Bioinformatic Methods for Metagenomics and Comparative Genetics in Veterinary Medicine

Martin Norling

Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics
Uppsala

Doctoral Thesis
Swedish University of Agricultural Sciences
Uppsala 2014

Acta Universitatis agriculturae Sueciae
2014:73

Cover: Cape Buffalo in the Masaai Mara National Reserve, Kenya
(photo: M. Norling)

ISSN 1652-6880
ISBN (print version) 978-91-576-8094-5
ISBN (electronic version) 978-91-576-8095-2

Print: SLU Service/Repro, Uppsala 2014

Bioinformatic Methods for Metagenomics and Comparative Genetics in Veterinary Medicine

Abstract

Good science includes innovation, investigation, and rigor. This thesis' first study is related to rigor. This study was performed at the International Livestock Research Institute (ILRI) in Nairobi, as part of the Arbovirus Incidence and Diversity (AVID) project. A field sample recording system was developed, which saves time and location metadata from the global positioning system (GPS), as well as a connected system monitoring the biobank, freezer, incubators and servers. This monitoring system more than once prevented loss of resources due to freezer failure by alerting responsible personnel, and was later published in 'Biopreservation and Biobanking'.

The sampling system was re-used in the second study, a Ugandan project aimed to identify African swine fever (ASF) in pigs. In this study Ndumu virus, a relatively unstudied virus previously only found in culicine mosquitoes, was discovered in domestic pigs.

For the third study, collaboration with ILRI continued with a study analyzing the 'Muguga Cocktail', the live vaccine currently used to control *Theileria parva*, a protozoan parasite causing East Coast Fever (ECF) in cattle. Live vaccines have many problems, such as high costs, difficult manufacturing, and the risk that misused vaccine will spread the disease. An in-depth study of the three parasite stocks included in the vaccine was performed, where genomic differences were identified with the goal of explaining the success of the vaccine, as well as identify a potential set of antigens which may in the future replace the live vaccine with a subunit vaccine.

Finally, for the fourth study, the metagenomic theme continued with the development of the MetLab, a tool for experimental design and analysis for viral metagenomics projects. The tool consists of three parts: (i) tools to estimate the sequencing needs of a metagenomic project, (ii) simulation tools, allowing users to simulate metagenomics sequencing data, and (iii) the system runs metagenomic analysis pipelines.

Keywords: bioinformatics, metagenomics, virology, infection biology, next generation sequencing

Author's address: Martin Norling, SLU, Department of Animal Breeding and Genetics,
P.O. Box 7023, 750 07 Uppsala, Sweden
E-mail: Martin.Norling@slu.se

Dedication

To all friends and family who waited, while I spent my time working instead of spending it with them.

A lot of my internal conflict and malaise comes from the tension between the life I ACTUALLY want to live, and the stories I'd love to be able to tell

Dinosaur Comics

Contents

List of Publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I Charles Masembe, George Michuki, Maria Onyango, Cecilia Rumberia, **Martin Norling**, Richard P Bishop, Appolinaire Djikeng, Stephen J Kemp, Alan Orth, Robert A Skilton, Karl Ståhl, and Anne Fischer (2012). Viral metagenomics demonstrates that domestic pigs are a potential reservoir for Ndumu virus. *Virology Journal* vol 9(11).

II **Martin Norling**, Absolomon Kihara, and Steve Kemp (2013). Web-Based Biobank System Infrastructure Monitoring Using Python, Perl, and PHP. *Biopreservation and Biobanking* vol 11(6), 355-358.

III **Martin Norling**, Richard P. Bishop, Roger Pelle, Weihong Qi, Sonal Henson, Kyle Tretina, David Odongo, Stephen Mwaura, Thomas Njoroge, Erik Bongcam-Rudloff Claudia A. Daubenberger, Joana C. Silva. Whole Genome Comparisons of the Three Major Stocks Comprising the Live Sporozoite Theileria parva Vaccine. (Manuscript)

IV **Martin Norling**, Oskar Karlsson, Erik Bongcam-Rudloff, Juliette Hayer. MetLab: An *In Silico* Experimental Design, Simulation and Validation Tool for Viral Metagenomics studies. (Manuscript)

Papers I-II are reproduced with the permission of the publishers.

