or qualitative, is often considered a classification problem. Binary outcome models are handled by logistic regression. Certainly one of the key goals of any regression design is finding the effect of outcome, or response dependent based on individual inputs, or predictor variables. A

p-value less than 0.05 is usually considered significant. However, it is important to note that these main relationships, as they are called, can be a result of interaction between different input predictor variables and these non-independent effects need to be reported. Below I will introduce methods that have been applied to this thesis, including random forest and linear regression models.

4.5.1 Random forest

Publication I had its roots in a project to manage and plot random forest, a learning method for non-linear data and classifying genome data is ensemble based random forest (RF) [Ho 1995, Breiman 2001]. Ensemble implies improving on final classification and regression performance is based on assembly of many weaker models. Random forest builds large number of decision trees using a subset of the data and on each decision node, a random subset of the variables are selected and split based on the variable contribution to either entropy or impurity reduction, or GINI index [Tuv et al. 2009]. In essence all trees are walked to their ends while splitting on variables that maximizes homogeneity, or the shortest path to the leaf. To be exact, purity score is 0 when the relevant values are homogeneous. Random forest has been shown to be accurate and good performant, particularly if cross variable correlations are minimized [Touw et al. 2013]. RF also overcomes the over-fitting limitation of single decision trees by random selection of variables on each node and also using bagging, where out of bag samples are used for prediction assessment and validation. Aggregating all variable GINI index scores over all trees and splits, variable importance can also be reported. Interpretation of RF model results can be challenging [Chen & Ishrawan 2012] due to multiple randomizing steps and recommended sample bagging. While RF is assumed to be a black box, individual trees are accessible and comprehensive analysis of decision paths for the best ranked variable paths can offer insights particularly in conjunction with interactive plotting.

4.5.2 Regression and classification

In Publication II, generalized linear regression model was applied on bacterial abundances between case and controls. The model, implemented in DESeq2 (Love et al. 2014) used negative binomial distribution and is equipped and recommended to account for excessive variation, typical of bacteria abundances between samples. Publication II, the association with virus exposures towards autoimmunity is tested with conditional logistics regression model included in Bioconductor [Huber et al. 2015]. Bacterial-phage networks were scored using both Pearson bivariate correlation [Pearson 1895] and Spearman correlation suitable for ordinal counts or ranks [Lehman 2005].

4.5.3 Clustering

The primary goal of unsupervised learning is finding hidden patterns within data without the aid of labels or known classes, such as case and control identifiers. Clustering and association rule learning are classes of unsupervised learning and widely applied in machine learning and data mining. Clustering methods in essence finds groups based on similarities. As similarities are defined in the context of the data, clustering also have many different implementations and strategies. Hierarchical clustering is employed as part of Publication III, Vipie web pipeline results, heatmap plotting function where sample and their accessions are plotted as heatmaps and placed in nested groups based on virome profile similarities.

4.5.4 Sensitivity and precision

As part of Publication III Vipie performance benchmarking and validation, we processed simulated mock data consisting of sequences from human, bacterial and viral. Sensitivity and precision was computed between Vipie calls with correct read origins.

Sensitivity or recall or True positive rate (TPR, 2.1) computing proportion of True positives (TP) accounting for False negatives (FN):

$$(2.1) \quad TPR = TP \div P = TP \div (TP + FN)$$

Precision or Positive predictive value (PPV, 2.2) is computed proportion of True positives accounting for False positives (FP):

$$(2.2) \quad PPV = TP \div (TP + FP)$$

F-measure (F, 2.3), a harmonic mean of precision and recall is defined:

$$(2.3) \quad F = 2 \times ((PPV \times TPR) \div (PPV + TPR))$$

4.5.5 Shannon diversity

For Publication III, viral alpha or intra-sample diversity is measured as Shannon [Shannon 1948] entropy (H, formula 2.4) where the proportion of bacterial individuals belonging to the serial accession within the dataset of interest. Shannon entropy quantifies the uncertainty in predicting the individual species identified taken at random relative to relevant species population within dataset.

$$(2.4) \quad H = - \sum_{i=1}^R (p_i \ln (p_i))$$

Shannon index originates from amount of entropy, uncertain information, within a given string of characters. Within Vipie, this score quantifies the amount of uncertainty in viral taxonomy accession taken at random. Other measures, such as indexes such as Fisher, Simpson and Chao [Chao 2003, Fisher et al. 1943, Simpson 1949] can be calculated using the virome profile matrix available as part of downloads.

5 RESULTS

In this chapter, I describe the published tools and results presented in this thesis. First, I describe the plotting usability of POMO and an application involving comparison of microbial community energy usage (Publication I). Second, the main findings of Publication II study focusing on bacteriome association and correlation networks with bacteriophages towards early age T1D autoimmunity are discussed. Third, Vipie multi-sample virome pipeline interface, features and accompany results are highlighted (Publication III).

5.1 Plotting omics networks for multiple organisms (Publication I)

POMO serves as a secured web based tool for genome-wide network visual exploration and promotes collaboration since the filtered results can be viewed in multiple layouts and shared as images or text inputs for future session inputs. The original motivation of Publication I was driven by need to visualize The Cancer Genome Atlas (TCGA) genetic and structural relationships, derived from omics data and produced by machine learning random forest [Breiman 2001] and comprehensive paired-paired feature correlations. In an effort to extend the TCGA configuration to mouse stem cells and also yeast networks produced by Hidden Markov Model [Baum & Petrie 1966], we realized that a generalized tool, fully web based that can handle plotting networks independent of learning methods and supporting different model organisms would be beneficial for investigators producing large scaled genomic interactions and structural aberration.

To add more usability and value to the labs, the graphs can be exported as scaled vector graphs (SVG), a high quality image format while offering built in graph examples, in simple text formats such as tab separated or simple interaction format (SIF) extensions for different model organisms.

5.1.1 Nodes, edges and styling

The essence and simplicity of a graph is that mathematically it is defined and consists only of nodes and edges. POMO supports this definition by allowing users to upload a text file where each line is an edge representing two interacting nodes, with an optional score or strength or color describing this interaction. Within the same line, the nodes can be described with types, such as gene expression or methylation sources. Using this definition,

and plotting reference chromosome regions as outer perimeter arcs, the nodes resolve or define to a position and inner arcs are drawn connecting the genetic positions. Additional chromosomal annotations, including copy number events, can be plotted as outer rings via uploaded text files. Edges can be further described with colors, strength of interaction using quantitative measures, such as rank or correlation values and those can be filtered from the browser. We further integrated Cytoscape web to allow for additional layouts such as tree and force directed.

5.1.2 Model organisms

Listed in Table 4 of the Materials chapter, Publication I has integrated the full genome references of human, mouse, rat, worm, yeast, zebra fish, Arabidopsis, rice, tomato and E. coli. At the time of the project and we believe as of today, it is the most comprehensive collection of its kind while also supporting custom organisms. POMO users can temporally define custom versions of model organisms and then plot the interactions using coordinate based node labels.

5.1.3 Brain cancer genomic rearrangements

Chromosomal wide context layouts are intuitive lenses into global genomic aberrations such as gene fusion and structural instabilities. Some of the key findings from TCGA glioblastoma (GBM) study were poor survival with certain chromothripsis and genomic instability events [Zheng et al. 2013] and selected results are plotted in Figure 6 below. The number of supporting reads per chromothripsis event is represented by edge colors, grey (< 50), red (50 to 100) and blue (> 100). The outer ring provides the cytoband and chromosomes while the optional inner rings represent relevant gene copy number gains and losses, the innermost ring provides estimated fusion event. Part B of the figure represents the same graph using Cytoscape view and reveals 4 sub graphs. In addition to installation free, POMO node label flexibility, allowing gene names and/or positional node labels and edge color within text files provides increased usability.

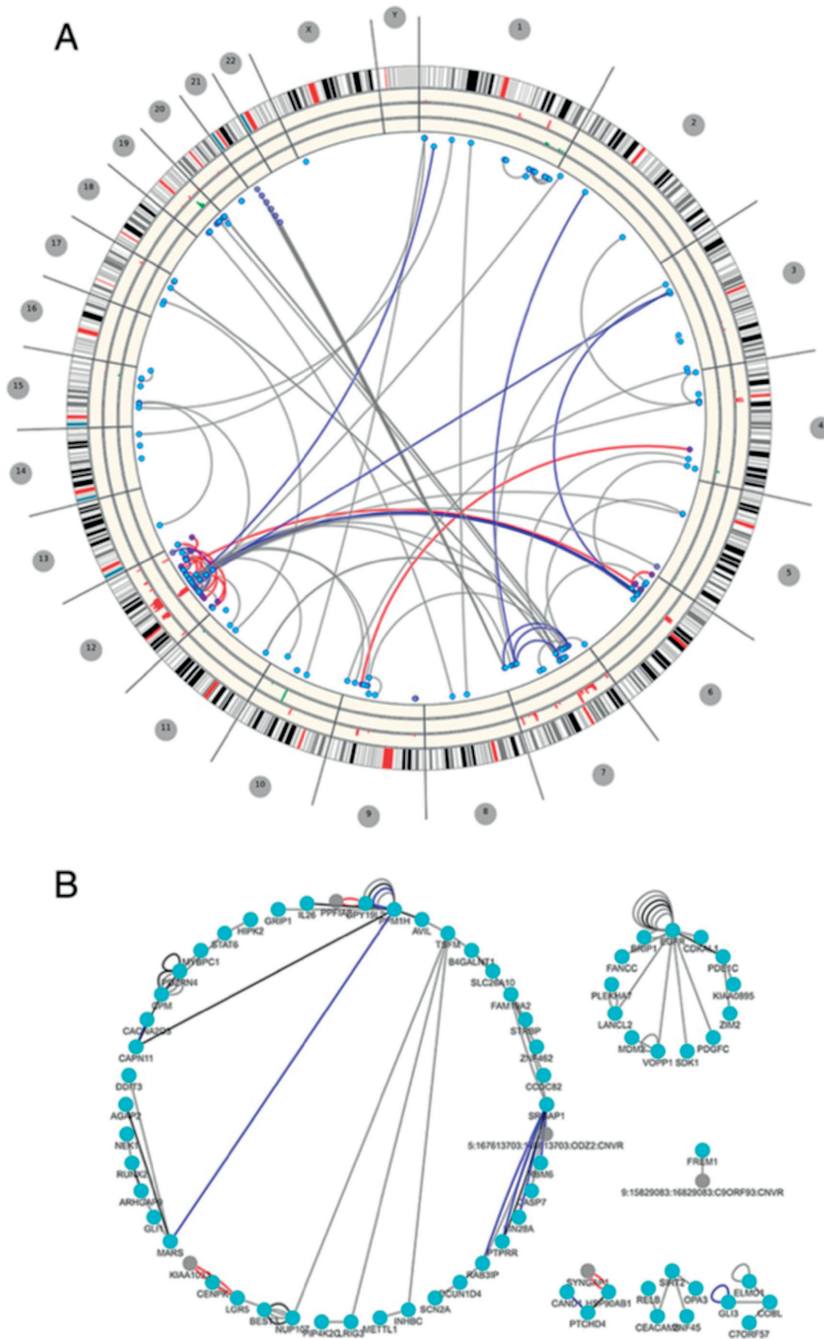


Figure 6. Plotting genomic structural variations. The figure depicts TCGA GBM rearrangements and chromothripsis events associated with poor survival from whole genome sequencing. A) Edges colors are used to describe supporting reads, with blue larger than 100, red 50 to 100 and grey less than 50. B) Cytoscape based network view showing the same data where the layout reveals 4 disjoint graphs. (Lin et al, 2013, Figure 3, BMC Genomics).

5.1.4 POMO Edge filtering and bundling

Genomic networks mined from omics data are often large and dense. We have implemented several filtering options including edge score comparisons and node label and type matching and the graph is refreshed with filtered results. Figure 2 in Publication 1 shows the original 45,791 graph of human embryonic stem cell (hESC) and human induced pluripotent stem cells (hiPSC) copy number correlations [Närvä et al. 2010, Hussein et al. 2011] and then a refreshed graph after selecting top 2000 based on absolute correlation, with pink edges depicting negative relationships.

Genome context views have a clear spatial limitation where nodes very close to each other will be blurred or not visible. To alleviate this limitation, POMO has a tabular view showing all edges and also edge bundling option, users can define central nodes, in effect hubs, where the proximal nodes within defined nucleotide distances are grouped. An example workflow is further described in POMO user guide, available on the website.

5.1.5 POMO example metagenomic application

POMO was applied in a study to investigate microbial community, from wastewater, usage of resources between generalist, physiologically versatile versus narrow niche specialist bacterial organisms [Muller et al. 2014] via integration of 16S NGS and metabolome data from LC-MS. Reproduced here as Figure 7 [Figure 3, Muller et al. 2014], mRNA expressions, protein and SNP rings of 10 reconstructed genomes, scaled using base lengths, were plotted using POMO and additional graphs were included in its supplement.

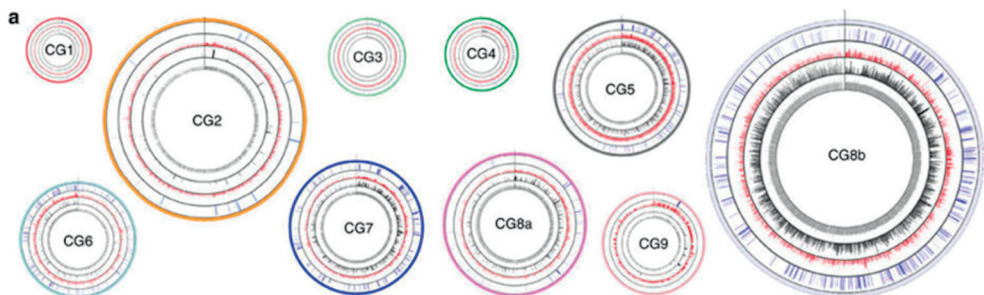


Figure 7. Applying POMO in metagenomics – Graphs reconstructed using microbial genome mRNA, protein expressions and SNPs detected data (Permission granted from Nature Communications).

5.2 Investigating microbiome associations in early islet autoimmunity (Publication II)

5.2.1 Viruses found in stool virome

Bacteriophages were found in 54% of the stool virome samples, crAssphage being the most abundant. Human viruses, including viruses with either DNA or RNA genomes, were found in 10.4% of the samples, and further validated using real RT-PCR [Kramna et al. 2015]. There were not any statistical significant associations with virome composition and autoimmunity.

Gut bacteriome

Using tightly matched case and control groups, bacterial alpha diversity in stool samples was assessed and found to be not statistically associated with autoimmunity. Beta diversity plots also did not reveal clusters or obvious patterns. Agreeing with literature, the most abundant phyla were Firmicutes, then Bacteroides, Proteobacteria and Actinobacteria.

At the OTU level, defined to be 97% sequence similarity over 250 bases, in effect allowing for maximum 7 base difference, four taxonomy units were found to be less abundant in cases compared to controls after adjusting for sample time of 3, 6 and 9 months prior to autoimmunity using DESeq2 generalized linear model with negative binomial distribution. The p-values were corrected for multiple testing via Benjamini-Hochburg [Benjamini and Hochburg 1995] correction and subsequent p-values less than 0.05 were considered significant. The mean abundances were relatively higher in control samples compared to cases, three of the OTUs mapped to phyla Bacteroides and one to Bifidobacterium, and at the species level *B. caccae*, *B. dorei* and *B. vulgatus*. While validating reported *Bacteroides dorei* dysbiosis autoimmunity [Davis-Richardson et al. 2015], we found that our strongest signal included a combination of *B. dorei* and *B. vulgatus* as the two species are just one nucleotide apart on the 16S V4 stretch of 253 bases. We repeated the analysis with 100% similarity, in essence no binning and using absolute sequence, and we confirmed the *B. vulgatus* inverse association towards islet autoimmunity, adjusted for sample age, but *B. dorei* was no longer significantly different between case and controls, across any of the sample ages. The *B. dorei* and *B. vulgatus* findings were further validated via PCR quantifications. OTUs mapped to *Bifidobacterium catenulatum* and *Bifidobacterium bifidum* were found to be significant though their mean abundances were 4 times less than *B. vulgatus* [Table 2 in Publication II].

5.2.2 Integrating virome with bacteriome

We constructed web based correlation (ρ) networks visualizing the abundance data of viral bacteriophages versus bacteria abundance. All quantities were normalized to 10,000 reads. The network was pruned so that edges were retained only if their absolute Spearman correlation was more than 0.3 and P values less than 0.001. After making the sample-wise correlation network we also made networks shifted by 3 months to investigate how bacterial abundance correlates with upcoming phage abundance and vice versa. The interactive networks are available at: [https:// http://compbio.uta.fi/phagenet/v2/index.html](https://http://compbio.uta.fi/phagenet/v2/index.html). A searchable tabular table is also provided for clarity and usability.

Figure 8 depicts case (top) and control (bottom) sample correlation networks of phage and bacteria. Spearman correlation was used to account for non-Gaussian distribution and spurious associations were pruned using thresholds significant p-value less than 0.001 and absolute correlation value greater than 0.3. The size of the circular nodes are scaled according to population magnitude where red and pink edges represent positive correlation, while blue for negative values. There are closed to 50% more edges in the case graph, 67 compared to the control network with 44. There are also double the number of viral phage nodes in the case network, represented as orange and dashed edges for added visibility. Shown in top figure, we found that CrAssphage was correlated with *Bacteroides dorei* ($\rho = 0.56$, p-value < 0.0001) but not with any other *Bacteroides* group members. Temporal based networks according to 9, 6, and 3 months before autoimmunity were constructed and we observed an increase in phage bacterial correlations getting closer to autoimmunity confirmation. Between case and control networks, we did not find any strong common edges though intersecting the 3 time point specific networks revealed strong, > 0.6 correlation. Interestingly, 9-month bacteria:6-month-phage networks showed the highest number of phage interactions, with seven while on the next iteration, 6-month-bacteria:3-month-phage the number was reduced by more than half to just three.

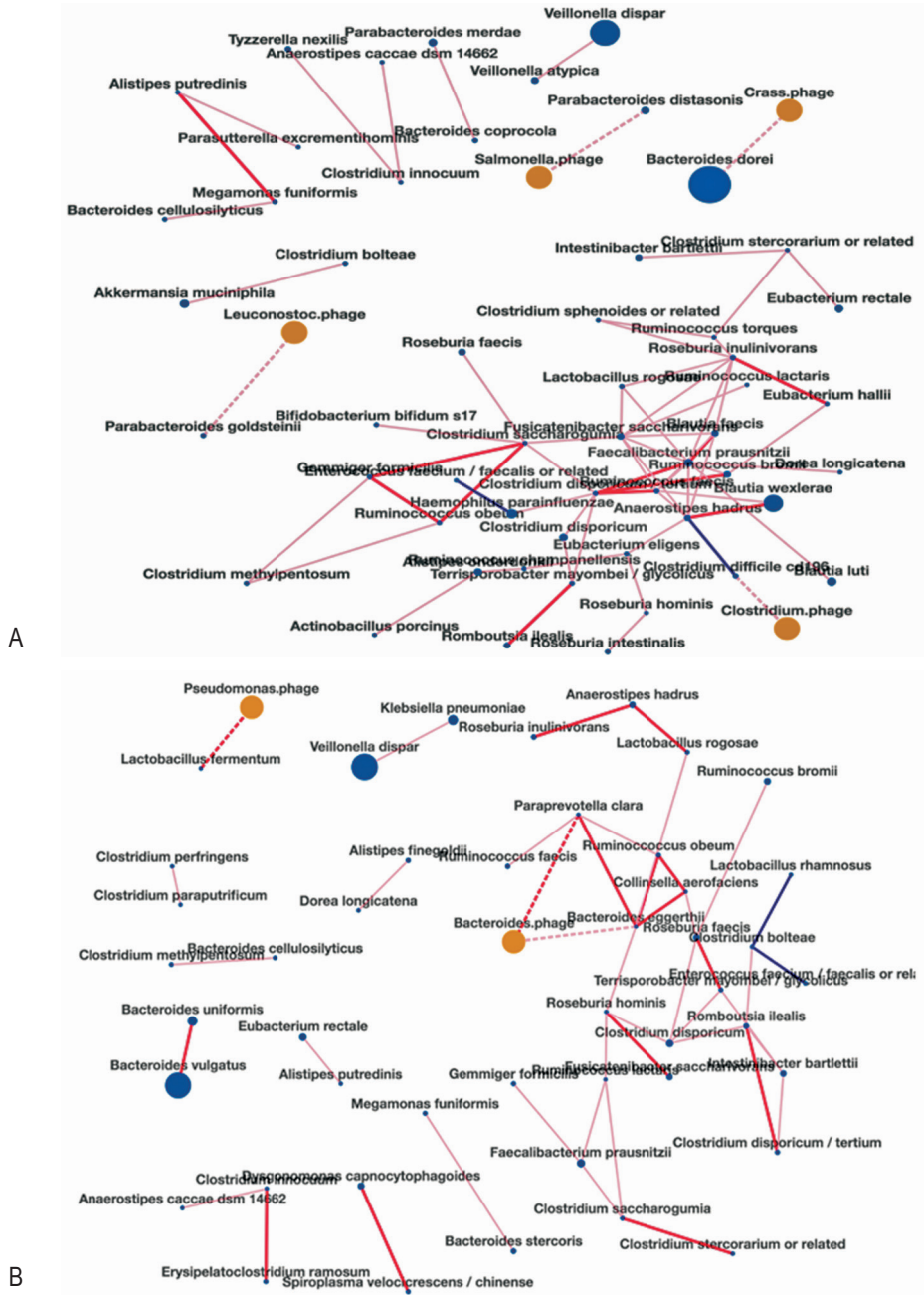


Figure 8. Correlated networks of bacteria and phages – Case graph (A, top) has 50% more edges and viral phage, in orange, nodes compared to control graph (B). Negative correlations are represented as blue, while red are strong positive correlations. These graphs are built using cytoscapejs and best accessed on web page <http://compbio.uta.fi/phagenet/v2> where correlation networks from 3, 6 and 9 months prior to autoimmunity are also available.

5.3 Vipie web based virus profiling pipeline for multiple samples (Publication III)

5.3.1 Mock data precision and recall results

Using approximately 20 million of simulated human and microbial reads and listed in Table 7, Vipie achieved good precision (96.85%), recall (95.36%) and F-measure (96.08%) results relative to MetaShot [Fosso et al. 2017]. The simulation data was listed in Materials chapter and built using ART sequence simulator [Huang et al. 2012]. While Vipie reports bacterial content based on 16S ribosomal matches, simulated data included bacterial reads from full genome and these contributed to Vipie's higher unclassified read percentage of 6.73%. Simulation inputs and scripts applied can be found on Vipie source code page. Not shown here but relative to MetaShot Table 1 [Fosso et al. 2017], Vipie viral precision and recall metrics were better than Kraken [Wood & Salzberg 2014] and MetaPhlan2 [Truong et al. 2015].

Table 7. (A) Comparison of read assignments between MetaShot and Vipie on simulated datasets consisting of 19,582,500 human (94.5%), 986,114 bacterial (4.8%) and 146,886 viral (0.7%) reads. Vipie percentages are based on random subsampling of one million reads and bacterial statistics are not reported and contributes to unclassified higher percentage. (B) Precision, Recall and F-measure are computed.

A				
	Assigned %^b		Correctly Assigned %^c	
	MetaShot	Vipie	MetaShot	Vipie
Human (host)	99.18	99.27	99.99	99.27
Viruses				
Family	97.74	99.98	98.53	93.39
Genus	97.39	98.99	99.75	93.33
Species	97.81	93.66	96.70	92.97

B				
	Human (host)		Virus	
	MetaShot	Vipie	MetaShot	Vipie
Precision (%)	100.00	100.00	98.30	96.85
Recall (%)	99.97	99.96	98.19	95.36
F-measure (%)	100.00	99.98	98.07	96.08
Unclassified (%)	1.04	0.73	3.94	6.73

a. <https://recascloud.ba.infn.it/index.php/s/nw4s9hqnF8QkBsK>

b. The percentage refers to the total number of reads assignable to the specific taxonomic rank.

c. The percentage refers to the relevant assigned reads.

5.3.2 Interface design, parameters and status communications

The architecture and main components of Vipie were drawn and described in previous chapter. Shown below in Figure 9, Vipie web interface design is aimed at ease-of-use and flexibility. The parameters are grouped into QC, de novo assembly, BLAST and Remap panels. Valid parameter values are enforced and described in User guide. Project name and virome fastq archive file upload are the only required inputs to start a job. Job statuses are sent to the registered email and on completion, a secured URL is sent.

The image shows a screenshot of the Vipie web interface with four panels of parameters:

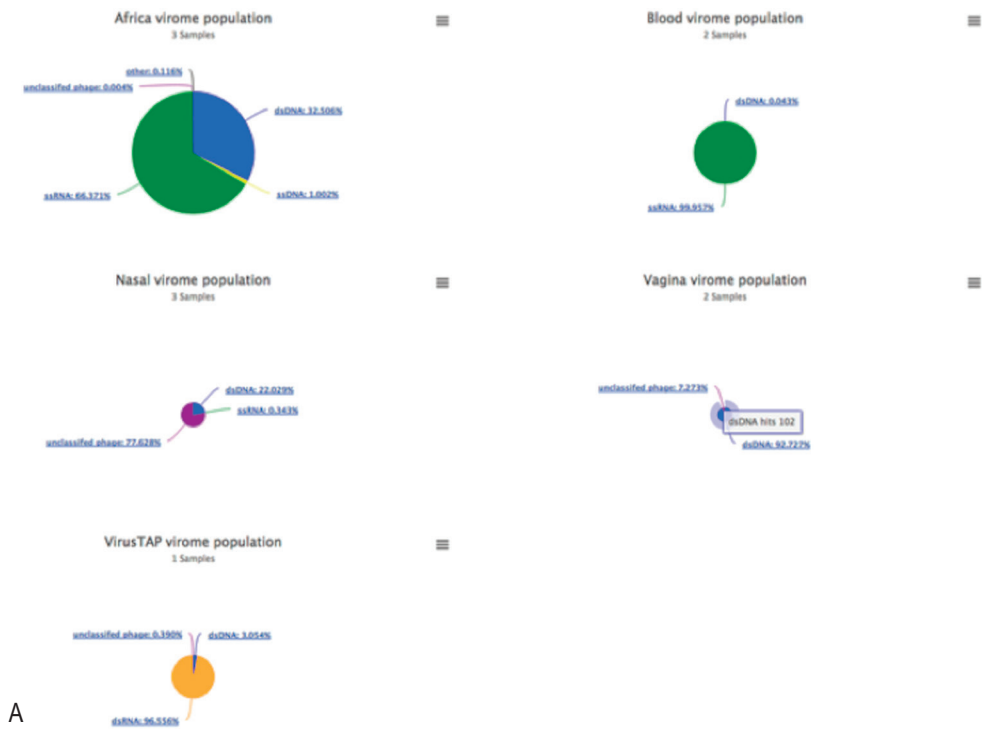
- QC parameters:**
 - Trim left/right: 10
 - Qual cutoff: 10
 - Insert/read length: 200
 - MAPQ/Phred: 10
 - Subsample: .75
- de novo assembly parameters (Optimize KMERGENE):**
 - de novo: Velvet (dropdown), amos-no (dropdown)
 - Kmer: 51
 - Expected coverage/cutoff: auto (dropdown), 20
 - Min contig length: auto
- BLAST parameters:**
 - Output format: tsv (dropdown)
 - Summary format: xlsx (dropdown)
 - Min percent similar: 80
 - E value: 0.0001
 - Number of alignments: 10
- REMAP/Reduction parameters:**
 - Minimum total matches: 5
 - Hits PER: 100000
 - Apply blacklist (vector/synthetic refs): yes (dropdown)
 - Remove human: yes (dropdown)
 - Remove ribosomes (bacteria/others): yes (dropdown)

Figure 9. Vipie web interface – along with NGS archived file, users can set pipeline parameters from the web where validation checks are automatic.

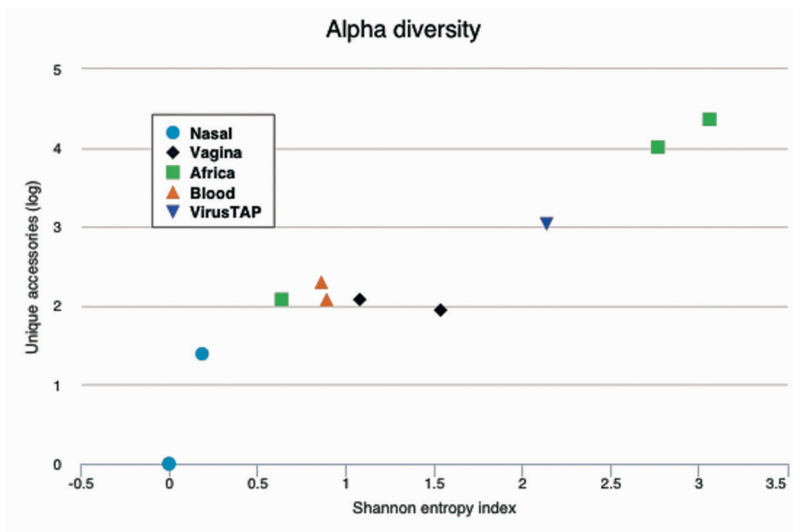
Result page is accessible from a secured URL containing interactive table, quality control and read reports, population pie charts and applicable diversity and clustered maps. Raw and intermediate output downloads are also supported. To allow for flexibility and promoting collaboration, the results page can be shared using the same encrypted URL since it does not require user authentication. Accompanying results from listed samples are captured in figures below and overall layout is divided into Population profiles, QC & Dark matter, Summary & Diversity and Viral hits table panels.

Group	Order	Family	SRS014466 Identified 0.07%	SRS072318 Identified 9.77%	S12 Identified 0.85%	SRS015072 Identified 0.04%	Diarrhea Identified 12.32%	SRS072313 Identified 0.60%	S14 Identified 31.30%	SRS072366 Identified 0.22%	SRS072261 Identified 0.14%	SRS072276 Identified 10.56%	S11 Identified 14.41%
			sample reads	sample reads	sample reads	sample reads	sample reads	sample reads	sample reads	sample reads	sample reads	sample reads	sample reads
ssDNA viruses	-	unclassified Anelloviridae	0	0	74	0	0	0	123	0	0	0	4
dsDNA viruses	Caudovirales	Siphoviridae	56	0	409	26	327	598	181	225	126	2	14106
Retro-transcribing viruses	-	Retroviridae	0	0	24	0	0	0	0	0	0	0	0
dsDNA viruses	Caudovirales	Podoviridae	0	0	59	0	0	0	66	0	0	0	0
ssRNA negative-strand viruses	Mononegavirales	Pneumoviridae	0	0	0	0	0	0	0	0	15	0	0
ssRNA positive-strand viruses	Picomavirales	Picomaviridae	0	0	71	0	0	0	30795	0	0	0	0
ssDNA viruses	-	Panoviridae	0	0	2	0	0	0	0	0	0	0	0
dsDNA viruses	Caudovirales	Myoviridae	9	0	0	8	0	0	20	0	0	0	0
ssDNA viruses	-	Microviridae	0	0	2	0	0	0	0	0	0	0	0
ssDNA viruses	-	Inoviridae	1	0	0	0	0	0	0	0	0	0	0
dsDNA viruses	Herpesvirales	Herpesviridae	0	0	0	0	0	0	0	0	0	6	0
ssRNA positive-strand viruses	Nidovirales	Coronaviridae	0	9771	0	0	0	0	0	0	0	10657	0
ssRNA negative-strand viruses	-	Bunyaviridae	0	0	23	0	0	0	0	0	0	0	14
ssDNA viruses	-	Anelloviridae	0	0	55	0	0	0	161	0	0	0	0
dsDNA viruses	Herpesvirales	Alloherpesviridae	0	0	0	0	0	0	0	0	0	0	254
artificial sequences	-	-	4	0	128	4	11994	0	18	0	0	0	36

Figure 10. Interactive table – the viral results are sortable and can be filtered based on text. Viral populations can be summed across taxonomy levels.

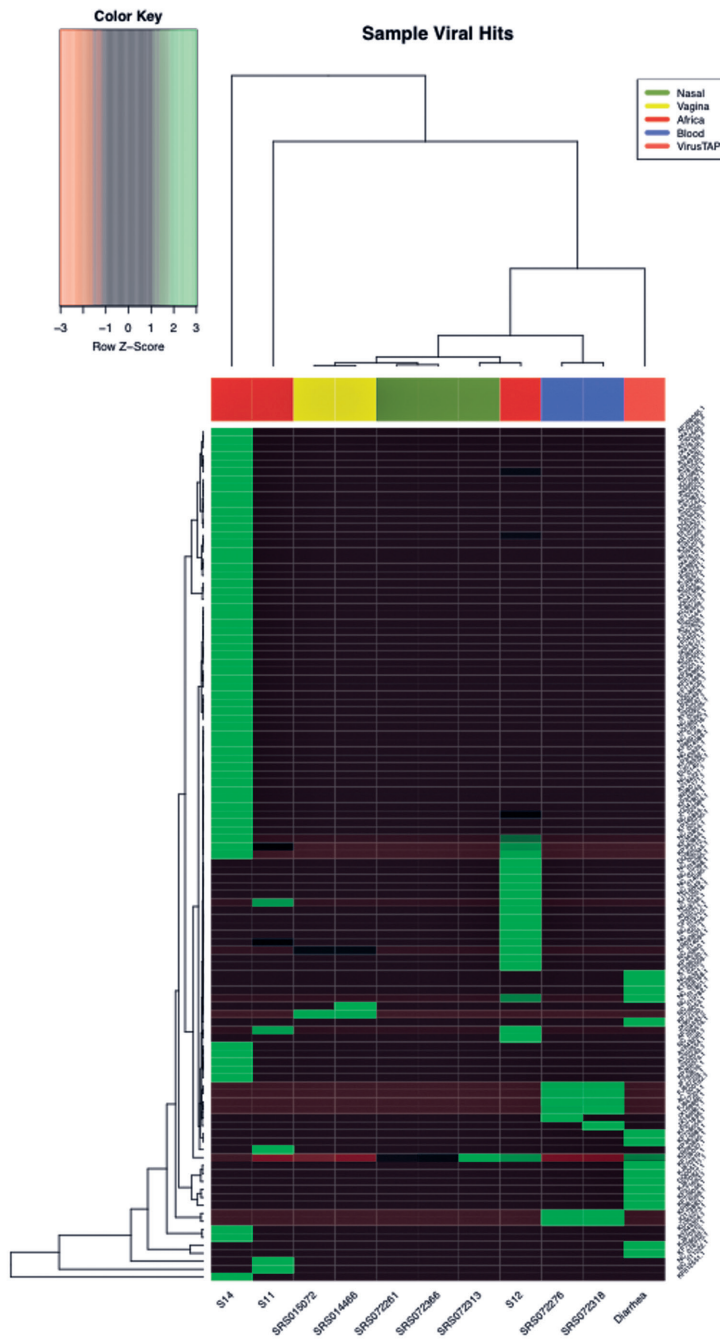


A



B

Figure 11. Vipe interactive visualization maps – (A) Virus populations are displayed as pie slices where sub-taxonomy information is shown via clicking. The pie chart sizes are relative to the total number of viruses relative to all samples in submission. (B) Sample alpha diversity scores are plotted with Shannon entropy and unique number of accessions.



C

(C) Clustered heat map is produced from R using viral population similarity and hierarchical structure, resulting in clustered grouping of HMP samples and distinct African and diarrhea sample virus profiles.

5.3.3 Web results and interactive features

Covered in materials chapter, Vipie performance and validation used 7 total virome samples from HMP (Healthy adult sample sites are stool, vagina and nasal.), 1 from VirusTAP and 3 from African rural stool samples. As expected, we found the highest amount and diversity of viruses in the African samples. Using the integrated viral table matrix as input, hierarchical clustering using paired wise similarity, correctly grouped the HMP samples by source site, and the VirusTAP diarrheal sample in its own group. The tool also correctly revealed high abundance of Rotavirus, the agent responsible for acute diarrhea. Figure 10 provides a figure of the interactive viral findings table where the samples are columns and the rows are viral accessions. Viral accessions can be collapsed according to taxonomy level where the sample viral hits are summed to include its sublevels. The cell background color is styled with red color and the intensity deepens according to viral hits. To account for read size variances, all sample hits are scaled to per 100,000 reads. Figure 11 shows the other main result plots with panel A clickable population pie charts, B showing Shannon alpha diversity described in Chapter 2 and in C, a hierarchical clustered sample heat map.

5.4 Performance and results management

With the integration of python parallel processing, end to end performance took 82 minutes for 11 samples. While the processing time is relative to server specification and load, for comparison, VirusTap the only other web based pipeline also performing de novo assembly required 17 minutes for 1 sample with prior removal of human reads. Currently, Vipie is the only web virome pipeline allowing multiple samples. Published Vipie results and visualizations can be accessed here: <https://binf.uta.fi/vipie/results.html?key=eLZPuObVoU>

The tabular accession result file, qc report and intermediate outcomes, including contigs and supporting reads, can be downloaded from similar secured encrypted links within 30 days of job completion. To promote and allow investigators the option to share their results with collaborators, the secured results can be accessed without user credentials. Since publication, more than 30 institutions in 20 countries have used Vipie. We discussed further role of Vipie in T1D autoimmunity research, potential limitations and improvements in the next chapter.

6 DISCUSSION

Exceeding global population growth, T1D incidence rates are worrisomely increasing and support for environmental factors is strengthening as multiple viral and bacterial associations have been published [Hyöty et al. 1995, Laitinen et al. 2014, Lernmark et al. 2016]. These disease trends are driving investigators to form large international cohorts, such as DIABIMMUNE (Finland, Estonia and Russia) and The Environmental Determinants of Diabetes in the Young (TEDDY) based in Finland, Germany, Sweden and USA to actively designed from-birth prospective cohorts with matched case and control children focusing on serial collection of stool, nasal and blood samples for microbiome and gene expression studies. These large omics repositories are motivating the need for continued development of novel bioinformatics tools and systematic application of existing methodologies. Historically T1D has been implicated with enterovirus [Knip et al. 2005, Hyöty et al. 1995] and urgently pressing virome research efforts is that enterovirus genetic sequences have been found in fresh pancreatic tissue donated by living and recently diagnosed T1D patients [Krogvold et al. 2015]. Interestingly and warranting further systematic investigations, enterovirus exposures have induced strong interferon pathway, key signature of host antiviral activity, responses and related networks in T1D susceptible children [Lietzen et al. 2017].

Publication II, supporting innovation and foresight of DIPP study, based in Finland, is one of the first analysis geared at investigating roles of gut virome and bacteriome towards development of T1D autoimmunity. Using stool samples taken approximately every three months from young Finnish children prior to autoimmunity seroconversion and tightly matched control, we found four operational taxonomic units to be significantly less abundant in case samples compared to controls – the most significant came from the *Bacteroides* and *Bifidobacterium phyla*. Human viruses were detected from NGS and confirmed using PCR in 10.4% of the stool samples from the same design. We constructed correlation networks between gut bacterial abundance to each other and viral phage populations (found in more than 50% of the samples) processed from virome using the same stool samples. From the correlation network, we found a novel bacterial interaction between crAssphage, and *Bacteroides dorei*. CrAssphage host is currently unknown as the virus was found computationally via cross assembly and its sequences are reported to be the most common in the human gut [Dutilh et al. 2014]. As phages exist to attack specific bacterial strains, they obviously have large impact on bacterial population and adversely

impact gut bacterial dysbiosis. We believe this is the first study to apply 16S sequence in conjunction with shotgun virome to identify crAssphage host.

Additionally, regarding gut bacteriome imbalance and our findings in Publication II, a prior report [Davis-Richardson et al. 2015] had reported higher abundance of Bacteroides phylum, notably species *Bacteroides dorei* in case samples [Davis-Richardson et al. 2014] compared to controls, also from DIPP, albeit only one DIPP clinic was included (University Hospital in Turku, a southwest Finnish city). The samples within Publication II were taken from Tampere and Oulu, about 200 and 500 kilometers from Turku. Reported in Publication II, the OTU threshold was refined to 100% and reinforce the autoimmunity statistical significant association with *Bacteroides vulgatus* and *Bacteroides caccae* abundance reduction but no longer significant for *Bacteroides dorei*. While it is possible that the disagreement originates in sequencer error or limited sensitivity, it can also be that gut imbalance of Bacteroides, as reported in both publications, is associated with autoimmunity and that the manifestation occurs in different species due to locale and phages fluctuations. Both studies also mostly only included gut samples taken prior to 2 years old, where the gut is still developing and highly impacted by antibiotics and diet. In support of a DIABIMMUNE study on hygiene hypothesis, Vatanen and colleagues [Vatanen et al. 2016] found less *Bacteroides* species in Russian children compared to Finnish and Estonian children. Instead, there were higher proportions of *Actinobacteria*, a gram positive phylum, and *Bifidobacterium* in the Russian population. It reasons that higher percentages of Finnish children, with higher risk of autoimmune diseases, had early colonization of *Bacteroides*. The quality of LPS of *Bacteroides* differs from that of *E. coli* which in turn was more prevalent in Russian population, leading to different innate immune stimulation. One issue is that children in Estonia, as reported had the same Bacteroides distribution but Finnish children have about 2 times disease incidence. Also complicating matters is that researchers from DIABIMMUNE project, [Yassour et al. 2016] reported that 20% of Finnish children born vaginally had very low Bacteroides during first 18 months of life. We found that there was only one nucleotide base difference within 253 bases of the 16S V4 amplicon region between *B. dorei* and *B. vulgatus*. Clearly exceeding 99% and there are likely other regions with high similarity between strains, this underscores potential limitations of 16S analysis based on similarity clustering and warrants additional checks prior to key conclusions based solely on standard 16S pipelines.

For Publication III and central to this thesis, we published Vipie, a high performance, easy to use web based virome pipeline capable of processing multiple virome samples. Vipie web is commonly used for general virome population profiling and processed job results are securely presented as interactive population maps, quality reports, and searchable tables. Using mock simulation data consisting of human, bacterial and viral, we reported comparatively good results in terms of precision and recall. The publication demonstrated Vipie's built in clustering function and correctly grouped the HMP healthy adult samples

and separated African rural children samples from one Japanese gastroenteritis based on relevant virome profiles.

The roots of Vipie originated from Publication II and its core components were developed and tested for the analysis of DIPP samples. As listed in Table 2 in chapter 2, there are several viral pipelines available but as there were not any easy to use pipelines allowing de novo assembly, Vipie was constructed and eventually web automated and published. MetaShot [Fosso et al. 2017], a Linux based pipeline, published around the same time period and impressively reported better precision (98.30, Vipie 96.85) and recall (98.19, Vipie 95.36) using mock simulation data originated from MetaShot. MetaShot performs a two step processing and the first step removes similar sequences to selected host, therefore viral reads possessing high similarity to human samples will be removed. These removals contribute to the difference in accuracy as Vipie reports but does not remove viral reads based on similarity to human. We feel that it is important to capture all possible viral reads, as a stretch of Cocksackievirus is highly similar to the predictive islet autoantibody GAD65 [Atkinson 1995]. Interestingly, human adenovirus protein homology to gluten leading to cross reaction have also been reported to contribute to celiac disease pathogenesis [Kagnoff et al. 1984].

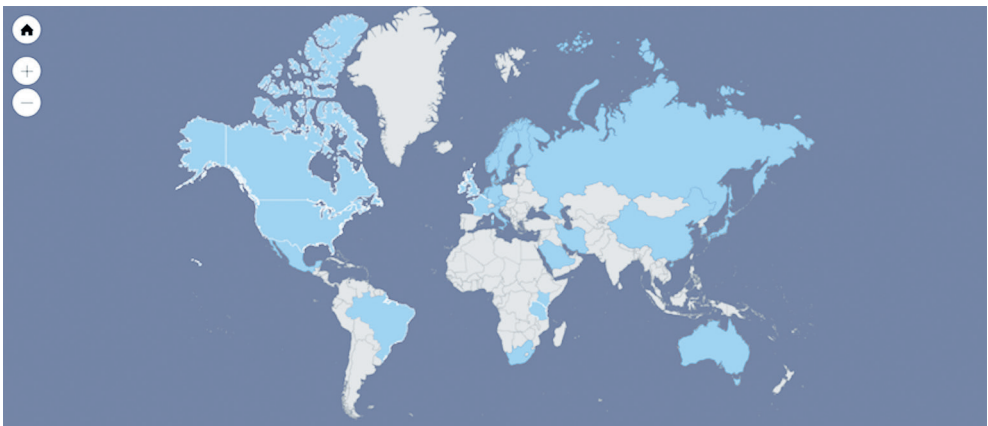


Figure 12. Vipie users – investigators from more than 20 countries across biomedical, disease control centers and interestingly agriculture and blood donation centers have registered and used Vipie.