Additional Publications

I Katarina Truvé, Oscar Eriksson, **Martin Norling**, Maria Wilbe, Evan Mauceli, Kerstin Lindblad-Toh, Erik Bongcam-Rudloff (2011) SEQscoring: a tool to facilitate the interpretation of data generated with next generation sequencing technologies. *EMBnet.journal*, vol 17(1)

II Harald Mischak, Walter Kolch, Michalis Aivaliotis, David Bouyssié, Magali Court, Hassan Dihazi, Gry H. Dihazi, Julia Franke, Jérôme Garin, Anne Gonzalez de Peredo, Alexander Iphöfer, Lothar Jänsch, Chrystelle Lacroix, Manousos Makridakis, Christophe Masselon, Jochen Metzger, Bernard Monsarrat, Michal Mrug, **Martin Norling**, Jan Novak, Andreas Pich, Andrew Pitt, Erik Bongcam-Rudloff, Justyna Siwy, Hitoshi Suzuki, Visith Thongboonkerd, Li-Shun Wang, Jérôme Zoidakis, Petra Zürbig, Joost P. Schanstra and, Antonia Vlahou (2010) Comprehensive human urine standards for comparability and standardization in clinical proteome analysis. *PROTEOMICS - Clinical Applications, vol 4(4)*

The contribution of Martin Norling to the papers included in this thesis was as follows:

I Bioinformatic analysis scripts.

II Design, evaluation and programming of the system, as well as writing part of the manuscript.

III Sequence assembly, analysis, and writing part of the manuscript.

IV Design, scripting and programming of the system, and writing part of the manuscript.

Abbreviations

Arbovirus	Arthropod borne virus
ASFV	African swine fever virus
CDS	Coding sequence
ECF	East Coast Fever
GPS	Global Positioning System
ILRI	International Livestock Research Institute
indel	Small insertion or deletion
ITM	Infection and treatment immunization
NDUV	Ndumu virus
NGS	Next-Generation Sequencing
PCR	Polymerase Chain Reaction
SNP	Single nucleotide polymorphism
TTV	Torque teno virus

1 Introduction

1.1 Investigation, Innovation, and Rigor

Science is often viewed as driven by genius and insight, where scientists formulate brilliant hypothesis, validate them by testing, and by those means gain insight into completely new knowledge. In truth though, while the basis of scientific generalization of perceived patterns are clearly investigation and hypothesis testing, in close second place comes innovation and rigor.

Innovation is natural when it comes to science. When investigating at the edge of human knowledge, there is an obvious need to develop new methods and improve old ones, both as the target of scientific research, and as a necessity for moving research forward.

Proper application of rigor is what ensures the reliability of conclusions. It aims to prevent erroneous information from being accepted due to chance outcomes. Many applications of this reliability are thoroughly embedded in both the scientific method in general, which stresses the importance of repeatability, and in the statistical applications commonly applied in research.

Remembering the importance of these concepts is itself important, as the appeal of scientific discovery can be very strong, especially in highly competitive research areas. Proper applications of rigor allow scientists to avoid reliance on new and exciting, but potentially dubious, information. In research, it is equally important to prevent unverified or erroneous information from obscuring or corrupting results, as it is to contribute with new information. This allows scientists to rely on previous findings, only occasionally needing to reproduce previous results, which is in many ways what allows research to progress efficiently.

1.2 Monitoring Infrastructure

Experimental biological science is highly dependent on environmental samples. Retrieving samples is often a difficult and expensive task though, and can be dependent on things like disease outbreaks, environmental conditions, etc. The problem of getting proper samples for a planned study means that samples may be retrieved ‘when possible’, and then stored until the study can be planned and afforded.

Storage of samples also poses a number of problems though. While sample storage is possible for extended periods, in for example liquid nitrogen, other problems appear over time. Conclusions drawn from experiments on a given sample is only as good as the knowledge and reliability of the sample. What if someone accidentally left a freezer door slightly open for a night, allowing some samples to thaw and re-freeze once he door was closed? What if the papers describing the sampling protocol were lost when moving to a new lab a few years after the sampling?