Shown in Figure 12, remarkably investigators from more than 30 institutes in 20 countries working in biomedical and agriculture industries, hospitals, disease control centers as well as blood donation centers have registered and processed jobs in Vipie. This warm reception speaks to the benefits of web based tools and also Vipie intuitive design – emphasizing ease of use, accuracy and insightful visualizations. At the same time there are many labs analyzing virome across different domains and as the viral references are generalized and relatively limited (7,477 full virus genomes on NCBI as of Jan 2018). As viruses

are diverse, lacking a universal biomarker and known to have faster mutation rates, most virome reads will not map and as a result, Vipie reports these unknown reads as dark matter. In Publication II and shown in Figure 13, more than 70% of the DIPP and HMP NGS reads were classified as dark matter. This is clearly a serious concern involving the entire viral research community. One possible way forward is to treat long unmapped contigs as quasi-virus reference genomes, generated from de novo assembly, and perform remapping with raw reads. Quasi-contigs with high matches and coverage would become valuable novel virus reference candidates. Heintz-Buschart and colleagues impressively clustered dark bacterial contigs in the context of gut microbial within a familial T1D study [Heintz-Buschart et al. 2016].

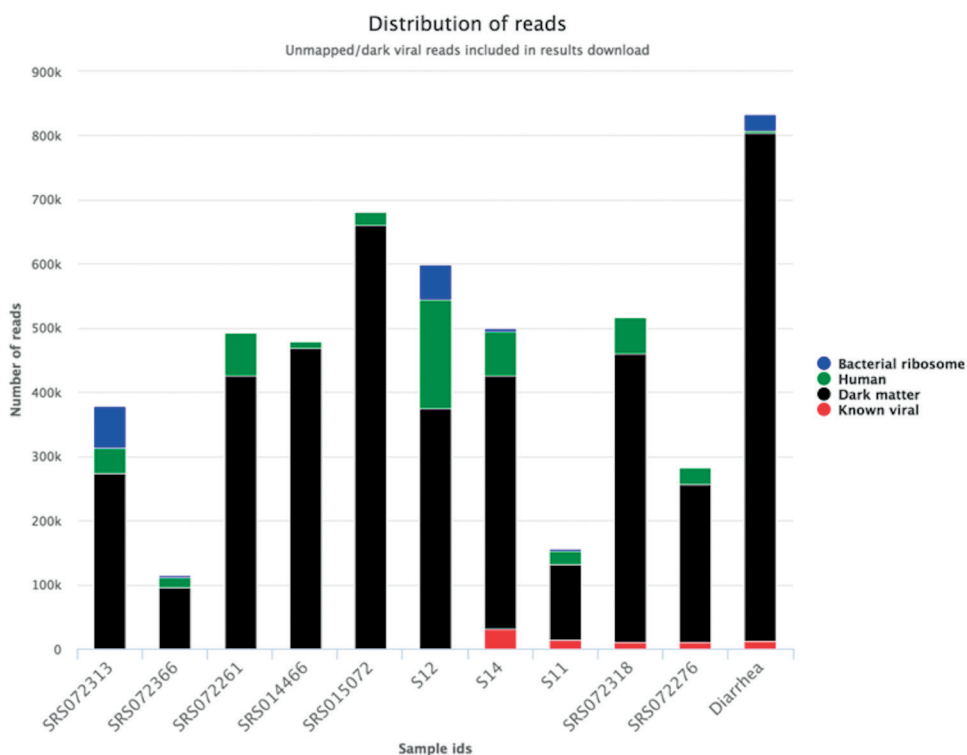


Figure 13. Viral dark matter – most virome samples contain many unmapped viral reads. Shown in black, dark matter reads dominate the proportion of reads in across all samples used in Vipie publication, including 7 Human Microbiome Project samples (name beginning with SRS) from vagina, nasal and blood.

Another limitation involves Vipie pipeline in Publication III as it identifies viruses based on their whole genome and do not further classify the matches. While virus genomes generally are simple and exist to infect other organisms, there are important structural regions such

as capsid area, the protein shell. Particularly relevant is within Enterovirus, the capsid consists of four proteins, VP₁, VP₂, VP₃ and VP₄ and they are the most distinctive and metagenomic reads mapped to a capsid area warrant higher confidences. Vipie publication was aimed at a general virus profiling design for wider consumption but within targeted disease analysis, we anticipate and recommend complementing Vipie findings with custom steps. The steps would include assessing assigned Vipie reads with selected genome regions, essentially epitopes or genomic windows of interest. To minimize the effect of base mutations, the regions can be comprehensively translated to proteins. These hits would also be more qualify to answer the important question of persistent or novel infections over time within the same host.

A highlight of Vipie has been the successful adoption of Vipie for TEDDY virome analysis using its high performance computation cluster. We have systematically processed and identify virus exposure profiles from more than 10,000 stool samples from case children confirmed for T1D autoimmunity and matched controls. These samples were collected monthly starting from 3 months old until 4 years old. In addition, thousands of nasal and blood samples, drawn every 3 months, have also been processed. We are currently applying conditional logistic models on the virome profiles. Concurrently, we are also integrating and assessing CVB serotypes with Vipie mapped reads to overcome specificity limitations as described earlier. Vipie, lightly refactored for Linux, is ideal for processing large scaled samples as it uses multi processor designed, centralized database and individual sample processing is independent of other samples. Interestingly, *Circoviridae* was recently found to be significantly less [Guo et al. 2017] in T1D autoimmunity control children though the study, using 10 monthly samples, only involved 11 children confirmed for T1D autoimmunity.

A comprehensive virome roadmap [Delwart 2013] project has been proposed and as stated previously, there are clearly multiple motivations and important benefits. For example, reovirus was recently implicated as a potential dietary interplay trigger for inflammation in celiac disease [Brouziat et al. 2017]. Notably children suffering from celiac disease have significantly higher risks, hazard ratio greater than 2, for T1D [Ludvigsson et al. 2006]. One central step would be to integrate metadata, including family background, health and dietary history from the large T1D related cohorts and annotate them to relevant microbiome and control NGS files produced. It would be revealing to find regional and seasonal viral exposure patterns prior to inflammation episodes. Signs of a potential leaky and aberrant gut could be phages sequences found in non-stool samples and large sudden changes in phage viruses between relevant time points. In addition to storing and annotating known viral sample profiles, a parallel effort could be building of the unknown dark viral matter central resource, capturing also the assembly methods and key parameters as they would be beneficial for contig validation and reproducibility. Long dark viral contigs with 100% similarity or sub-stretches are good novel virus candidates. In addition, reports have cited that certain genome mutations and deletions, such as 5'

end of Coxsackievirus [Tracy et al. 2015] can increase survival fitness in pancreas. Also interesting is that Coxsackievirus virulence, the potential lethality towards the host cells, has been shown to be variable based on single amino acid changes [Halim & Ramsingh 2000]. Taken all together, a comprehensive viral roadmap will offer an optimal resource for detection of acute versus persistent infections, strain genotyping and mutation virulence assessment. The metadata collected will also be valuable for future statistical modelling efforts and temporal study on mutation rates and selection. As the sensitivity of single cell sequencing improves [Haque et al. 2017], their applications together with improvements of selectively staining infected donated pancreatic islet beta cells will shed more light on key current viral associations and mutually drive the need for in bioinformatics.

In addition to virome analysis, this thesis also contributes to novel and intuitive bioinformatics visualization. Publication I introduces POMO, an easy to use web based application designed to plot genomic networks for multiple model organisms. POMO was started in support of displaying and interactively filtering large scaled TCGA pairwise and random forest [Breiman 2001] genomic networks in Circos like layout. With redesign and code refactoring, we generalized the tool to support networks independent of mining and statistical methods. We also extended the reference databases beyond human to benefit investigators working on other organisms including fly, mouse, rat, worm, yeast, *Arabidopsis* and *E. coli*. POMO plotting can also be applied to metagenomics, involving multiple constructed genomes and their energy usages represented by protein activity and polymorphism rates [Muller et al., 2015]. Visualizations from genomic networks and annotations are automatically plotted, filtered and downloaded across multiple layouts on web browsers. POMO is also capable of visualizing comparative genomic results, shown in Publication I Figure 4, demonstrating human and mouse phenology from obesity-abnormal food intake study [McGary et al. 2010]. Very recently, CGDV [Jha et al. 2017], a web tool for circular visualization of omics data was published, like POMO, the application has support for multiple model organisms and it also cited that Circos installation and usage can be challenging for biologists.

7 ACKNOWLEDGEMENTS

The study was conducted in Computational Biology and Virology groups and Faculty of Medicine and Life Sciences in University of Tampere. I am very grateful to my supervisors Reija Autio, Professor Heikki Hyöty and Professor Matti Nykter. They have been truly great and uniquely admirable. Intelligent, resourceful and genuinely caring. I am lucky, personally and academically, to have connected with them. Greetings to all current and former labmates, my apologies for not naming everyone. The thesis has been generously funded by ADELE, BMT graduate school, DIPP and TEDDY studies.

Special thanks to Professor Ondrej Cinek, along with Lenka Kramna, for introducing and patient guidance with virome and microbiome analysis.

Thank you to reviewers Docent Tarja Sironen and Docent Christophe Roos for being cordial and thesis help via insightful corrections and comments. Thanks to Jukka Intosalmi for Finnish abstract translation. Thank you to Jaakko Nevalainen and Niina Lietzen for their advisory committee roles.

I like to personally thank Ilya Shmulevich and his lab for opening the realm of computational biology and always being so supportive. Infocore and TCGA/RE mates Ryan, Dick, Hector, Kalle, Lesley, Brady, Vesteinn, Sheila, Theo, Timo.

Grateful to Aimée Dudley and her lab for teaching me some yeast fluffy genetics and cool memories. Cathy, Adrian, Gareth, Amy, Cecilia. And Teresa rest in peace.

Thank you Ilya for allowing me to come to Finland, Olli Yli-Harja for indulging the whim. Virpi K, Matti Annala, Virve L and Ulla S for helping with transition.

Thank you Patrick May for hiring me at LCSB, giving me plenty of freedom and Alexander Skupin for sharing our Esch flat. Professors Karsten Hiller and Reinhard Schneider for being flexible and supportive.

TEDDY EAP – Elaheh Moradi, Kirsi Granberg, Suvi Luoto, Sami Oikarinen, Tomi Häkkinen. Thanks to Heini Huhtala and Juha Kesseli for helpful statistics discussions. Kudos Katri Lindors, I hope MP199 will see the light.

Along the meandering trail, I have been blessed meeting good friends dispersed here and there – Susie Purves, Maria Liivrand and Nico Bouvy, Intosalmi family, Patrick May, Skupin family, Mafalda Galhardo, Mike Ross, Petteri Vakkila, Kai Baer, Tuomas

Luukkonen, Tiblu Tökkäri. Kiitos Telakan henkilökunnalle. Some favorite places – Finnish libraries and Kauppi woods. Sweetie Bill Moomau (nico and galinghas). My parents, siblings and their families. Love to Ninni Luhtasaari and her family.

Tampere, May 2018

Jake Lin

8 REFERENCES

- Abbas A (2006) Basic Immunology. Elsevier ISBN 978-1-4160-2974-8.
- Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methé B, Schloss PD, Gevers D, Mitreva M, Huttenhower C. (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol.* 2012 Jun;8(6):e1002358.
- Adler K, Mueller DB, Achenbach P, Krause S, Heninger AK, Ziegler AG, Bonifacio E (2011) Insulin autoantibodies with high affinity to the bovine milk protein α casein. *Clin Exp Immunol* 164: 42–49.
- Alonso-Aleman D, et al. (2014) Further Steps in TANGO: improved taxonomic assignment in metagenomics. *Bioinformatics* 30(1):17–23.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.
- Anthony SJ, Epstein JH, Murray KA, Navarrete-Macias I, Zambrana-Torrel CM, Solovyov A, Ojeda-Flores R, Arrigo NC, Islam A, Ali Khan S, Hosseini P, Bogich TL, Olival KJ, Sanchez-Leon MD, Karesh WB, Goldstein T, Luby SP, Morse SS, Mazet JA, Daszak P, Lipkin W (2013) A strategy to estimate unknown viral diversity in mammals. *MBio.* Sep 3;4(5):e00598-13. doi: 10.1128/mBio.00598-13.
- Ascher H, Krantz I, Kristiansson B (1991) Increasing incidence of coeliac disease in Sweden. *Arch Dis Child*; 66: 608–11.
- Atkinson MA (1997) Molecular Mimicry and the Pathogenesis of Insulin- dependent Diabetes Mellitus: Still Just an Attractive Hypothesis. *Annals of Medicine*, 29:5, 393-399, DOI: 10.3109/07853899708999368.
- Atkinson MA (2012) The Pathogenesis and Natural History of Type 1 Diabetes. *Cold Spring Harb Perspect Med Nov*; 2(11): a007641. doi: 10.1101/cshperspect.a007641.
- Atkinson MA, Eisenbarth GS, Michels AW (2014) Type 1 diabetes. *Lancet.* Jan 4; 383(9911):69-82
- Attar N (2016) Viral evolution: More of the world's a phage. *Nature reviews Microbiology* doi:10.1038/nrmicro.2016.58.
- Baum LE, Petrie T (1966) Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics.* 37 (6): 1554–1563.
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B.* 57 (1): 289–300. MR 132539.
- Berger B, Peng J, Singh M (2013) Computational solutions for omics data. *Nat Rev Genet* 2013, 14:333–346.
- Bergman EN (1990) Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. *Physiol Rev.* Apr; 70(2):567–90.
- Blinkova O, Victoria J, Li Y, et al. (2010) Novel circular DNA viruses in stool samples of wild-living chimpanzees. *J Gen Virol* 91:74–86.
- Botstein D, Chervitz SA, Cherry JM (1997) Yeast as a Model Organism. *Science* Aug 29; 277(5330): 1259–1260.

- Bouziat R, Hinterleitner R, Brown JJ, Stencel-Baerenwald JE, Ikizler M, Mayassi T, Meisel M, Kim SM, Discepolo V, Pruijssers AJ, Ernest JD, Iskarpatyoti JA, Costes LM, Lawrence I, Palanski BA, Varma M, Zurenski MA, Khomandiak S, McAllister N, Aravamudhan P, Boehme KW, Hu F, Samsom JN, Reinecker HC, Kupfer SS, Guandalini S, Semrad CE, Abadie V, Khosla C, Barreiro LB, Xavier RJ, Ng A, Dermody TS, Jabri B (2017) Reovirus infection triggers inflammatory responses to dietary antigens and development of celiac disease. *SCIENCE* 07 APR: 44-50 doi: 10.1126/science.aah5298.
- Bray JR, Curtis JT (1957) An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs* 27:325-349.
- Breiman L (2001) Random Forests. *Machine Learning*, 45(1):5-32.
- Brennan CW, Verhaak RG, et al., TCGA Research Network (2013) The somatic genomic landscape of glioblastoma. *Cell* Oct 10;155(2):462-77. doi: 10.1016/j.cell.2013.09.034. Erratum in: *Cell* Apr 24;157(3):753.
- Bressler R, Lin J, Eakin A, Robinson T, Kreisberg R, Rovira H, Knijnenburg T, Boyle J, Shmulevich I (2012) Fastbreak: a tool for analysis and visualization of structural variations in genomic data. *EURASIP J Bioinform Syst Biol*. Oct 9; (1):15. doi: 10.1186/1687-4153-2012-15.
- Brugman S, Klatter FA, Visser JT, Wildeboer-Veloo AC, Harmsen HJ, Rozing J, Bos NA (2006) Antibiotic treatment partially protects against type 1 diabetes in the Bio-Breeding diabetes-prone rat. Is the gut flora involved in the development of type 1 diabetes? *Diabetologia*. Sep; 49(9):2105-8.
- Buschard K (2011) What causes type 1 diabetes? Lessons from animal models. *APMIS Suppl*. Jul;(132):1-19. doi: 10.1111/j.1600-0463.2011.02765.x.
- Caine EA, Moncla LH, Ronderos MD, Friedrich TC, Osorio JE (2016) A Single Mutation in the VP1 of Enterovirus 71 Is Responsible for Increased Virulence and Neurotropism in Adult Interferon-Deficient Mice *J. Virol*. October vol. 90 no. 19 8592-8604.
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data *Nature Methods*, 7(5):335-336.
- Cardwell CR, Stene LC, Joner G, Cinek O, Svensson J, Goldacre MJ, Parslow RC, Pozzilli P, Briggs G, Stoyanov D, Urbonaite B, Sipetić S, Schober E, Ionescu-Tirgoviste C, Devoti G, de Beaufort CE, Buschard K, Patterson CC (2008) Caesarean section is associated with an increased risk of childhood-onset type 1 diabetes mellitus: a meta-analysis of observational studies. *Diabetologia* May; 51(5):726-35.
- Chao A, Shen TJ (2003) Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*. 10 (4): 429-443. doi:10.1023/A:102609620472.
- Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* Sep; 6(9):677-81.
- Chen X, Ishwaran H (2012) Random Forests for Genomic Data Analysis. *Genomics* 99(6):323-329. doi:10.1016/j.ygeno.2012.04.003.
- Chikhi R, Medvedev P (2013) Informed and Automated k-Mer Size Selection for Genome Assembly. *Bioinformatics* Vol 30, Issue 1, 31-37.

- Cohn A, Sofia AM, Kupfer SS (2014) Type 1 diabetes and celiac disease: clinical overlap and new insights into disease pathogenesis. *Current diabetes reports* 14:517.
- Coleman TJ, Gamble DR, Taylor KW (1973) Diabetes in mice after Coxsackie B4 virus infection. *Br Med J* 3:25–27.
- Concannon P, Rich SS, Nepom GT (2009) Genetics of type 1A diabetes. *N Engl J Med* 360: 1646–1654.
- Coppieters KT, Boettler T, von Herrath M (2012) Virus infections in type 1 diabetes. *Cold Spring Harb Perspect Med*. Jan;2(1):a007682. doi: 10.1101/cshperspect.a007682.
- Crick FHC (1958) On Protein Synthesis. In F.K. Sanders. *Symposia of the Society for Experimental Biology, Number XII: The Biological Replication of Macromolecules*. Cambridge University Press. pp. 138–163.
- Dabelea D, Pihoker C, Talton JW, D’Agostino RB, Fujimoto W, Klingensmith GJ, Lawrence JM, Linder B, Marcovina SM, Mayer-Davis EJ, et al. (2011) Etiological approach to characterization of diabetes type: The SEARCH for Diabetes in Youth Study. *Diabetes Care* 34: 1628–1633.
- Davis-Richardson AG, Ardisson AN, Dias R, Simell V, Leonard MT, Kempainen KM, Drew JC, Schatz D, Atkinson MA, Kolaczowski B, Ilonen J, Knip M, Toppari J, Nurminen N, Hyöty H, Veijola R, Simell T, Mykkänen J, Simell O, Triplett EW (2014) *Bacteroides dorei* dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Front Microbiol*. Dec 10;5:678. doi: 10.3389/fmicb.2014.00678. eCollection 2014.
- Delwart E (2013) A Roadmap to the Human Virome. *PLoS Pathog* 9(2): e1003146. doi:10.1371/journal.ppat.1003146.
- Donohoe DR, Garge N, Zhang X, Sun W, O’Connell TM, Bunger MK, Bultman SJ (2011) The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metab*. May 4;13(5):517–26. doi: 10.1016/j.cmet.2011.02.018.
- Duffy S, Shackelton LA, Holmes EC (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9, 267–276.
- Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GG, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, Felts B, Dinsdale EA, Mokili JL, Edwards RA (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun*. Jul 24;5:4498. doi: 10.1038/ncomms5498.
- Edgar, RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.
- Eisen, JA (2007) Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbes. *PLoS Biology*. 5 (3): e82. doi:10.1371/journal.pbio.0050082.
- Fernandez-Cassi X, Timoneda N, Gonzales-Gustavson E, Abril JF, Bofill-Mas S, Girones R (2017) A metagenomic assessment of viral contamination on fresh parsley plants irrigated with fecally tainted river water. *Int J Food Microbiol*. Sep 18;257:80–90. doi: 10.1016/j.ijfoodmicro.2017.06.001.
- Filippi CM, von Herrath MG (2008) Viral trigger for type 1 diabetes: pros and cons. *Diabetes* Nov;57(11):2863–71. doi: 10.2337/db07-1023.
- Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.*, 12, 42–58.

- Fosso B, Santamaria M, D'Antonio M, Lovero D, Corrado G, Vizza E, Passaro N, Garbuglia AR, Capobianchi MR, Crescenzi M, Valiente G, Pesole G (2017) MetaShot: an accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data. *Bioinformatics*, Volume 33, Issue 11, 1 June, pp. 1730–1732.
- Gale EA (2002) The Rise of Childhood Type 1 Diabetes in the 20th Century. *Diabetes*, Dec; 51(12): 3353–3361.
- Gardner SG, Bingley PJ, Sawtell PA, Weeks S, Gale EAM (1997) Rising incidence of insulin dependent diabetes in children aged under 5 years in the Oxford region. *BMJ* 315:713–717.
- Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, Gavin AC (2010) Visualization of omics data for systems biology. *Nat Methods*. Mar;7(3 Suppl):S56–68. doi: 10.1038/nmeth.1436.
- Ghazarian L, Diana J, Beaudoin L, Larsson PG, Puri RK, van Rooijen N, Flodström-Tullberg M, Lehuen A (2013) Protection against type 1 diabetes upon Coxsackievirus B4 infection and iNKT-cell stimulation: role of suppressive macrophages. *Diabetes* Nov;62(11):3785–96. doi: 10.2337/db12-0958.
- Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*; 11:R86 3.
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. *Science*. Oct 25; 274(5287):546, 563–7.
- Gronenborn B, Messing J (1978) Methylation of single-stranded DNA in vitro introduces new restriction endonuclease cleavage sites. *Nature*, 272, 375–377.
- Gundersen E (1927) Is Diabetes of Infectious Origin? *The Journal of Infectious Diseases*, 41:197–202.
- Halim S, Ramsingh AI (2000) A Point Mutation in VP1 of Coxsackievirus B4 Alters Antigenicity, *Virology* 269, 86–94 doi:10.1006/viro.2000.0188.
- Haller MJ, Schatz DA (2016) The DIPP project: 20 years of discovery in type 1 diabetes. *Pediatr Diabetes* 17:5–7. doi:10.1111/pedi.12398.
- Haque A, Engel J, Teichmann SA, Lönnberg T (2017) A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*. Aug 18;9(1):75. doi: 10.1186/s13073-017-0467-4.
- Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, Wampach L, Schneider JG, Hogan A, de Beaufort C, Wilmes P (2016) Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol*. Oct 10;2:16180. doi: 10.1038/nmicrobiol.2016.180.
- Huang W, Li L, Myers JR, Marth GT (2012) ART: a next-generation sequencing read simulator. *Bioinformatics* Feb 15;28(4):593–4. doi: 10.1093/bioinformatics/btr708.
- Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* 12, 115.
- Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res*. Sep;21(9):1552–60. doi: 10.1101/gr.120618.111.

- Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature*. Jun 13;486(7402):207–14. doi: 10.1038/nature11234.
- Hussein SM, Batada NN, Vuoristo S, Ching RW, Autio R, Narva E, Ng S, Sourour M, Hamalainen R, Olsson C, Lundin K, Mikkola M, Trokovic R, Peitz M, Brustle O, Bazett-Jones DP, Alitalo K, Lahesmaa R, Nagy A, Otonkoski T (2011) Copy number variation and selection during reprogramming to pluripotency. *Nature*, 471:58–62.
- Hyöty H, Hiltunen M, Knip M, Laakkonen M, Vahasalo P, Karjalainen J, Koskela P, Roivainen M, Leinikki P, Hovi T (1995) A prospective study of the role of coxsackie B and other enterovirus infections in the pathogenesis of IDDM. Childhood Diabetes in Finland (DiMe) Study Group. *Diabetes* 44: 652–657.
- Ilonen J, Reijonen H, Knip M, Simell O (1996) Population-based screening for IDDM susceptibility as a source of HLA-genotyped control subjects. *Diabetologia* 1996;39:123.
- Iweala OI, Nagler CR (2006) Immune privilege in the gut: the establishment and maintenance of non-responsiveness to dietary antigens and commensal flora. *Immunol Rev*. Oct;213:82–100.
- James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning with Applications in R. Springer ISBN 978-1-4614-7138-7.
- Jakobsson HE, Jernberg C, Andersson AF, Sjölund-Karlsson M, Jansson JK, Engstrand L (2010) Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLOS ONE*, 5(3): 1–12, 03.
- Jha V, Singh G, Kumar S, Sonawane A, Jere A, Anamika K (2017) CGDV: a webtool for circular visualization of genomics and transcriptomics data. *BMC Genomics*201718:823. <https://doi.org/10.1186/s12864-017-4169-5>.
- Jun HS, Yoon JW (2003) A new look at viruses in type 1 diabetes. *Diabetes Metab Res Rev*. Jan–Feb;19(1):8–31.
- Kagnoff MF, Austin RK, Hubert JJ, Bernardin JE, Kasarda DD (1984) Possible role for a human adenovirus in the pathogenesis of celiac disease. *J Exp Med*. Nov 1;160(5):1544–57.
- Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28, 27–30.
- Karvonen M, Viik-Kajander M, Moltchanova E, Libman I, LaPorte R, Tuomilehto J (2000) Incidence of childhood type 1 diabetes worldwide. Diabetes Mondiale (DiaMond) Project Group. *Diabetes Care* Oct;23(10):1516–26.
- Knip M, Veijola R, Virtanen SM, Hyöty H, Vaarala O, Akerblom HK (2005) Environmental triggers and determinants of type 1 diabetes. *Diabetes Dec*; 54 Suppl 2():S125–36.
- Kondrashova A, Mustalahti K, Kaukinen K, Viskari H, Volodicheva V, Haapala AM, Ilonen J, Knip M, Maki M, Hyöty H, and EpiVir Study (2008) Lower economic status and inferior hygienic environment may protect against celiac disease. *Annals of Medicine*, 40(3):223–231.
- Kramna L, Kolarova K, Oikarinen S, Pursiheimo J, Ilonen J, Simell O, Veijola R, Knip M, Hyöty H, Cinek O (2015) Gut virome sequencing in children with early islet autoimmunity. *Diabetes Care*. 2015;38:930–933.
- Krogvold L, Edwin B, Buanes T, Ludvigsson J, Korsgren O, Hyöty H, Frisk G, Hanssen KF, Dahl-Jørgensen K (2014) Pancreatic biopsy by minimal tail resection in live adult patients at the onset of type 1 diabetes: experiences from the DiViD study. *Diabetologia*. Apr; 57(4):841–3.

- Krogvold L, Edwin B, Buanes T, Frisk G, Skog O, Anagandula M, Korsgren O, Undlien D, Eike MC, Richardson SJ, Leete P, Morgan NG, Oikarinen S, Oikarinen M, Laiho JE, Hyöty H, Ludvigsson J, Hanssen KF, Dahl-Jørgensen K (2015) Detection of a low-grade enteroviral infection in the islets of langerhans of living patients newly diagnosed with type 1 diabetes. *Diabetes May*;64(5):1682–7. doi: 10.2337/db14-1370.
- Krzywinski, M, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an Information Aesthetic for Comparative Genomics. *Genome Res* 19:1639–1645. doi:10.1101/gr.092759.109.
- Laitinen OH, Honkanen H, Pakkanen O, Oikarinen S, Hankaniemi MM, Huhtala H, Ruokoranta T, Lecouturier V, André P, Harju R, Virtanen SM, Lehtonen J, Almond JW, Simell T, Simell O, Ilonen J, Veijola R, Knip M, Hyöty H (2014) Coxsackievirus B1 is associated with induction of β -cell autoimmunity that portends type 1 diabetes. *Diabetes Feb*;63(2):446–55. doi: 10.2337/db13-0619.
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat.Methods* 9, 357–359.
- Lazarova DL, Bordonaro M, Carbone R, Sartorelli AC (2004) Linear relationship between Wnt activity levels and apoptosis in colorectal carcinoma cells exposed to butyrate. *Int J Cancer Jul* 1; 110(4):523–31.
- Lee HS, Burkhardt BR, McLeod W, Smith S, Eberhard C, Lynch K, Hadley D, Rewers M, Simell O, She JX, Hagopian B, Lernmark A, Akolkar B, Ziegler AG, Krischer JP, TEDDY study group (2014) Biomarker discovery study design for type 1 diabetes in The Environmental Determinants of Diabetes in the Young (TEDDY) study. *Diabetes Metab Res Rev Jul*;30(5):424–34. doi: 10.1002/dmrr.2510.
- Lehman, A (2005) *Basic Univariate And Multivariate Statistics: A Step-by-step Guide*. Cary, NC: SAS Press. p. 123. ISBN 1-59047-576-3.
- Lehuen A, Diana J, Zaccane P, Cooke A (2010) Immune cell crosstalk in type 1 diabetes. *Nat Rev Immunol Jul*;10(7):501-13. doi: 10.1038/nri2787.
- Lernmark Å (2016) Environmental factors in the etiology of type 1 diabetes, celiac disease, and narcolepsy. *Pediatr Diabetes Jul*;17 Suppl 22:65–72. doi: 10.1111/pedi.12390.
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754–60.
- Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–9.
- Lietzen N, An LTT, Jaakkola MK, Kallionpää H, Oikarinen S, Mykkänen J, Knip M, Veijola R, Ilonen J, Toppari J, Hyöty H, Lahesmaa R, Elo LL (2018) Enterovirus-associated changes in blood transcriptomic profiles of children with genetic susceptibility to type 1 diabetes. *Diabetologia Feb*;61(2):381–388. doi: 10.1007/s00125-017-4460-7.
- Lin HC, Wang CH, Tsai FJ, Hwang KP, Chen W, Lin CC, Li TC (2015) Enterovirus infection is associated with an increased risk of childhood type 1 diabetes in Taiwan: a nationwide population-based cohort study. *Diabetologia Jan*;58(1):79–86. doi: 10.1007/s00125-014-3400-z.
- Lodish H, Berk A, Zipursky SL (2000) *Viruses: Structure, Function, and Uses*, Molecular Cell Biology. 4th edition. New York: W. H. Freeman.
- Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics Oxf Engl* 26:2347–2348.
- Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15, p. 550. doi: 10.1186/s13059-014-0550-8.

- Lozupone CA, Hamady M, Kelley ST, Knight R (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* 73(5):1576–85. doi:10.1128/AEM.01996-06.
- Ludvigsson JF, Ludvigsson J, Ekbom A, Montgomery SM (2006) Celiac disease and risk of subsequent type 1 diabetes: a general population cohort study of children and adolescents. *Diabetes Care* Nov;29(11):2483–8.
- Lupton, JR (2004) Microbial Degradation Products Influence Colon Cancer Risk: the Butyrate Controversy. *J Nutr.* Feb;134(2):479–82.
- Mangani C, et al. (2015) Effect of complementary feeding with lipid-based nutrient supplements and corn-soy blend on the incidence of stunting and linear growth among 6- to 18-month-old infants and children in rural Malawi. *Matern Child Nutr.* Dec;11 Suppl 4:132–43. doi: 10.1111/mcn.12068.
- McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey, MR, O'Brien, JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright, GD (2013) The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, 57 (7):3348–3357.
- McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, Caporaso JG (2012) The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience* 1:7. doi:10.1186/2047-217X-1-7.
- McMurdie P J, Holmes S (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8: e61217.
- McMurdie P J, Holmes S (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 10: e1003531.
- Medawar PD, Medawar JS (1983) *Aristotle to Zoos A Philosophical Dictionary of Biology.* Cambridge, MA: Harvard University Press.
- Mills CE, Robins JM, Lipsitch M (2004) Transmissibility of 1918 pandemic influenza. *Nature* 432 (7019): 904–6. doi:10.1038/nature03063.
- Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A (2015) Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand Genomic Sci.* Mar 30;10:18. doi: 10.1186/1944-3277-10-18.
- Mueller E, Pinel N, Laczny C, Hoopmann M, Narayanasamy S, Lebrun L, Roume H, Lin J, May P, Hicks N, Heintz-Buschart A, Wampach L, Liu C, Price L, Gillece J, Guignard C, Schupp J, Vassis N, Baliga N, Moritz R, Keim P, Wilmes P (2015) Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nature Communications* DOI: 10.1038/ncomms6603.
- Narzisi G, Mishra B (2011) Comparing De Novo Genome Assembly: The Long and Short of It. Aerts S, ed. *PLoS ONE* 6(4):e19175. doi:10.1371/journal.pone.0019175.
- Närvä E, Autio R, Rahkonen N, Kong L, Harrison N, Kitsberg D, Borghese L, Itskovitz-Eldor J, Rasool O, Dvorak P, Hovatta O, Otonkoski T, Tuuri T, Cui W, Brustle O, Baker D, Maltby E, Moore HD, Benvenisty N, Andrews PW, Yli-Harja O, Lahesmaa R (2010) High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. *Nat Biotechnol* 28:371–377.
- Palmer JP, Hampe CS, Chiu H, Goel A, Brooks-Worrell BM (2005) Is latent autoimmune diabetes in adults distinct from type 1 diabetes or just type 1 diabetes at an older age? *Diabetes* 54: 62–67.

- Palsson B, Zengler K (2010) The challenges of integrating multi-omic data sets. *Nat Chem Biol* 6:787–789.
- Patterson CC, Dahlquist GG, Gyürüs E, Green A, Soltész G, EURODIAB Study Group (2009) Incidence trends for childhood type 1 diabetes in Europe during 1989–2003 and predicted new cases 2005–20: a multicentre prospective registration study. *Lancet Jun* 13;373(9680):2027–33. doi: 10.1016/S0140-6736(09)60568-7.
- Patterson KD, Pyle GF (1991) The geography and mortality of the 1918 influenza pandemic. *Bull Hist Med*. Spring;65(1):4–21. <http://www.jstor.org/stable/44447656>.
- Pearson K (1895) Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.
- Pociot F, Lernmark Å (2016) Genetic risk factors for type 1 diabetes. *Lancet Jun* 4;387(10035):2331–2339. doi: 10.1016/S0140-6736(16)30582-7.
- Pugliese A, Yang M, Kusmarteva I, Heiple T, Vendrame F, Wasserfall C, Rowe P, Moraski JM, Ball S, Jebson L, Schatz DA, Gianani R, Burke GW, Nierras C, Staeva T, Kaddis JS, Campbell-Thompson M, Atkinson MA (2014) The Juvenile Diabetes Research Foundation Network for Pancreatic Organ Donors with Diabetes (nPOD) Program: goals, operational model and emerging findings. *Pediatr Diabetes Feb*; 15(1):1–9.
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res*. 41 (D1): D590–D596.
- Rampelli S, Soverini M, et al. (2016) ViromeScan: a new tool for metagenomic viral community profiling. *BMC Genomics* 17:165.
- Redondo MJ, Fain PR, Eisenbarth GS (2001) Genetics of type 1A diabetes. *Recent Prog Horm Res* 56: 69–89.
- Rodríguez-Díaz J, et al. (2014) Presence of human enteric viruses in the stools of healthy Malawian 6-month-old infants. *J Pediatr Gastroenterol Nutr*. 58(4):502–4. doi:10.1097/MPG.000000000000215.
- Roux S, Faubladier M, et al. (2011) Metavir: a web server dedicated to virome analysis. *Bioinformatics* 27(21):3074–5.
- Savage DC (1977) Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol*. 31():107–33.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541.
- Schwikowski B, Uetz P, Fields S (2000) A network of protein-protein interactions in yeast. *Nat Biotechnol*. Dec;18(12):1257–61.
- Shannon CE (1948) A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423 and 623–656.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. Nov;13(11):2498–504.
- Simpson EH (1949) Measurement of diversity. *Nature* 163: 688. doi:10.1038/163688a0.

- Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen E (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* Jan 4;46(D1):D661–D667. doi: 10.1093/nar/gkx1064.
- Stene LC, Nafstad P (2001) Relation between occurrence of type 1 diabetes and asthma *Lancet* 2001;357:607.
- Strachan DP (1989) Hay fever, hygiene, and household size. *British Medical Journal*, 299(6710):1259.
- Streisand R, Monaghan M (2014) Young children with type 1 diabetes: challenges, research, and future directions. *Current Diabetes Reports* 01 Jan 14(9):520.
- Szymanski M, Zielezinski A, Barciszewski J, Volker A, Wojciech E, Karlowski M (2016) 5SRNAdb: an information resource for 5S ribosomal RNAs. *Nucleic Acids Research*, Volume 44, Issue D1, 4 January, Pages D180–D183. doi.org/10.1093/nar/gkv1081.
- Tattersall RB (2009) *Diabetes The biography*. Oxford University Press.
- TEDDY Study Group (2008) The Environmental Determinants of Diabetes in the Young (TEDDY) Study. *Ann N Y Acad Sci.* Dec;1150:1–13. doi: 10.1196/annals.1447.062.
- The Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature.* Feb 15;409(6822):860–921.
- Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA (2013) Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief Bioinform.* May;14(3):315–26. doi: 10.1093/bib/bbs034.
- Tracy S, Smithee S, Alhazmi A, Chapman N (2015) Coxsackievirus can persist in murine pancreas by deletion of 5' terminal genomic sequences. *J Med Virol.* Feb;87(2):240–7. doi: 10.1002/jmv.24039.
- Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N (2015) MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, 12, 902–903.
- Vaarala O (2012) Gut Microbiota and Type 1 Diabetes. *Rev Diabet Stud.* Winter; 9(4): 251–259.
- Vatanen T, Kostic AD, d’Hennezel E, Siljander H, Franzosa EA, Yassour M, Kolde R, Vlamakis H, Arthur TD, Hämmäläinen AM, Peet A, Tillmann V, Uibo R, Mokurov S, Dorshakova N, Ilonen J, Virtanen SM, Szabo SJ, Porter JA, Lähdesmäki H, Huttenhower C, Gevers D, Cullen TW, Knip M, DIABIMMUNE Study Group, Xavier RJ (2016) Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell Jun 2;165(6):1551.* doi: 10.1016/j.cell.2016.05.056.
- Von Bertalanffy L (1950) An outline of general system theory. *British Journal for the Philosophy of Science* 1, 134–165.
- Wilcoxon, F (1945) Individual comparisons by ranking methods. *Biometrics Bulletin* 1 (6): 80–83.
- Wommack KE, Bhavsar J, et al. (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci.* 6(3):427–39.
- Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, 15, R46.
- Wooley JC, Ye Y (2009) Metagenomics: Facts and Artifacts, and Computational Challenges. *J Comput Sci Technol.* Jan;25(1):71–81.
- Yamashita A, Sekizuka T, Kuroda M (2016) VirusTAP: Viral Genome-Targeted Assembly Pipeline. *Front Microbiol.* Feb 2;7:32. doi: 10.3389/fmicb.2016.00032.

- Yassour M, Vatanen T, Siljander H, Hämäläinen AM, Härkönen T, Ryhänen SJ, Franzosa EA, Vlamakis H, Huttenhower C, Gevers D, Lander ES, Knip M; DIABIMMUNE Study Group, Xavier RJ (2016) Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science Translational Medicine*, Volume 8, issue 343, 343ra81. DOI: 10.1126/scitranslmed.aad0917.
- Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–9.
- Zhao G, Vatanen T, Droit L, Park A, Kostic AD, Poon TW, Vlamakis H, Siljander H, Härkönen T, Hämäläinen AM, Peet A, Tillmann V, Ilonen J, Wang D, Knip M, Xavier RJ, Virgin HW (2017) Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proc Natl Acad Sci USA*. Jul 25;114(30):E6166–E6175. doi: 10.1073/pnas.1706359114.

9 ORIGINAL ARTICLES

SOFTWARE

Open Access

POMO - Plotting *Omics* analysis results for Multiple Organisms

Jake Lin^{1,2,3*}, Richard Kreisberg³, Aleksi Kallio², Aimée M Dudley³, Matti Nykter⁴, Ilya Shmulevich³, Patrick May^{1,3} and Reija Autio^{2*}

Abstract

Background: Systems biology experiments studying different topics and organisms produce thousands of data values across different types of genomic data. Further, data mining analyses are yielding ranked and heterogeneous results and association networks distributed over the entire genome. The visualization of these results is often difficult and standalone web tools allowing for custom inputs and dynamic filtering are limited.

Results: We have developed POMO (<http://pomo.cs.tut.fi>), an interactive web-based application to visually explore *omics* data analysis results and associations in circular, network and grid views. The circular graph represents the chromosome lengths as perimeter segments, as a reference outer ring, such as cytoband for human. The inner arcs between nodes represent the uploaded network. Further, multiple annotation rings, for example depiction of gene copy number changes, can be uploaded as text files and represented as bar, histogram or heatmap rings. POMO has built-in references for human, mouse, nematode, fly, yeast, zebrafish, rice, tomato, *Arabidopsis*, and *Escherichia coli*. In addition, POMO provides custom options that allow integrated plotting of unsupported strains or closely related species associations, such as human and mouse orthologs or two yeast wild types, studied together within a single analysis. The web application also supports interactive label and weight filtering. Every iterative filtered result in POMO can be exported as image file and text file for sharing or direct future input.

Conclusions: The POMO web application is a unique tool for *omics* data analysis, which can be used to visualize and filter the genome-wide networks in the context of chromosomal locations as well as multiple network layouts. With the several illustration and filtering options the tool supports the analysis and visualization of any heterogeneous *omics* data analysis association results for many organisms. POMO is freely available and does not require any installation or registration.

Keywords: *Omics*, Association, Visualization, Ortholog, Phenolog, Genome-wide, Network, Model organism

Background

Modern high-throughput technologies measuring different *omics* types are constantly producing masses of new data [1-3]. Simultaneously, the various analysis algorithms and association analyses methods applied to these measurements are providing many different types of results [2-6]. Thus, the integration of the data and subsequent visualization of these results are becoming increasingly important and challenging [7].

The different types of analysis algorithms are resulting in various types of associations within the data. Often these methods include correlation-based or integrative data mining algorithms [6], and the results can include genomic feature to genomic feature associations across multiple data types, such as gene expression and chromosome rearrangements. The features can, for example, be genes or genomic positions such as regulatory regions, or they can be also clinical or sample annotations resulting for example from differential expression analysis [3,8]. While the different values or types of data are related with each other, it also becomes necessary and challenging to be able to visualize different types of data and the results of their analysis [7,9,10]. Generally, the results of various analyses are given as text lists and visual illustrations are

* Correspondence: jake.lin@uni.lu; reija.autio@tut.fi

¹Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg, Luxembourg

²Department of Signal Processing, Tampere University of Technology, Tampere, Finland

Full list of author information is available at the end of the article

confounded by different formats, software platforms, and dependencies. However, because most of the genomic data can be organized by its genomic location, it is straightforward and advantageous to utilize the genomic position as a parameter in visualization. Since the majority of resulted *omics* associations can be linked to the physical chromosome positions, genome-wide illustrations can provide new insights to the investigator [9].

Traditional genome browsers such as Integrative Genomic Viewer [11], UCSC Genomic Browser [12] and GBrowse [13] are very useful for viewing biological data with multi-scaled linear tracks but they are not ideal to view gene networks. Cytoscape [14] fills this need and is adept at displaying network interactions and has released CytoscapeWeb [15] and Cytoscape.js beta libraries designed for web programming integration. Given that structural rearrangement events are likely more informative in the context of ordered chromosome circular layout context, there are a limited number of software tools available for circular illustration of the genomic association data, of which Circos [16] is most often used. Circos provides command line options to plot various types of data together into assorted attractive but static circular plots. Circos software requires local installation along with several mandatory Perl core and third party modules. The recent introduction of RCircos [17] successfully draws Circos images with R but implies that its usage is limited to experienced R programmers. DNAPlotter [18] plots interactive user-defined circular and linear genomic tracks. This standalone tool, improved from other published genomic viz tools such as CGView [19], GenomeDiagram [20], GenomePlot [21], GenoMap [22] and Microbial Genome Viewer [23] by combining Jemboss [24] and Artemis [25], flexibly accepts custom text files and relational databases, and the plotted tracks can be filtered and exported. DNAPlotter requires installation and does not support associations. Galaxy [26], web-based and very comprehensive for biomedical analysis and sharing, recently introduced Circster [27] a web-based Circos like visualization as part of its comprehensive pipeline. While Galaxy is available both publically and as a local install, Galaxy visualization functions are only available downstream of its workflows and thus limited to its ecosystem. As such, visualizing *omics* data with such a program requires a certain level of computational experience and multiple programs to illustrate, share and filter the data analysis results. In contrast, the UCSC Interaction Browser [28] and WikiPathways [29] both allow for web visualization and organization of network interactions, but they do not have genomic chromosomal context association views and they lack support for several important model organism references. In addition, as *omics* data includes often thousands of feature values, and there are at total thousands to millions resulted associations, it is vital to support filtering options for

exploration and detection of sub-networks from dense and cluttered networks.

To address these issues, we have developed POMO, Plotting *Omics* analysis results for Multiple Organisms. POMO is a free web-based software suite that permits the illustration of associations inferred from *omics* data as filterable circular genome-wide, Cytoscape Web and grid views. Aiming to parallel the diversity of systems biology research, POMO software has built in reference support for human [30] and the following model organisms: mouse [31], zebrafish [32], worm [33], fly [34], rice [35], tomato [36], *Arabidopsis* [37], *S. cerevisiae* [38] and *E. coli* [39] (See Table 1 for resources). In addition, the program accepts parameters for integration and plotting of genomic homologies and orthologous features of multiple strains of the same organism or closely related species. Multiple text file formats are supported, and associations can be directly uploaded or referenced as URL addresses using modern web browsers. POMO supports the plotting of an unlimited number of rings to highlight genomic annotations and regions of interest, and all results remain private and can be exported and shared as SVG image or TSV text files. The web based (<http://pomo.cs.tut.fi>) program is a freely available user-friendly tool for genome-wide biological research that does not require any installation or registration. With the wide selection of data visualization options, POMO is a unique tool for all the researchers working with *omics* data analysis, which can be used, for example, to visualize and filter the genomic networks in the context of chromosomal locations as well as multiple network layouts.

Implementation

It is widely accepted that visual networks are valuable for detecting and exploring patterns in large datasets. Genomic network visualizations with multiple perspectives, particularly within chromosomal context can offer insights of key proximal nodes and possible sub-networks. Data mining algorithms produce genome-wide association sets where individual associations are described with either a numerical ranking or weight. The option to filter and iteratively visualize these large data sets is of key importance in exploring and understanding the genomic associations. Our web application addresses and extends these requirements by combining different data types and including the reference genomes of multiple organisms by utilizing modern web programming technologies and components. POMO allows immediate visualization of genome-wide associations and annotations directly from text files while offering grid, Cytoscape and genomic circular context views. Within the genomic circular context, chromosomes are drawn as segments of the circumference; its length is normalized dependent on the nucleotide base length of the displayed organism. *Omics* nodes,

Table 1 Supported organism references

Organism	Species/build	Source	URL
Human	H. Sapiens (GRCh37.p11)	ENSEMBL	http://www.ensembl.org/Homo_sapiens/Info
Fly	D. melanogaster (BDGP5)	Fly base	http://flybase.org/
Mouse	M. musculus (GRCm38.p1)	MGI	http://www.informatics.jax.org/
Worm	C. elegans (WBcel235)	Worm base	http://wormbase.org/
Yeast	S. cerevisiae (EF4)	SGD	http://www.yeastgenome.org/
Zebra fish	D. rerio (Zv9)	ZFIN	http://www.zfin.org/
Arabidopsis	A. thaliana (TAIR10)	TAIR	http://www.arabidopsis.org/
Rice	O. sativa (MSU6)	MSU	http://rice.plantbiology.msu.edu/
Tomato	S. lycopersicum (SL2.40)	SolGenomics	http://solgenomics.net/
E. Coli	K-12 (MG1655)	Ecocyc	http://ecocyc.org/

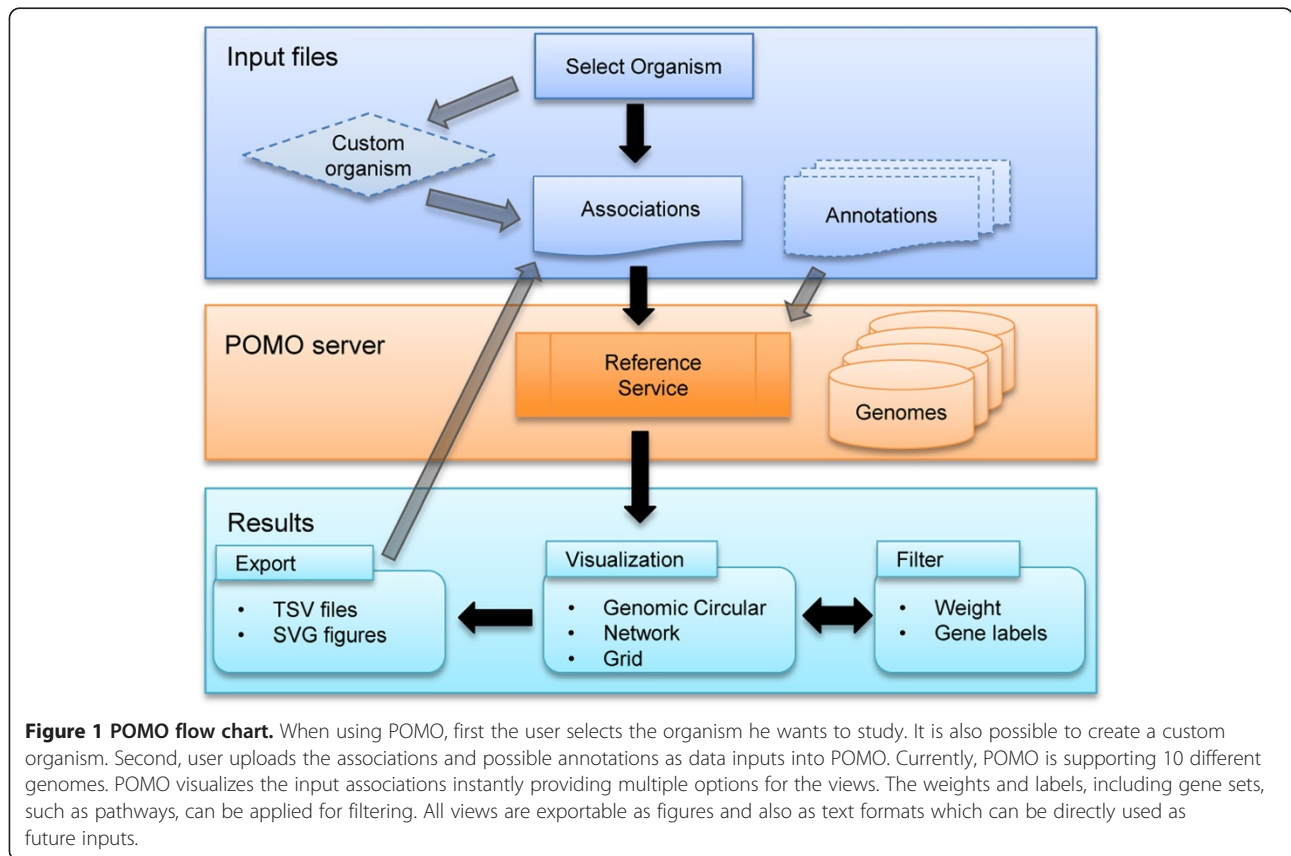
which can be labelled as gene names or ids or explicit genomic positions, will be oriented/mapped to these segments, and the associations are represented as an edge between two genomic locations or genes. For additional visual differentiation, the notations are color encoded for different *omics* data types, such as gene expression, copy number variations, or proteomics data. Multiple annotation rings, with support for bar, histogram and heatmap graphs, can also be appended. Outer glyphs are used for representation of genomic features to unmapped nodes, which have no genomic location, such as phenotypic traits or disease state features.

Many labs studying data originating from *omics* studies of different organisms are lacking the personnel and expertise to write customized software for visualizing genome-wide associations. The inclusion of multiple organisms into POMO addresses this need by enhancing the utility and usability of visualization software. POMO supports the newest genome builds of the following organisms: human, mouse, nematode, fly, yeast, zebrafish, *Arabidopsis*, rice, tomato and *E. coli* (Table 1).

Additionally, POMO provides an interface for a custom/new organism selection. This option allows users to define a new organism, which can be for example an existing organism that POMO does not yet support, parts of an existing organism (chromosomes or contigs), or combination of several species. As outlined in Figure 1, unsupported or custom references can be defined and their associations plotted and exported. In addition, POMO enables pairwise between-organism comparison allowing visualization of in-between associations of genes or genomic locations between different organisms, such as human-mouse or yeast-yeast. The resultant views can be exported as an SVG and converted to publication resolution quality images using free tools like Inkscape. This function will assist labs with communicating and sharing their association findings. The exported filtered text associations can be used as immediate POMO inputs as well.

Further, POMO supports direct URL referencing of associations, such as cloud-based files stored on GoogleDrive or DropBox, and thus researchers can communicate their insights visually with fellow collaborators. POMO does not store any upload data thus preserving and addressing security and privacy.

POMO is designed for illustrating *omics* associations directly from text files in circular genomic, network and tabular contexts with dynamic built in organism reference and annotation support. Following graph syntax from math, an edge is defined as two nodes having a link or association. In POMO, this edge can be ranked with a numeric weight, such as a p-value or correlation, or the user can directly mark this association with a color. Input associations can be derived from any data mining method as long as node labels are either gene names, identifiers such as ENSEMBL and ENTREZ or chromosome based positions. This flexibility allows for network nodes to be in non-coding DNA range which leads to complete inclusivity. Non-gene coding events such as promoter sites, copy number variation and other aberrations can easily be integrated and visualized. The program supports mixing gene and non-gene position based node labels. POMO node labels can be either ENSEMBL/ENTREZ id or gene label or position based. Position based nodes are labelled in the form chr:start:end. The nodes may be enhanced with a source type, such as genotype (GENO), gene expression (GEXP) or proteomics (PROT) data. These optional node annotations are encoded to a set of colors that lead to richer and differentiable graphical details. In addition, POMO supports multiple genome wide annotation rings, where the rings are defined in a text file and then uploaded. The syntax allows for pairing of values or colors to a gene or a segment in the chromosome. Syntax details and examples are provided in the Additional file 1. As exhibited in Additional file 1: Figure S10, annotation rings can be represented as bars, histograms and heat maps. Unmapped (PHENO) phenotype



associations are visually portrait as outer glyph ticks, where the position represents the genomic position linked to the unmapped feature.

POMO inputs are text files containing genomic results such as interactions or associations. Each edge defines two nodes and the nodes are labelled with a gene name or ENSEMBLE or ENTREZ identifiers. The user can mix the node labels freely and Additional file 1: Table S1 provides more details and examples. Edges can optionally be rank with weights and also directly marked up with an HTML supported color. The supported delimiters along with the file type extensions are spaces (.txt), tabs (.tsv) and commas (.csv). Simple Interaction Format (.sif), which allows for multiple associations to be placed on one line, is also supported. We have also extended the sif format to allow an optional weight or color column.

Utilizing HTML5 FileReader API and modern web browsers, the tool allows uploading of association and annotation text files and then upon chromosome position translation immediately plots the resultant graph. Publicly accessible cloud hosted *omics* association files can be read by POMO as an URL parameter. For testing and efficient plotting of small networks, one can declare association edges directly inside the URL parameter. Details and syntaxes are provided in the user guide,

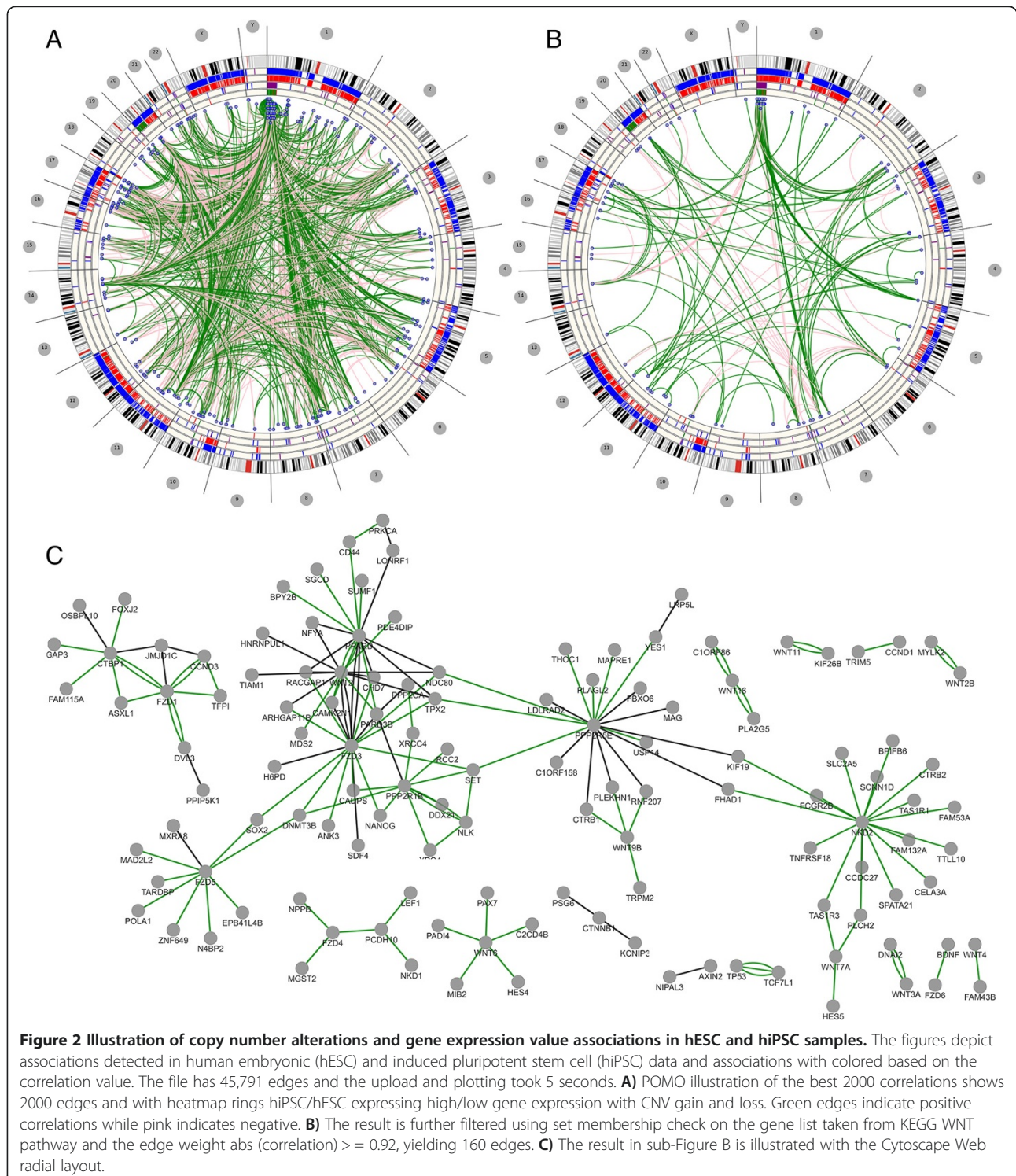
Additional file 1. The software includes comprehensive dialogs and messages to report if certain association node labels cannot be mapped to the selected reference. Association weight filtering can be accomplished if numeric values are provided. Moreover, POMO also allows for label set filtering, meaning, *e.g.*, that a list of gene labels, such as members of a particular pathway, can be used to find subsets of the graph. The circular, grid and network views are automatically refreshed on each filtered submit and their iterated graph images can be exported as SVG image file, suitable for publishing or posters with its high definition presentation.

The interface dialog windows are programmed with custom listeners and AJAX events for seamless dynamic document updates. Since JavaScript allows for functions as parameters, these dynamic functions are then being utilized on callback functions upon user selection of organism and file format selection and upload. Object instantiations are also linked to different user interface selections, such as organism determines the genome browser a node click resolution. This Web 2.0 application includes extensive usage of ExtJS layout and panels, and jQuery AJAX with JSON Objects for data exchange. POMO circular visualization is built on top of VisQuick that utilizes Protovis [40] while the network view

integrates CytoscapeWeb with custom discrete mappers and data population functions. Built on modern web software principles that include integration of python libraries and SQLite databases, the application can be deployed to all major web servers independent of platform and operating system.

Results and discussion

Big data is a large and routine part of modern day genomics research; along with troves of public databases, labs are generating different types of genome-wide data from new experiments and various instruments. Various sets of associations, often heterogeneous, are being extracted and by



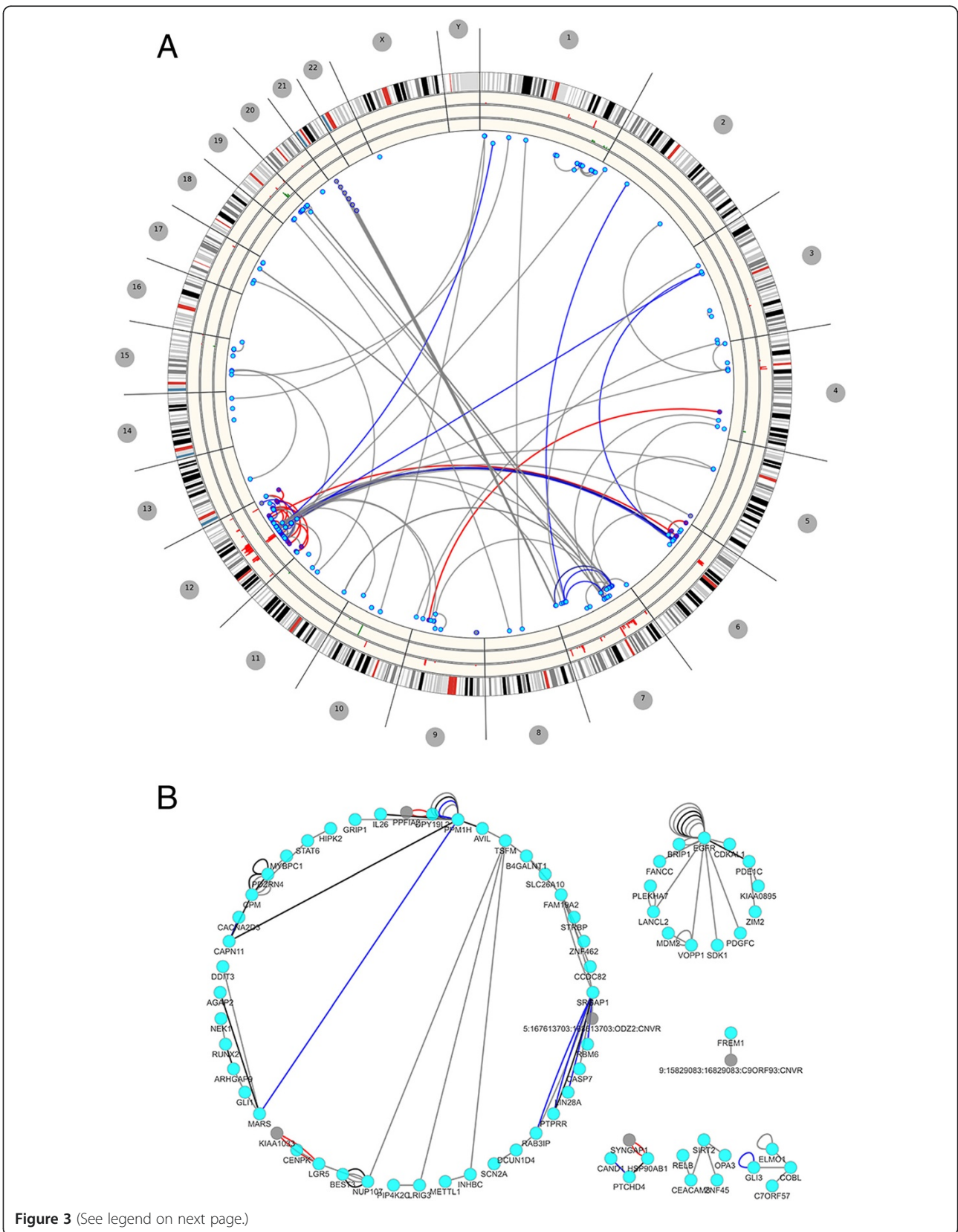


Figure 3 (See legend on next page.)

(See figure on previous page.)

Figure 3 Plotting genomic structural variations. The figures depict TCGA GBM [43,44] rearrangements and chromothripsis findings from whole genome sequencing. **A)** Edge colors are used to describe number of supporting reads, with gray < 50 and blue greater than 100. Histogram rings are depicting copy number gain and loss ratios while the inner most ring accounts for possible gene fusion events. **B)** The result is showing the network view of the same data where the single edge associations are filtered.

using POMO the investigators can gain insights from the different visual perspectives and layouts. Of particular interest is the genomic circular layout, where nodes are spatially mapped to chromosome arcs on the circumference and the associations are represented as edges between the genome-anchored nodes. Proximal and high degree nodes are revealed instantly, as well as sparse disjoint associations. With usage of filtering by association weight with multiple operators, gene label, or list of gene labels that can be for example pathways, investigators can intuitively find insights from previous uninformative dense networks. It is well known that genome wide visualizations, particularly in circular context, can have limited spatial capacities and dense graphs are not informative. To address this, POMO allows for filtering and edge bundling functions. The edge bundling allows for a node range window and groups the edges if the start and end nodes are within this window. Optionally, a score threshold can be set to exclude valued edges from the bundling (See Additional file 1 for more usage details).

POMO can serve as a tool for genome-wide network visual exploration and communicative collaboration since the filtered results can be shared as exported files, images or directly as an URL. Clicking on nodes will open specific Genome Browsers on the selected region window of the specific organism. In the following scenario, POMO is used for integrating and visualizing copy number gains and losses in relation to correlation associations in application of human embryonic stem cells (hESC) and human induced pluripotent stem cells (hiPSC) samples [41,42] (Figure 2). The rings in POMO plot are illustrating the copy number variations together with genes whose expression values have been identified to be associating with the copy number variation. In Figure 2, after the outermost cytoband, the first ring is indicating the areas whose copy number has been altered in hESC samples, while the next ring illustrates the genes whose high expression is associated with gain in copy number (red) and whose low expression is associated with loss in copy number (green) of the same samples [41]. Similarly the fourth ring illustrates the copy number alterations in hiPSC samples [42] and lastly the associated genes with them in the same samples (unpublished observations, Laurila *et al.* submitted). The edges demonstrate correlations between the detected genes computed through all the expression data. Based on the genome-wide figure it is easy to see how there are several genes with copy number alteration in both hiPSC and hESC samples in the chromosome 1, that

are highly correlating with other altered genes and are also a part of WNT pathway.

Genome-wide contexts can be particularly helpful in viewing chromosomal arrangements. Figure 3 depicts TCGA glioblastoma multiforme (GBM) [43] rearrangement and chromothripsis events associated with poor survival [44]. Using data in the accompanied supplement, chromothripsis results are represented as red edges while blue edges demonstrate rearrangements with supporting reads of greater than 100, where grey represents supporting reads of lower than 50. Chromosome region 12q14-15 is considered as a breakpoint-enriched region where oncogenes *CDK4* and *MDM2* are noted to amplify frequently [44]. The inner red ring of the figure demonstrates these elevated amplifications where the next two inner rings represent gains (green) and then genes with evidence involved in fusions.

Another case study is the visualization of high quality yeast protein-protein interactions labelled with ENSEMBL gene ids [45,46]. Released as part of Cytoscape, the file contains 6888 edges and can be directly uploaded into POMO without any data manipulation. A full workflow, including file upload and resolution of chromosome positions using POMO's reference translator service took 1.9 seconds and then 1 second to plot the default but configurable limit of the first 2000 edges [See Additional file 1: Figure S13]. This is consistent with our randomized testing of 1000 edge sets where the genomic translation service performs around 500 milliseconds and then almost instantaneous plotting. See Table 2 for more details on browser/OS comparisons. Though web based software has a dependence on network connectivity, we have successfully tested the service from different locations. For clarity, plot limits can be set easily with a pull down list and filtering, whether it is label set or scoring based, is always applied on the full association set. The actual plotting relies on browser/client memory. Furthermore, the export of filtered associations can serve as inputs on future POMO sessions. The different views are all updated dynamically and synced with the latest uploaded and filtered results. Users can toggle between the tree, circle, radial and force-directed layouts in the Cytoscape Web view.

POMO also allows the user to visualize genomic associations between two related organisms, or two distinct strains within the same POMO supported organism. Figure 4A exhibits phenolog [47] orthologs of obesity-abnormal food intake between human and mouse. Edge

Table 2 Performance benchmarking on yeast protein-protein associations

	Process	Firefox	Chrome
Windows 7 4 GB RAM 2.6 GHz	Upload/server translation	1.5 seconds	1.5 seconds
	Browser plotting	1 second	1 second
Mac OS 10.8 8 GB RAM 1.8 GHz	Upload/server translation	1.5 seconds	1 second
	Browser plotting	1 second	0.5 second

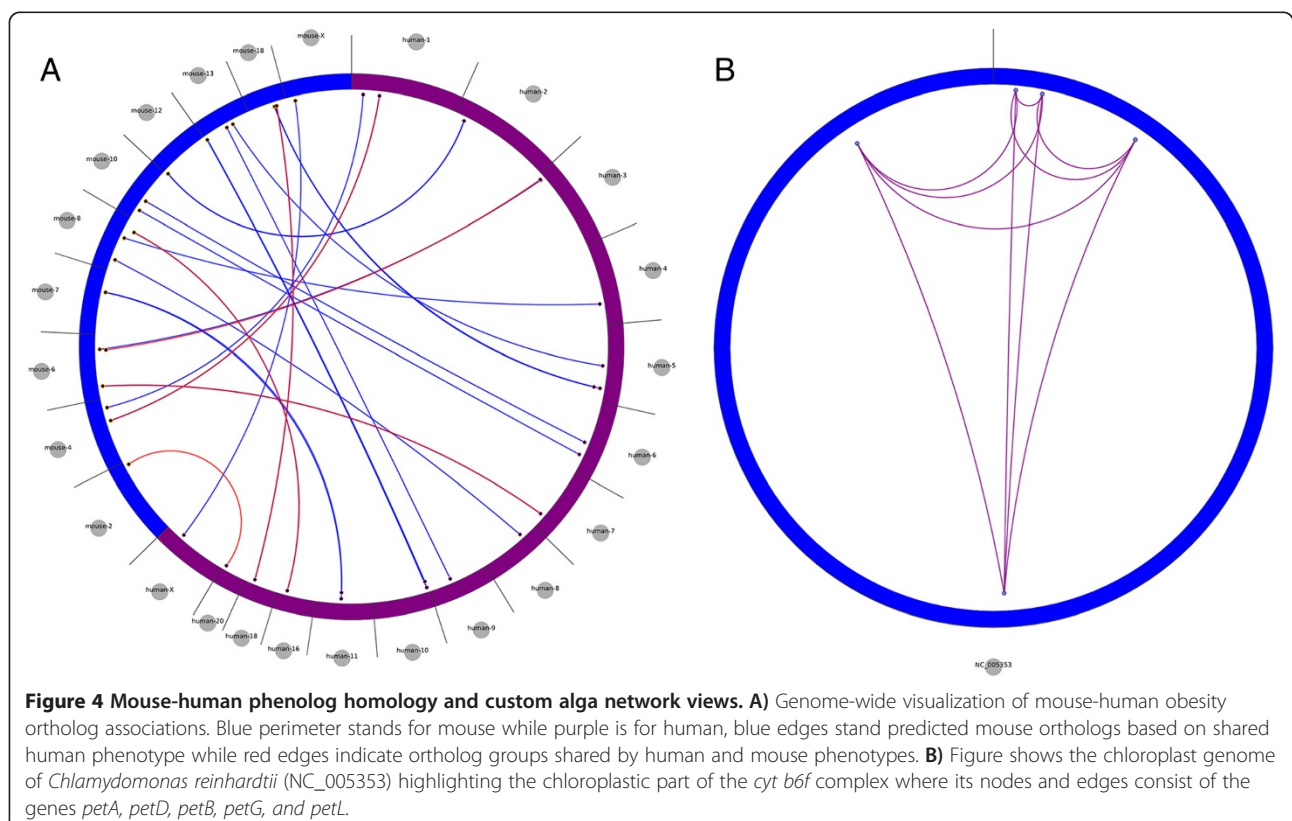
The network consists of 3025 nodes and 6888 edges. The times given here include uploading/translation (genomic identifiers to chromosome positions) and plotting. Generally, Chrome is faster at plotting while uploads will depend on network speed and geographical location. These tests were done in Finland and Germany on wireless connections. We recommend using relative recent releases of Firefox and Chrome because of HTML5 file uploading and JavaScript libraries dependencies.

colors are used to differentiate predicted orthologs and shared orthologs based on observed phenotypes. Using the same interface and selecting custom organism, the user selects the organisms to contrast, and then the input file association node labels are resolved based on the selected reference. Following this workflow, an unsupported organism can be defined by indicating its chromosomes and base lengths. Figure 4B demonstrates the custom function to illustrate the chloroplast genome of the green alga *Chlamydomonas reinhardtii* (NC_005353) [48], highlighting the associations of genes in the *cyt b6f*

complex, which mediates electron transfer between photosystems (PS) II and I, cyclic electron flow around PSI, and state transitions [49]. More information concerning custom organism options is described in detail in the Additional file 1.

Conclusions

POMO, freely available for non-commercial research, was designed for life science researchers to easily plot, filter and share genome-wide *omics* data and associations using an intuitive web interface. In supporting different labs studying different organisms, a comprehensive set of model organism genome references are fully integrated to allow for flexible association notations. The unique property, only available in POMO, is allowing the user to illustrate various organisms or closely related organisms together within single view. POMO also includes a detailed user guide, and several example associations and annotations are provided. In future, we will add support for other further organisms and appreciative of user feedbacks to improve the views and interface. For maximal visual impact, different visualization views and network layouts are supported and can be seamlessly toggled with simple clicks. Upon filtering, each view is dynamically filtered and text exports can serve as future inputs while the SVG image export can be converted to publishing quality



presentations. POMO is an open sourced project and the code, builds and documentations are available at <http://pomo.googlecode.com>. In sum, as genome-wide visualizations, particularly interactive and web based, can help researchers to confirm theories and formulate new research questions, POMO can significantly facilitate researchers in finding new biological discoveries among their *omics* data.

Availability and requirements

Project name: POMO: Plotting *Omics* analysis results for Multiple Organisms

Project home page: <http://pomo.cs.tut.fi>

Operating system(s): Platform independent

Programming language: Python 2.6+, JavaScript, HTML5, SQLite 3.7+

License: POMO is available free of charge to academic and non-profit institutions.

Any restrictions to use by non-academics: Please contact authors for commercial use.

Additional file

Additional file 1: POMO User Guide.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JL, PM and RA conceptualized and initiate the project. JL, RK, PM and RA designed POMO that JL and AK implemented. RK designed and implemented VisQuick. JL designed, implemented and populated the reference databases and translation service stack. JL, RA, and PM drafted the paper. AD, MN and IS contributed important ideas and advices. RA, PM and IS supervised the project. All authors read and approved the final manuscript.

Acknowledgement

This work was supported by a strategic partnership between the ISB and the University of Luxembourg. The work of RA has been funded by Academy of Finland Finnish Programme no 134117 and 135257. The work of PM has been funded by "le plan Technologies de la Santé par le Gouvernement du Grand-Duché de Luxembourg" through Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg.

Author details

¹Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg, Luxembourg. ²Department of Signal Processing, Tampere University of Technology, Tampere, Finland. ³Institute for Systems Biology, Seattle, USA. ⁴Institute of Biomedical Technology, University of Tampere, Tampere, Finland.

Received: 19 September 2013 Accepted: 18 December 2013

Published: 24 December 2013

References

1. Kircher M, Kelso J: High-throughput DNA sequencing – concepts and limitations. *Bioessays* 2010, **32**:524–536.
2. Schatz MC, Langmead B, Salzberg SL: Cloud computing and the DNA data race. *Nat Biotechnol* 2010, **28**:691–693.
3. Berger B, Peng J, Singh M: Computational solutions for omics data. *Nat Rev Genet* 2013, **14**:333–346.
4. Palsson B, Zengler K: The challenges of integrating multi-omic data sets. *Nat Chem Biol* 2010, **6**:787–789.
5. Kirwan GM, Johansson E, Kleemann R, Verheij ER, Wheelock AM, Goto S, Trygg J, Wheelock CE: Building multivariate systems biology models. *Anal Chem* 2012, **84**:7064–7071.
6. Liu Y, Devescovi V, Chen S, Nardini C: Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC Syst Biol* 2013, **7**:14.
7. Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T: Visualizing genomes: techniques and challenges. *Nat Methods* 2010, **7**(3 Suppl):S5–S15.
8. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M: Mapping complex disease traits with global gene expression. *Nat Rev Genet* 2009, **10**:184–194.
9. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, Gavin AC: Visualization of omics data for systems biology. *Nat Methods* 2010, **7**(3 Suppl):S56–S68.
10. Theocharidis A, van Dongen S, Enright AJ, Freeman TC: Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nat Protoc* 2009, **4**:1535–1550.
11. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: Integrative genomics viewer. *Nat Biotechnol* 2011, **29**:24–26.
12. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: The Human Genome Browser at UCSC. *Genome Res* 2002, **12**:996–1006.
13. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: The generic genome browser: a building block for a model organism system database. *Genome Res* 2002, **12**:1599–1610.
14. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campillo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang P-L, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, et al: Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2007, **2**:2366–2382.
15. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD: Cytoscape Web: an interactive web-based network browser. *Bioinformatics Oxf Engl* 2010, **26**:2347–2348.
16. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: Circos: an information aesthetic for comparative genomics. *Genome Res* 2009, **19**:1639–1645.
17. Zhang H, Meltzer P, Davis S: RCircos: an R package for Circos 2D track plots. *BMC Bioinforma* 2013, **14**:244.
18. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J: DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* 2009, **25**:119–120.
19. Stothard P, Wishart DS: Circular genome visualization and exploration using CGView. *Bioinformatics* 2005, **21**:537–539.
20. Pritchard L, White JA, Birch PRJ, Toth IK: GenomeDiagram: a python package for the visualization of large-scale genomic data. *Bioinformatics* 2006, **22**:616–617.
21. Gibson R, Smith DR: Genome visualization made fast and simple. *Bioinformatics* 2003, **19**:1449–1450.
22. Sato N, Ehira S: GenoMap, a circular genome data viewer. *Bioinformatics* 2003, **19**:1583–1584.
23. Kerkhoven R, van Enckevort FHJ, Boekhorst J, Molenaar D, Siezen RJ: Visualization for genomics: the microbial genome viewer. *Bioinformatics* 2004, **20**:1812–1814.
24. Carver TJ, Mullan LJ: JAE: JemBoss Alignment Editor. *Appl Bioinformatics* 2005, **4**:151–154.
25. Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, Parkhill J, Rajandream M-A: Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* 2008, **24**:2672–2676.
26. Goecks J, Nekrutenko A, Taylor J, Team TG: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010, **11**:R86.
27. Goecks J, Eberhard C, Too T, Team TG, Nekrutenko A, Taylor J: Web-based visual analysis for high-throughput genomics. *BMC Genomics* 2013, **14**:397.
28. Wong CK, Vaske CJ, Ng S, Sanborn JZ, Benz SC, Haussler D, Stuart JM: The UCSC interaction browser: multidimensional data views in pathway context. *Nucleic Acids Res* 2013, **41**:W218–W224.
29. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR: WikiPathways: building research communities on biological pathways. *Nucleic Acids Res* 2012, **40**(Database issue):D1301–D1307.

30. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, García-Girón C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kähäri AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, et al: **Ensembl 2013**. *Nucleic Acids Res* 2013, **41**:D48–D55.
31. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, the Mouse Genome Database Group: **The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse**. *Nucleic Acids Res* 2012, **40**:D881–D886.
32. Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, Frazer K, Haendel M, Howe DG, Mani P, Ramachandran S, Schaper K, Segerdell E, Song P, Sprunger B, Taylor S, Van Slyke CE, Westerfield M: **The Zebrafish Information Network: the zebrafish model organism database**. *Nucleic Acids Res* 2006, **34**(suppl 1):D581–D585.
33. Chen N, Harris TW, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Canaran P, Chan J, Chen C-K, Chen WJ, Cunningham F, Davis P, Kenny E, Kishore R, Lawson D, Lee R, Muller H-M, Nakamura C, Pai S, Ozersky P, Petcherski A, Rogers A, Sabo A, Schwarz EM, Van Auken K, Wang Q, Durbin R, Spieth J, Sternberg PW, et al: **WormBase: a comprehensive data resource for Caenorhabditis biology and genomics**. *Nucleic Acids Res* 2005, **33**(suppl 1):D383–D389.
34. Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelts VB, Wilson RJ, the FlyBase consortium: **FlyBase: improvements to the bibliography**. *Nucleic Acids Res* 2013, **41**:D751–D757.
35. Kawahara Y, de la Bastide M, Hamilton J, Kanamori H, McCombie WR, Ouyang S, Schwartz D, Tanaka T, Wu J, Zhou S, Childs K, Davidson R, Lin H, Quesada-Ocampo L, Vaillancourt B, Sakai H, Lee SS, Kim J, Numa H, Itoh T, Buell CR, Matsumoto T: **Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data**. *Rice* 2013, **6**:4.
36. Tomato Genome Consortium: **The tomato genome sequence provides insights into fleshy fruit evolution**. *Nature* 2012, **485**:635–641.
37. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E: **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools**. *Nucleic Acids Res* 2012, **40**:D1202–D1210.
38. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simson M, Weng S, Wong ED: **Saccharomyces Genome Database: the genomics resource of budding yeast**. *Nucleic Acids Res* 2012, **40**:D700–D705.
39. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, Fulcher C, Huerta AM, Kothari A, Krummenacker M, Latendresse M, Muniz-Rascado L, Ong Q, Paley S, Schroeder I, Shearer AG, Subhraveti P, Travers M, Weerasinghe D, Weiss V, Collado-Vides J, Gunsalus RP, Paulsen I, Karp PD: **EcoCyc: fusing model organism databases with systems biology**. *Nucleic Acids Res* 2013, **41**(Database issue):D605–D612.
40. Bostock M, Heer J: **Protovis: a graphical toolkit for visualization**. *IEEE Trans Vis Comput Graph* 2009, **15**:1121–1128.
41. Närvä E, Autio R, Rahkonen N, Kong L, Harrison N, Kitsberg D, Borghese L, Itskovitz-Eldor J, Rasool O, Dvorak P, Hovatta O, Otonkoski T, Tuuri T, Cui W, Brustle O, Baker D, Maltby E, Moore HD, Benvenisty N, Andrews PW, Yli-Harja O, Lahesmaa R: **High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity**. *Nat Biotechnol* 2010, **28**:371–377.
42. Hussein SM, Batada NN, Vuoristo S, Ching RW, Autio R, Narva E, Ng S, Sourour M, Hamalainen R, Olsson C, Lundin K, Mikkola M, Trokovic R, Peitz M, Brustle O, Bazett-Jones DP, Alitalo K, Lahesmaa R, Nagy A, Otonkoski T: **Copy number variation and selection during reprogramming to pluripotency**. *Nature* 2011, **471**:58–62.
43. Network CGA: **Comprehensive molecular characterization of human colon and rectal cancer**. *Nature* 2012, **487**:330–337.
44. Zheng S, Fu J, Vegesna R, Mao Y, Heathcock LE, Torres-Garcia W, Ezhilarasan R, Wang S, McKenna A, Chin L, Brennan CW, Yung WKA, Weinstein JN, Aldape KD, Sulman EP, Chen K, Koul D, Verhaak RGW: **A survey of intragenic breakpoints in glioblastoma identifies a distinct subset associated with poor survival**. *Genes Dev* 2013, **27**:1462–1472.
45. Von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**:399–403.
46. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne J-B, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in *Saccharomyces cerevisiae***. *Science* 2002, **298**:799–804.
47. McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM: **Systematic discovery of nonobvious human disease models through orthologous phenotypes**. *Proc Natl Acad Sci USA* 2010, **107**:6544–6549.
48. Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB: **The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats**. *Plant Cell* 2002, **14**:2659–2679.
49. May P, Christian J-O, Kempa S, Walthert D: **ChlamyCyc: an integrative systems biology database and web-portal for *Chlamydomonas reinhardtii***. *BMC Genomics* 2009, **10**:209.

doi:10.1186/1471-2164-14-918

Cite this article as: Lin et al.: POMO - Plotting Omics analysis results for Multiple Organisms. *BMC Genomics* 2013 **14**:918.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



SOFTWARE

Open Access



Vipie: web pipeline for parallel characterization of viral populations from multiple NGS samples

Jake Lin¹ , Lenka Kramna², Reija Autio³, Heikki Hyöty^{1,4*}, Matti Nykter^{1*} and Ondrej Cinek^{2*}

Abstract

Background: Next generation sequencing (NGS) technology allows laboratories to investigate virome composition in clinical and environmental samples in a culture-independent way. There is a need for bioinformatic tools capable of parallel processing of virome sequencing data by exactly identical methods: this is especially important in studies of multifactorial diseases, or in parallel comparison of laboratory protocols.

Results: We have developed a web-based application allowing direct upload of sequences from multiple virome samples using custom parameters. The samples are then processed in parallel using an identical protocol, and can be easily reanalyzed. The pipeline performs de-novo assembly, taxonomic classification of viruses as well as sample analyses based on user-defined grouping categories. Tables of virus abundance are produced from cross-validation by remapping the sequencing reads to a union of all observed reference viruses. In addition, read sets and reports are created after processing unmapped reads against known human and bacterial ribosome references. Secured interactive results are dynamically plotted with population and diversity charts, clustered heatmaps and a sortable and searchable abundance table.

Conclusions: The Vipie web application is a unique tool for multi-sample metagenomic analysis of viral data, producing searchable hits tables, interactive population maps, alpha diversity measures and clustered heatmaps that are grouped in applicable custom sample categories. Known references such as human genome and bacterial ribosomal genes are optionally removed from unmapped ('dark matter') reads. Secured results are accessible and shareable on modern browsers. Vipie is a freely available web-based tool whose code is open source.

Keywords: *Metagenomics, Viromes, Virus, Assembly, NGS analysis, Visualization, Parallel processing, Viral dark matter*

Background

The use of virome metagenomics has been growing rapidly due to the increasing demands to study the whole virome in clinical samples and to evaluate the evolution of viral quasispecies during acute and chronic infections. The application of virome sequencing techniques become useful not only in infectious disease research, but also in association studies of primarily non-infectious conditions, i.e. in diseases where the

agent is presumed to modify the risk of the disease, which effect is detectable upon investigation of a large number of subjects only. These applications require an approximation of virus quantity, similar to what has long been utilized in bacteriome profiling.

As viruses lack a common sequence signature, metagenomics sequencing of random viral libraries remains the only feasible way of an unbiased assessment of the whole virome. Presently, the need for accurate quantification and interpretation of viral population metrics across a set of samples creates a substantial challenge for this kind of metagenomics studies. Prime obstacles for virome investigators are the large genetic heterogeneity and also that the majority of bioinformatic tools are command line based and overtly technical, being computationally demanding, with complicated dependencies,

* Correspondence: heikki.hyoty@uta.fi; matti.nykter@uta.fi; ondrej.cinek@lfmotol.cuni.cz

¹BioMediTech and Faculty of Medicine and Life Sciences, University of Tampere, PB 100FI-33014 Tampere, Finland

²Department of Pediatrics, 2nd Faculty of Medicine, Charles University and University Hospital Motol, V Úvalu 84, 150 06 Praha 5, Czech Republic

Full list of author information is available at the end of the article



and producing text based outputs that are not easily interpretable [1–5]. Recently released web based applications Taxonomer [6], VirusTAP [7], Virome [8] and Metavir [9, 10] have addressed some of the issues (especially those of user interaction), but mostly operate only on single sample experiments with different workflows. Requiring local dependencies and installation, ViromeScan [11] and MetaShot [12] works on multiple samples. Some of these tools were designed for long (>300) reads or assembled contigs [8–10], which is limiting as modern metagenomics projects including Human Microbiome Project (HMP) [1, 2] produce mostly high-throughput short paired reads. Table 1 provides an overview of the primary features and strategies of these different tools, including our work.

We aimed to open the possibility of creating a table of viral quantities of multiple samples assessed in parallel by exactly identical processes. Here we introduce Vipie, a web based viral diversity population tool accepting as input a set of files from virome metagenomics NGS analyses of multiple samples. Here we present the workflow and results using NGS samples from Human Microbiome Project and other metagenomics studies. Functional on all modern browsers, the high performance pipeline is freely available for academic usage.

Implementation

Our pipeline processes de-multiplexed paired FASTQ files, the most typical product of metagenomics sequencing. Several steps are then performed in parallel for all samples: quality control (QC), de-novo assembly of putative genomic contigs, taxonomic classification of the assembled contigs and orphan singleton reads by performing Blast queries against a local custom virus database derived from Genbank, and finally remapping of the sequencing reads onto reference sequences identified by this taxonomic classification. Default analysis parameters can be easily modified (e.g. the QC stringency, or the de novo assembly algorithm).

Depicted in Fig. 1, Vipie pipeline uses multi processor architecture with integration of PostgreSQL for performance and data management while providing secured interactive results and allowing web form parameters for QC, assembly and scoring. The individual parameters and its default values are listed in the user guide. Trimming and quality control are parameter based applying Galaxy project utilities [13, 14]. We have integrated leading de-novo assembly tools - Velvet [15], MetaVelvet [16], IDBA [17] and MEGAHIT (SOAPDENOV0) [18] and ABySS [19]; these methods and tools are further described and reviewed [5, 20–22]. Taxonomic identification is performed using BLAST [23] against a local NCBI database restricted to whole virus genomes. The final step of the parallel analysis remaps the raw reads

using BWA [24] onto a list of best matches from the BLAST queries, and lists the count of original reads matching to each of these references. In cases where reads match equally well to multiple viruses, the score is divided among such best matches to express importantly the ambiguity in assignation of the motifs shared among viral taxa, and the uncertainty of the presently available classification.

De-novo contigs and reads that do not match to any currently known virus, optionally filtered for human genome and known ribosomal DNA, can be retrieved for further analysis as this ‘dark matter’ of the virome presumably containing novel viruses. Our pipeline allows a direct export of these unmapped reads owing to three-step filtering strategy. Reads unmatched to known viruses are first deprived of sequences that match to ribosomal DNA of bacterial, archeal and fungal origin. This is performed by remapping the reads by the BWA program to databases of 16S, 23S and 5S rDNA (a copy of ftp.ncbi.nlm.nih.gov/genomes/TARGET, and a reduced database of 5S rDNA <http://www.combio.pl/rrna/>) [25]. The next step remaps the reduced set of reads to the human genome. This step yields the potential dark matter of the human genome, mixed with a small proportion of bacterial genomic DNA. Our pipeline does not filter out these bacterial genomic reads, as they may contain novel lysogenic (dormant) phages.

VIPIE’s reference virus database was built from three sources and clustering the sequences to the 97% level of identity further reduced the complexity. First, all viruses were downloaded from the *refseq* database at the NCBI (<https://ftp.ncbi.nih.gov/refseq/release/viral/>), and reduced to 97% identity by using the CD-HIT program (<https://github.com/weizhongli/cdhit/>) [26]. Then, all virus sequences labeled as “complete”, with the “txid10239” (superkingdom Viruses) in the “Orgn” field were retrieved from Genbank. The query retrieved approximately 80,000 sequences from the database, which were subsequently reduced to the 97% similarity by using the CD-HIT program. Finally, similarly to previous two databases, phages were merged and clustered from the European Bioinformatics Institute (EBI) repository (ftp.ebi.ac.uk/pub/databases/fastfiles/embl_genomes/genomes/Phage/).

The web form, interface dialogs and results are programmed to HTML5 standards and using JavaScript and modern, open source JavaScript libraries (<https://jquery.org>, <https://datatables.net>) for browser compatibility. Biopython [27] is used for sequencing parsing and formatting. Parallel processing is achieved via python (<https://www.python.org>) subprocess module implementation and uses PostgreSQL (<https://www.postgresql.org>) schema for job tracking and results merging. Standard SMTP library is used for notification, hence the email registration requirement. Clustered heatmaps are implemented

Table 1 Comparison of the existing virome pipelines tools

Pipeline Tool	Vipie	VirusTAP [8]	Virome [16]	Metavir [14]	Taxonomer [6]	MetaShot [12]
Primary goal	Parallel analysis of multiple viral metagenomes from web and suited for molecular epidemiology studies.	Identification of viruses in a sample, after a thorough elimination of known non-viral sequences.	Classification of all putative ORF found in a viral metagenome, characterization of viral communities.	Analysis of virome, diversity metrics and marker gene phylogenies.	Ultra fast metagenomics analysis focusing on detection of microorganisms, including virus and bacterial.	Highly accurate and comprehensive workflow for host-associate microbiome classification on multiple samples.
Web based	Yes.	Yes.	Yes (Flash required).	Yes.	Yes.	No.
Outputs	Interactive table, plots and raw downloads. Clustered heatmaps with dynamic group assignment re-plots.	Contig based hits and seamless web BLAST interface.	Rich collection of sample source virome ORF and sequence categories.	Comparative analysis of viromes and annotations including networks, nonmetric distance and tree maps.	Interactive pie charts with kingdoms in bins and also impressive sunburst flare sub classifiers.	A Krona graph and Interactive Taxonomy HTML table along with csv file.
Source data	Paired-end reads; <i>fastq</i> format.	Paired-end reads. Accepts also single-end reads; <i>fastq</i> format.	<i>sff</i> , or <i>fastq</i> ; intended for the 454-generated metagenomes.	Reads (>300 bases) or assembled contigs.	Paired-end reads in <i>fastq</i> and <i>fasta</i> formats.	Paired-end reads in <i>fastq</i> format.
Trimming and filtering	YES, as the first step.	YES, as the first step.	YES; quality based; duplicate filtering; contamination	Not specified.	Not specified.	YES, as the first step.
De-novo assembly	YES, a choice of assemblers.	YES, a choice of assemblers; done after subtraction steps.	No.	No.	No.	No.
Subtraction of human ref. and bacterial ribosomal sequences	Optional, only for the output of dark matter sequences.	YES, also other host databases available (mouse etc.). No for ribosomal.	Not specified for human. Ribosome is removed using BLAST against rDNA db.	Not specified.	Not subtracted but reported as part of detection.	Yes, reports identification of human host reads and bacterial mappings.
Means of virus identification	(a) BLAST against a pan-viral database. (b) Remapping of original reads to the identified candidates.	BLAST search against the NCBI nt database.	Protein BLASTP upon two databases. Several tiers of classification of the ORFs.	Not specified.	Taxonomer Binner DB with 21 bp kmers unique identifiers to known viruses.	Custom similarity workflow with hamming distance.
Virome database for identification	A custom database containing 20759 human, animal, plant and bacterial viruses.	Eukaryotic viruses only. Four custom databases available for download.	UniRef 100 peptide database, five annotated protein databases, MetaGenomes On-line.	GAAS tool (https://sourceforge.net/projects/gaas/).	Binner DB needs to be built using kAnalyze [42] (https://sourceforge.net/projects/kanalyze/files/).	TANGO [43] and NCBI Taxonomy [44].
Action when a read maps to different viruses	Score is split among the hit reference sequences.	Not specified.	Not specified.	Not specified.	Assigns as ambiguous.	Parsed for human endogenous retrovirus otherwise classify as ambiguous and discarded.

Most tools use BLAST [23] for initial detection of known references. Vipie uniquely allows web parallel analysis of multi-samples and accounts read hits to multiple viral references for comprehensive population profiling

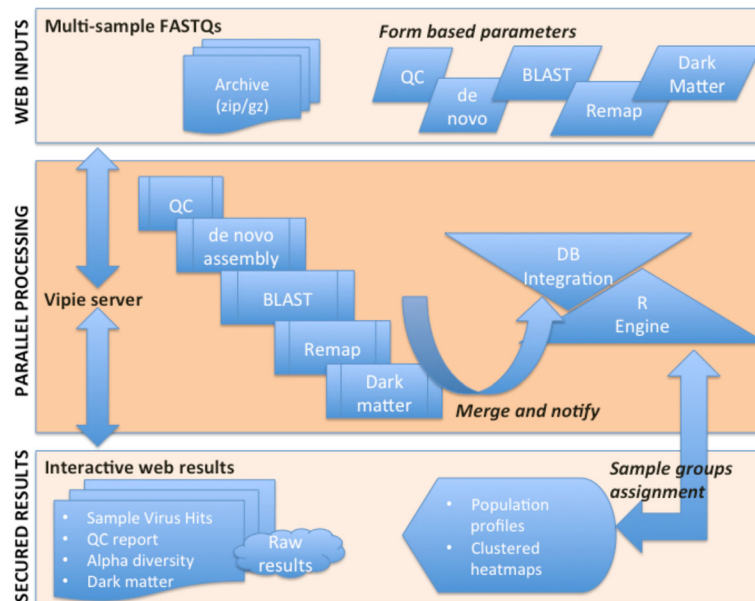


Fig. 1 Vipie web flow chart. For efficiency, sample based paired FASTQ files are uploaded as a zipped archive with optional mapping file. Illumina BaseSpace archive downloads can be used without changes. All pipeline parameters can be entered using the web form. The default values and use case are listed in the user guide available at home page along with example multi-sample archive input

with R ggplot2 [28] while other summary and alpha diversity statistics are computed using custom python scripts. Population maps and read distribution count summary charts are created using highcharts.js (<https://www.highcharts.com>) and custom event handlers for interactivity. Vipie is an ongoing open sourced project and available at <https://sourceforge.net/projects/vipie>.

Results

Input samples and interactive results

The pipeline utility is here demonstrated on set of 11 samples where the input and results are available to all users. The sample set consists of (a) blood, nasal, stool and vagina data from Human Metagenome Project (HMP), (b) diarrhea sample from gastroenteritis outbreak (DRA004165 DNA Data Bank Japan [29, 30]) used in VirusTAP and (c) stool data from in-house ongoing African metagenomics project [31, 32]. Table 2 lists relevant accession identifiers, sources and number of reads along with result links. As the compressed archived exceeds 1.2 gigabytes, a smaller subsampled archive consisting of 20% is available for download on the homepage and the original compressed FASTQ archived is available on <https://sourceforge.net/projects/vipie/files/data> [33]. End-to-end processing of the 11 samples took 82 min, processing 29,778,980 reads that includes assembly, scoring, and clustering and removal of human reference and known ribosomal references. The performance time was measured after the archive was uploaded as file upload depends fully on local network

speed. The interactive results, with population profile maps and filterable viral hit tables are accessible at: <https://binf.uta.fi/vipie/results.html?key=eLZPuObVoU>. Result links are accessible without registration and designed to be shared among collaborators whereas job history and active jobs are visible only to registered investigators. The results are divided into panels of Population profile & group assignment, QC & Dark matter report, Summary & alpha diversity, and Viral hits table. Raw results, including unmapped dark matter reads that do not match to any known virus can be also downloaded.

Figure 2 shows group-based population pie charts and alpha diversity as measured by Shannon entropy [34]. The population pie chart sizes are relative to total number of hits and their slices are fully interactive as clicking on the slices traverses the taxonomy levels. The tool found 167 unique accessions across the samples and an easy to use searchable and sortable sample hits table is provided and best experienced from the browser, where the table can be collapsed based on taxonomy and sample viral hits can be downloaded as a text file ready for Excel import.

Our user guide provides screenshots and directions on filtering the sample hits table and using the filtering function, we found Human Herpes hits on a HMP blood sample SRS072276, where herpes in hematological samples have been reported in a prior microbiome and hematopoiesis report [35]. Our results showed that virus population profiles are unique across body sites, reported also in ViromeScan and visually shown

Table 2 NGS samples used in Vipie validation from Human Microbiome Project, Africa study, and diarrhea sample sourced in Japan gastroenteritis outbreak. ViromeScan listed 20 HMP samples but only Stool types of 4 samples passed QC

AccessionId	Source	Sample Type	Number of Reads ^a	Sample used in Vipie-ViromeScan-VirusTAP validation	Vipie Results ^b
SRS072276	HMP	Blood	438,879	Yes-No-No	1,2
SRS072318	HMP	Blood	753,994	Yes-No-No	1,2
SRS019033	HMP	Retroauricular	1,285,003	Yes-No-No	1
SRS016944	HMP	Retroauricular	1,619,439	Yes-No-No	1
SRS012902	HMP	Stool	2,039,473	Yes-Yes-No	1
SRS014923	HMP	Stool	2,009,179	Yes-Yes-No	1
SRS014466	HMP	Vagina	367,077	Yes-No-No	1,2
SRS015072	HMP	Vagina	495,256	Yes-No-No	1,2
SRS072313	HMP	Nasal	320,672	Yes-No-No	2
SRS072261	HMP	Nasal	367,384	Yes-No-No	2
SRS072366	HMP	Nasal	114,414	Yes-No-No	2
S11	Africa	Stool	1,634,821	Yes-No-No	2
S12	Africa	Stool	1,191,427	Yes-No-No	2
S14	Africa	Stool	1,143,784	Yes-No-No	2
DRA004165	Japan	Diarrheal	1,108,688	Yes-No-Yes	2

In addition to those stool samples, Vipie test archive includes 4 other HMP sample types. Result links with performance time are also provided

^aInput archive of Result 2 samples (subsampling 20% 225 MB) available at: https://binf.uta.fi/vipie/data/vipie_archive_ssampling.zip

^bResults 1: <https://binf.uta.fi/vipie/results.html?key=2HSPXukkDS> (66 min)

Results 2: <https://binf.uta.fi/vipie/results.html?key=eLZPuObVoU> (82 min)

in the clustered maps. Interestingly, in the stool sample SRS012902, crAssphage [36] was by far the highest virus detected. Figure 3 shows the clustered heatmap generated in R, and it correctly clustered healthy HMP sample types together [11] while Japanese gastroenteritis and African samples showed profoundly different signatures.

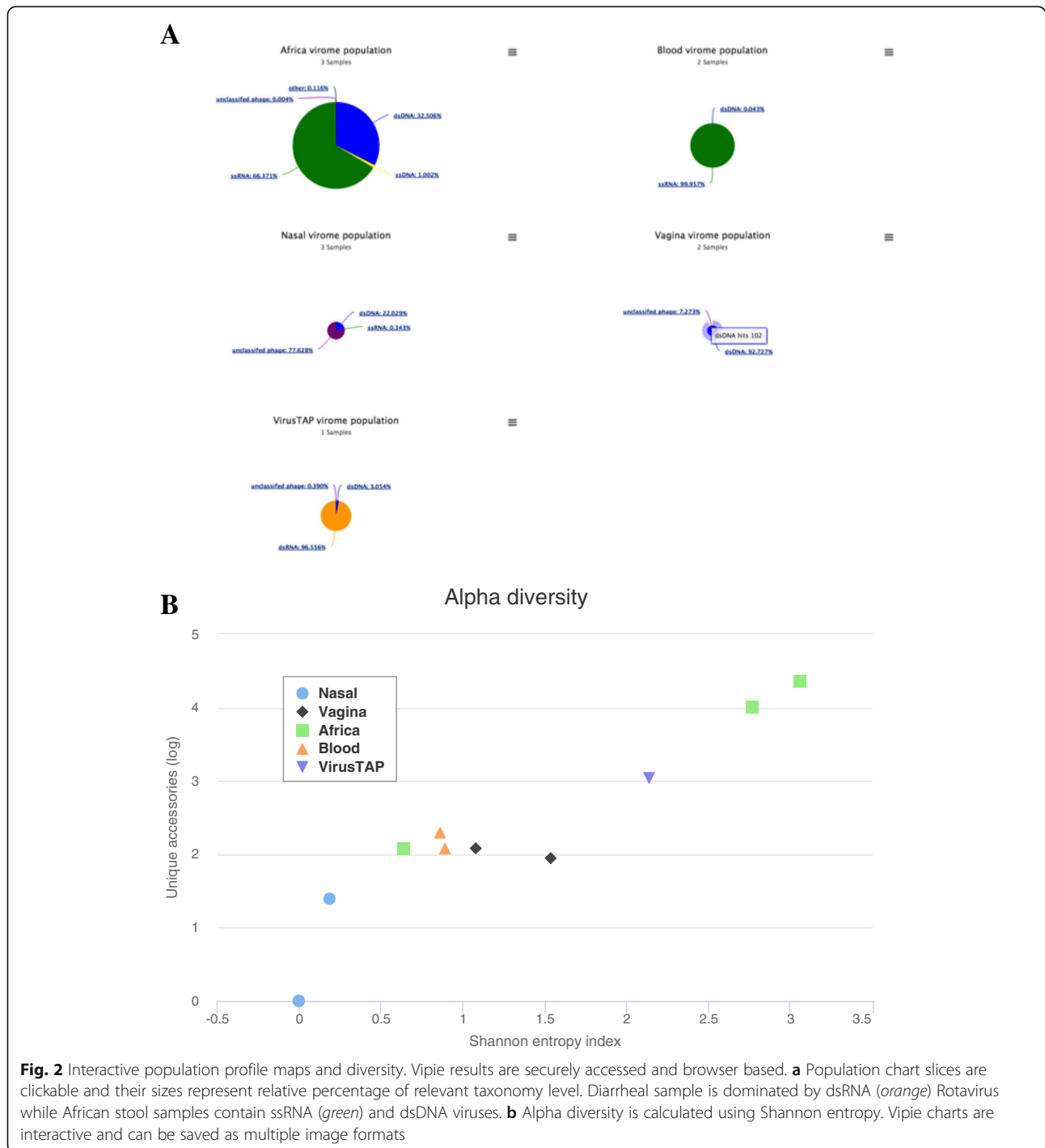
Comparisons

We first compared our performance to that of ViromeScan. While ViromeScan states that it supports multiple samples, it requires local installation with 50+ gigabytes of database requirements. The 20 HMP samples used for its validation, only the stool samples passed QC [37] and likely due to timing, the other sample types were not available on HMP download page. Our summary and cluster findings of stool samples and retroauricular, with the highest diversity, samples agree with ViromeScan and other HMP findings of ~5.5 genera per sample [38]. We were unable to reproduce the herpes associations reported with vagina samples as those samples are no longer available. Input parameters, interactive maps, QC report (Fig. 4a) and viral hits of the 11 samples are accessible at <https://binf.uta.fi/vipie/results.html?key=eLZPuObVoU> and Table 2 contains accession ids along with sample read sizes.

Then performance of Vipie was compared to VirusTAP. Its web based de novo assembly dedicated pipeline required 17 min to process the DRA004165 sample from a study of gastroenteritis [29] in Japan. VirusTAP capably

detected 11 Human rotaviruses where this result is cited and also available as its example results. Vipie using the same input detected similar findings of 14 Human rotaviruses strains (shown in Additional file 1: User guide Figure 10B) and also interestingly *Streptococcus* phage strains. Using the same sample, our pipeline required 32 min due to post assembly remapping with custom scoring and then unmapped origin filtering. Because of Vipie's parallel computing design, the archive of 11 samples and more than 10 times the amount of reads, took just 82 min. The more comprehensive findings also highlight the scoring split strategy on read hits on multiple viruses and investigation of unmapped viral read origins shown in Fig. 4b.

Furthermore, benchmarking was assessed and compared with the recently published MetaShot, using its simulated artificial dataset with a very high share of human sequences mixed with low amounts of many different viral sequences. Table 3 below shows the similar precision and recall results of the two tools. Vipie has a slightly higher percentage of unclassified viral reads likely due to subsampling of the initial dataset, and due to the fact that we optimized the virus BLAST database by removing sequences that were less distant than 3% from its closest relative; similar reduction of taxonomic complexity is known from e.g. bacteriome profiling. The script and Vipie results used for computing this statistics are available with README in Vipie project page on SourceForge. We are grateful to MetaShot authors for permission to use their simulated data, constructed using ART [39].



Discussion

Vipe interface is implemented with HTML5 standards and utilizes open source JavaScript libraries. Unlike older and Adobe Flash based applications, Vipe does not require additional installations and supports all modern HTML5 compliant browsers while offering a consistent user experience. The input parameter form is designed to be clean and to group into processed components

where each element has custom validation rules. The component details and rules are listed in the user guide. Secured and interactive analysis results are accessed with encrypted links and to promote collaboration, can be shared without registration. Sample based alpha diversity is provided, using Shannon entropy index [34] (Fig. 2) as a representative of diversity methods [35]. Vipe intuitively offers web based, form or file upload sample group

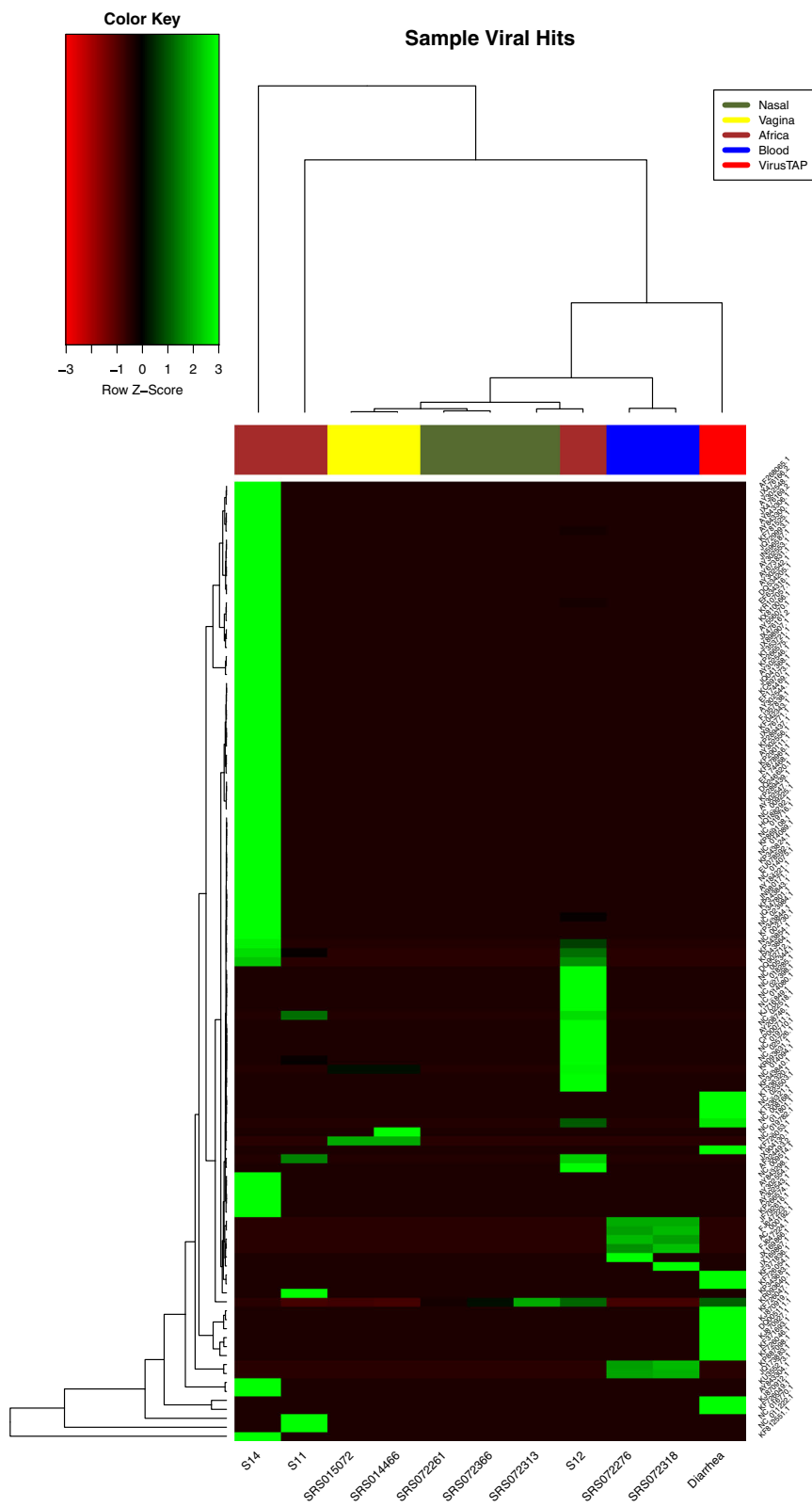


Fig. 3 (See legend on next page.)

(See figure on previous page.)

Fig. 3 Clustered heatmap of HMP, African and Japanese diarrheal samples. Public NGS data from different consortiums provide opportunities for advanced comparative virome analysis. Healthy HMP sample types clustered correctly (nasal, vaginal, blood samples) while a Japanese sample (gastroenteritis dataset from the VirusTAP report) and African samples (known to be positive for multiple viruses) showed different signatures. HMP samples can be identified using the legend on upper right, with *olive green* for nasal, *yellow* for vagina and *blue* for blood. Samples from rural Africa and VirusTAP (Japan) are marked in *colors brick and red*

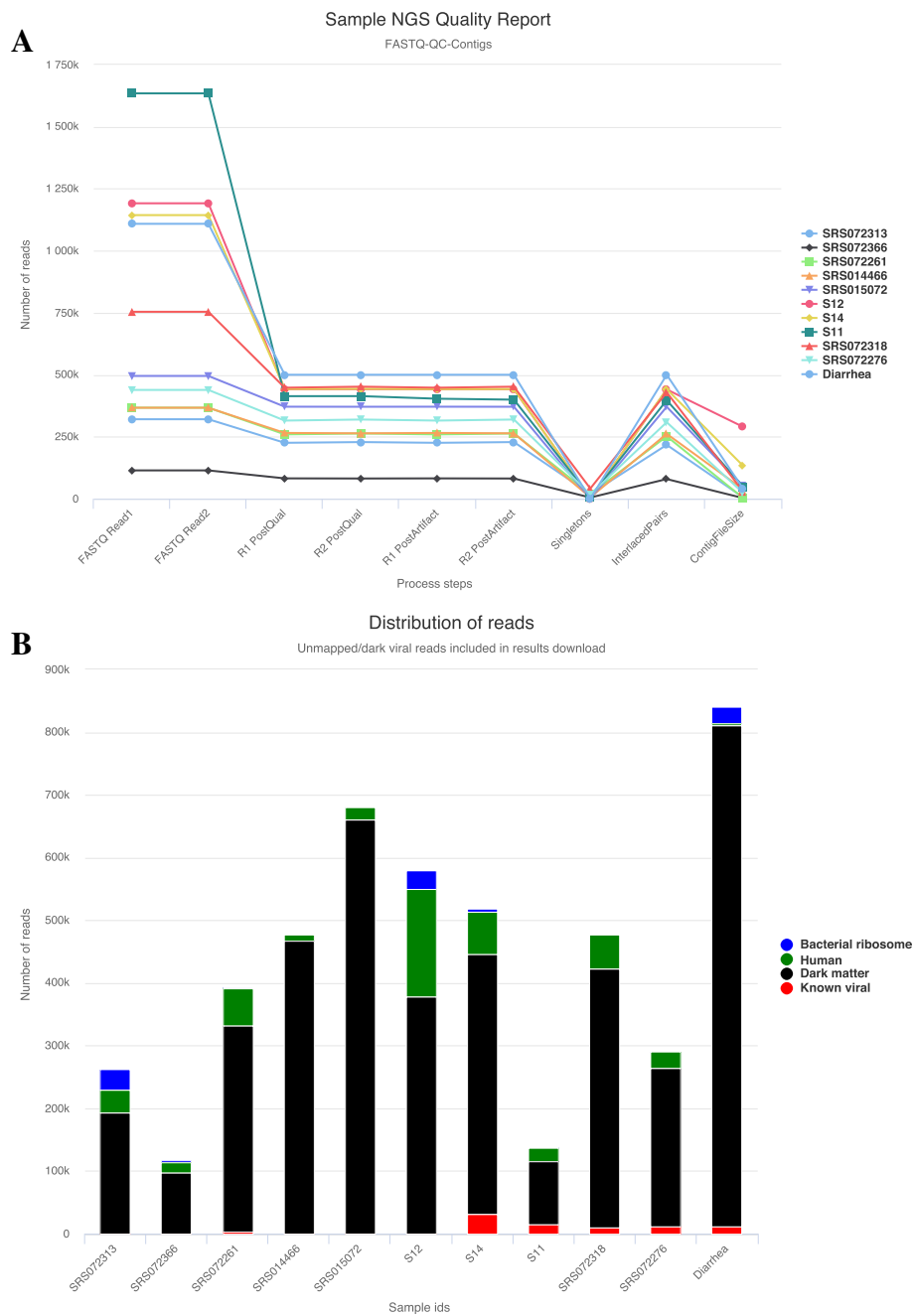


Fig. 4 QC and distribution of reads including dark viral matter. **a** The chart shows the number of NGS reads retained per sample through QC, interlacing and de novo assembly. **b** Sample reads, along the x-axis and their aligned origins are shown as stacked bars. Shown in *black*, unmapped viral 'dark matter' is of high interest across virology studies. *Blue* bars represent bacterial ribosome, *green* for human while *red* is for known viral matches

Table 3 (A) Read assignment benchmark assessment of MetaShot and Vipie on simulated dataset^a consisting of 19 582 500 human (94.5%), 986 114 bacterial (4.8%) and 146 886 viral (0.7%) reads. Vipie percentages are based on random subsampling of 1 000 000 reads and bacterial statistics are not reported as Vipie reports information on bacterial ribosome only (the bacterial genomic DNA is not filtered out, as it might lead to loss of dormant phage sequences). (B) Precision, Recall and F-measure are calculated on the same data. Input reads and assessment script are available on SourceForge^b

A	Assigned % ^c		Correctly Assigned % ^d	
	MetaShot	Vipie	MetaShot	Vipie
Human (host)	99.18	99.27	99.99	99.27
Viruses				
Family	97.74	99.98	98.53	93.39
Genus	97.39	98.99	99.75	93.33
Species	97.81	93.66	96.70	92.97
B	Human (host)		Virus	
	MetaShot	Vipie	MetaShot	Vipie
Precision (%)	100.00	100.00	98.30	96.85
Recall (%)	99.97	99.96	98.19	95.36
F-measure (%)	100.00	99.98	98.07	96.08
Unclassified (%)	1.04	0.73	3.94	6.73

^a<https://recascloud.ba.infn.it/index.php/s/nw4s9hqnf8QkBsK>

^b<https://sourceforge.net/projects/vipie/files/validation/k>

^cThe percentage refers to the total number of reads assignable to the specific taxonomic rank

^dThe percentage refers to the relevant assigned reads

reassignment where population and clustered maps are reanalyzed and dynamically redrawn. The pipeline produces a cross tabulation similar to the operational taxonomic unit (OTU) tables from bacteriome profiling, additional statistics is doable with advance R packages such as phyloseq [40] and deseq2 [41].

Often, published pipelines emphasize that their performance is by orders of magnitude faster than existing strategies [7, 8] and that the tasks can be completed in the order of minutes to single hours in a situation where existing viruses account only for a minor fraction of the total read count. We believe that the present Vipie pipeline offers fast data processing for most relevant applications, including real-time assessment of viral repertoire in clinical samples. For comparison, VirusTAP processing, up to assembly with 1 sample (~2 million reads, 172 MBs) took 17 min (Input upload time is not included as it is dependent completely on local network speed.). Vipie process the same sample in 32 min including assembly, cross validation scoring/remapping, known reference filtering and viral dark matter processing. Parallel implementation is ideal for multi-sample processing and input set of 11 samples (Table 2), consisting of ~30 million reads, 1.22 GBs compressed and processed in 82 min. There is no concurrent limit on the

number of samples eligible for processing other than a small database overhead. Job completion time has a direct relationship to the sample with the highest read depth and it is well known that interlacing and assembly are high memory tasks. The de novo assembly step implements random subsampling on user defined read percentage, default of 75% with a maximum of 1,000,000 NGS reads per sample. Very large archives can suffer from network timeouts on file upload. In overcoming this scenario, we have successfully deployed Vipie on cluster computing environment and analyze thousands of samples consisting of terabytes of data using SLURM, the default utility for Linux high performance computing. We believe that our strategy offers a good balance between bearable algorithm speed on most machines, and availability of multiple sample processing.

Importantly, the pipeline offers a set of files with bacterial, human, and unknown sequences (the “dark matter” of the virome). Dark matter reads are the remaining unmapped reads after filtering for human and bacterial ribosomes. It has been long known that the unknown dark matter is extremely valuable in virome analysis [9] and in focus with the recent discovery of new bacteriophage virus *crAssphage* while its bacterial host still unknown [36]. Many components of this “dark matter” of the virome have been observed across studies, and are likely to represent existing viruses, yet their taxonomy is presently unknown. The lack of taxonomic classification however should not preclude their use as provisional entities, exposures that are testable and quantifiable in epidemiological studies. Figure 4b shows an interactive sample based chart consisting of stacked bars representing the percentage of reads mapped to human, bacterial ribosomes, known viruses and dark matter. It is apparent that these unmapped reads dominated these NGS samples and deeper advanced analyses are necessary. As such, viral dark matter raw reads are part of downloads.

An often-overlooked aspect is the uncertainty in virus identification. The Genbank database contains many similar isolates of almost every relevant virus serotype. This means that most reads or contigs would map to multiple different sequenced virus isolates. In single sample studies this does not pose any problem - the taxonomy is concluded as the highest scoring hit, or the first of a set of similarly high scoring organisms. This however cannot be done when a pipeline processes multiple samples at the same time: due to the known intrinsic variability of the viruses, even a single subject may produce two different samples where different virus quasi-species may prevail that will preferentially map to two different virus reference sequences. There are two possible solutions to the problem: the ViromeScan pipeline employed one where the databases are smaller with a limited scope. Unfortunately, the strategy towards their

construction was not described in the paper, but clearly only the most important serotypes represent each virus species - e.g. only 92 sequences cover the whole repertoire of human DNA viruses. In Vipie we chose a different strategy: we decided to build a representative virus database of all available sequences (clustered to a 97% similarity level for the sake of algorithm speed), and all multiple equally likely mapping hits are resolved by splitting the mapping score among the different hits. At higher taxonomic levels of family or genus this is not visible, but when descending to the level below species (to individual reference sequences), the uncertainty is expressed by the existence of a whole block of candidate viral reference sequences to which the sample distributes many of its reads. This should express that the found virus is similar to many references, but neither is fully identical. This strategy has proven feasible in our benchmarking experiment when we reached parameters reasonably close to the specialized single-sample taxonomy tool MetaShot [12], while offering the possibility of parallel assessment of multiple viromes in one run. We assigned 3.73% less reads to their correct species (MetaShot 96.70%, VIPIE 92.97%) - this may be (a) the effect of clustering our representative virus database; some reads falling into species or serotype specific viral regions may thus remain unidentified; (b) the consequence of subsampling - VIPIE uses subsampling to 1 million reads maximum, whereas the simulated MetaShot data set is more than 20 times larger, with most of the viruses in trace amounts.

Conclusions

Virome NGS datasets are unique in several aspects. Firstly, unlike in amplicon libraries in bacteriome profiling, there are no clearly outlined methods of taxonomic classification and of quantification of the viral agents. Secondly, unlike work on e.g. RNA sequencing in humans and animals, there is no well-defined reference set of viral sequences. Therefore the virome characterization must rely on an insufficient knowledge of existing viruses, and on still uncertain techniques of taxonomic sorting - first because the taxonomy of viruses is still rapidly evolving.

When studying an association of existing or novel viral agents with a condition (as is a disease, an ecological variable, or a human intervention), it is imperative to keep the analytical conditions identical across the data set, and to attempt a truly unbiased relative quantification of the viral agents present therein. This can be safely achieved only if all samples of the dataset are processed by an identical protocol - and if they are quantified against a common set of reference sequences. The reference set should be a union of all possible references of the whole study set. Our pipeline performs such quantification: it identifies all agents present in the

dataset and in the final step it attempts remapping of the original reads from every sample to this whole reference set. This enables employing the ensuing virus quantity tables in downstream analyses similarly to the well-established analyses of bacterial profiles from 16S rDNA mass sequencing.

Availability and requirements

Project name: Vipie: web pipeline for parallel characterization of viral population from multiple NGS samples

Project home page: <https://binf.uta.fi/vipie>

Source code: <https://sourceforge.net/projects/vipie>

Operating system(s): Platform independent

Programming language: Python 2.7+, R 3.3, JavaScript, HTML5, PostgreSQL 9+

License: Vipie is available free of charge to academic and non-profit institutions.

Any restrictions to use by non-academics: Please contact authors for commercial use.

Additional file

Additional file 1: Vipie User Guide. (DOCX 3189 kb)

Abbreviations

HMP: Human microbiome project; NGS: Next generation sequencing; OTU: Operational taxonomic unit; QC: Quality control

Acknowledgements

We like to thank the authors of MetaShot and VirusTAP for assisting with validation and access to their test archive. In addition, we are grateful to Dr Per Ashorn for usage of Malawian virome samples.

Funding

The work has been supported by the University of Tampere's BioMediTech Doctoral School, National technology Agency in Finland and Ministry of Health of the Czech Republic, AZV 15-31426A.

Authors' contributions

OC, JL and HH conceptualized and initiate the project. JL, LK and OC designed Vipie that JL and OC implemented. JL, MN, HH and OC drafted the paper. RA contributed important ideas on R engine and statistical integration. OC, MN and HH supervised the project. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Human virome samples from HMP and VirusTAP projects have been previously published and adhered fully to the principles of the Declaration of Helsinki. The unpublished African metagenomic virome samples used for validation (randomly selected) come from a Malawian population study comprised of healthy 6-month-old rural infants. The trial adhered to the principles of the Declaration of Helsinki. Written informed consent was obtained from the mothers of all participants and the trial protocol was reviewed and approved by the College of Medicine research and ethics committee (University of Malawi) and the ethical committee of the Pirkanmaa Hospital District (Finland). The Malawian clinical trial is registered at ClinicalTrials.gov with identifier of NCT0052446.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹BioMediTech and Faculty of Medicine and Life Sciences, University of Tampere, PB 100FI-33014 Tampere, Finland. ²Department of Pediatrics, 2nd Faculty of Medicine, Charles University and University Hospital Motol, V Úvalu 84, 150 06 Praha 5, Czech Republic. ³School of Social Sciences, University of Tampere, Kalevantie 4, 33100 Tampere, Finland. ⁴Fimlab Laboratories, Pirkanmaa Hospital District, Tampere, Finland.

Received: 28 January 2017 Accepted: 25 April 2017

Published online: 15 May 2017

References

- The Human Microbiome Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486(7402):207–14.
- Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JL. The human microbiome project. *Nature*. 2007;449(7164):804–10.
- Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. *Nat Rev Microbiol*. 2017;15(3):183–92. doi:10.1038/nrmicro.2016.182.
- Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet*. 2005;6:805–14. doi:10.1038/nrg1709.
- Shapton TJ. An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci*. 2014;5:209.
- Flygare S, Simon K, et al. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol*. 2016;201617:111.
- Yamashita A, et al. VirusTAP: viral genome-targeted assembly pipeline. *Front Microbiol*. 2016;7:32.
- Wommack KE, Bhavsar J, et al. VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci*. 2012;6(3):427–39.
- Roux S, Faubladier M, et al. Metavir: a web server dedicated to virome analysis. *Bioinformatics*. 2011;27(21):3074–5.
- Roux S, et al. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinf*. 2014;15:76.
- Rampelli S, Soverini M, et al. ViromeScan: a new tool for metagenomic viral community profiling. *BMC Genomics*. 2016;17:165.
- Fosso B, et al. MetaShot: an accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data. *Bioinform*. 2017. doi: 10.1093/bioinformatics/btx036.
- Afgan E, Taylor J, Anton Nekrutenko A, Goecks J, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*. 2016;44(W1):W3–W10. doi:10.1093/nar/gkw343.
- Blankenberg D, the Galaxy Team, Taylor J, Nekrutenko A, et al. Dissemination of scientific software with galaxy ToolShed. *Genome Biol*. 2014;15:403. doi:10.1186/gb4161.
- Zerbina DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 2008;18:821–9.
- Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet : An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res*. 2012;40(20):e155.
- Peng Y, et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*. 2013;28:1420–1.
- Li D, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015;31(10):1674–6.
- Simpson K, et al. ABySS: A parallel assembler for short read sequence data. *Genome Res*. 2009;19(6):1117–23. doi:10.1101/gr.089532.108.
- Paszkiwicz K, Studholme DJ. De novo assembly of short sequence reads. *Brief Bioinform*. 2010;11(5):457–72. doi:10.1093/bib/bbq020.
- Tritt A, Eisen JA, Facciotti MT, Darling AE. An integrated pipeline for de novo assembly of microbial genomes. *PLoS One*. 2012;7(9):e42304. doi:10.1371/journal.pone.0042304.
- Li Y, et al. VIP: an integrated pipeline for metagenomics of virus identification and discovery. *Sci Rep*. 2016;6:23774. doi:10.1038/srep23774.
- Altschul SF, et al. Basic local alignment search tool. *J Mol Biol*. 1990;215:403.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–60.
- Szymanski M, Zielezinski A, et al. 5SRNAdb: an information resource for 5S ribosomal RNAs. *Nucleic Acids Res*. 2016;44(D1):D180–3.
- Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460–1.
- Cock PA, Antao T, Chang JT, Bradman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25:1422–3.
- Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2009.
- Kimura H, et al. A food-borne outbreak of gastroenteritis due to genotype G1P[8] rotavirus among adolescents in Japan. *Microbiol Immunol*. 2014; 58(9):536–9. doi:10.1111/1348-0421.12176.
- DNA Data bank of Japan [http://getentry.ddbj.nig.ac.jp/\(DRA004165\)](http://getentry.ddbj.nig.ac.jp/(DRA004165)) Accessed 01 Dec 2016.
- Rodríguez-Díaz J, et al. Presence of human enteric viruses in the stools of healthy Malawian 6-month-old infants. *J Pediatr Gastroenterol Nutr*. 2014; 58(4):502–4. doi:10.1097/MPG.0000000000000215.
- Mangani C, et al. Effect of complementary feeding with lipid-based nutrient supplements and corn-soy blend on the incidence of stunting and linear growth among 6- to 18-month-old infants and children in rural Malawi. *Matern Child Nutr*. 2015;11 Suppl 4:132–43. doi:10.1111/mcn.12068.
- Vipie project SourceForge <https://sourceforge.net/projects/vipie/files/data/> Accessed 15 Mar 2017
- Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27:379–423 and 623–656.
- Simpson EH. Measurement of diversity. *Nature*. 1949;163:688. doi:10.1038/163688a0.
- Dutilh BE, Edwards RA, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun*. 2014;5:4498. doi:10.1038/ncomms5498.
- NIH Human Microbiome Project website. <http://www.hmpdacc.org/HMASM/HMASM-690.csv>. Accessed 01 Jan 2017
- Wylie KM, Mihindukulasuriya KA, Zhou Y, Sodergren E, Storch GA, Weinstock GM. Metagenomic analysis of double-stranded DNA viruses in healthy adults. *BMC Biol*. 2014;12:71.
- Huang W, et al. ART: a next-generation sequencing read simulator. *Bioinformatics*. 2012;28:593–4.
- McMurdie PJ, Holmes S. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*. 2013;8(4): e61217. <http://dx.doi.org/10.1371/journal.pone.0061217>.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550. <http://dx.doi.org/10.1186/s13059-014-0550-8>.
- Audano P, Vannberg F. KAnalyze: a fast versatile pipelined k-mer toolkit. *Bioinformatics*. 2014;30:2070–2.
- Alonso-Alemany D, et al. Further steps in TANGO: improved taxonomic assignment in metagenomics. *Bioinformatics*. 2014;30(1):17–23.
- Sayers EW, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2009;37(Database issue):D5–15.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