This problem can be addressed by using proper storage procedures, both of the physical samples and of their associated meta-data, as well as proper monitoring procedures to ensure that no change happens without getting logged and compensated for (De Souza & Greenspan, 2013).

1.2.1 Sampling Systems

Maintaining the integrity of samples in proper storage is important, but even more value can be added by ensuring that as much additional external information, ‘metadata’, as possible is collected during sampling. As the person who does the sampling may not be the same person that designs the experiment and draws the conclusions, collecting as much data as possible might be crucial to ensure maximum value from the sampling, as they cannot know exactly what information might be needed.

A simple way of getting more information from samples, without making the sampling itself more complicated, is to use an automated system to log samples. Such a system should automatically save as much data as possible, such as position and time data from the Global Positioning System (GPS). Preferably, this system should also be integrated with the sample storage and logging systems to have as few steps as possible where errors could be introduced, as well as making sure that the storage systems handles the sample information in an optimal manner.

Finally, a third system integrating results from studies done on samples allows as many projects as possible to gain maximum value from every investment.

1.3 African Livestock parasites and Poverty Reduction

A major factor in poverty reduction strategies is the setup and maintenance of stable sources of food and income. Keeping and breeding livestock is a way to achieve both these goals, while still needing minimal investment in surrounding infrastructure to initiate. This strategy is often prevented by the presence of local parasites able to completely wipe out populations of imported livestock.

1.3.1 *Theileria Parva* and East Coast Fever

Theileria parva is an intracellular protozoan parasite native to eastern, central, and southern sub-Saharan Africa. The parasite is spread through ixodid ticks and infects Cape buffalo (*Syncerus caffer*) and cattle (*Bos Taurus)*. While buffalo remain asymptomatic throughout the infection, susceptible cattle develop high fever and generally die within 14-21 days of the original infection (McKeever & Morrison, 1990).

In 1992 the disease was measured to cause an approximate $170M loss annually, killing over one million cattle. A more recent study in Sudan measured the total cost of two outbreaks affecting 3460 animals at $134.000. *T parva* is also considered a major limiting factor on the domestic cattle production, as imported cattle are more susceptible to infection, and thus, the ability to cross-breed or import high yield cattle breeds is limited (Marcellino *et al.*, 2011; Mukhebi *et al.*, 1992b; Mukhebi *et al.*, 1992a).

T. parva infection was originally controlled by acaricide treatment to prevent tick feeding, but efficiency has been reduced by ticks developing resistance, as well as environmental and food-safety concerns (George *et al.*, 2004), the infection and treatment immunization method (ITM) was developed. ITM is based on injecting animals with a possibly lethal dose of parasite sporozoites, while simultaneously giving a large dose of oxytetracycline to reduce the infection, allowing the animal to develop immunity as it recovers.

In 1975, a mix of three parasite strains was discovered to provide unusually broad and long-term protection was discovered (Radley *et al.*, 1975a; Radley *et al.*, 1975b; Radley *et al.*, 1975c). This vaccine, commonly known as the 'Muguga Cocktail', has since then been deployed successfully in several countries. There are several drawbacks to using live parasite vaccines though, including high production costs, and risk of contributing to ECF spread if the vaccine is mishandled (Di Giulio *et al.*, 2009). Several studies have tried to identify the specific antigens that induce immunity, in order to create a vaccine that is cheaper to produce, and safer to handle, but none have yet succeeded in creating an equally effective vaccine.

T. parva belongs to the phylum *apicomplexa*, mostly known for *plasmodium falciparum*, the causative agent of Malaria in humans. This phylum also includes a number of other important parasites, including *Babesia*, another cattle parasite, *Cryptosporidium parvum*, a mammalian intestinal parasite, and *Toxoplasma gondii*, a parasite that infects almost every warm-blooded animal, and which a third of the human population is estimated to have been exposed too (Pappas *et al.*, 2009).

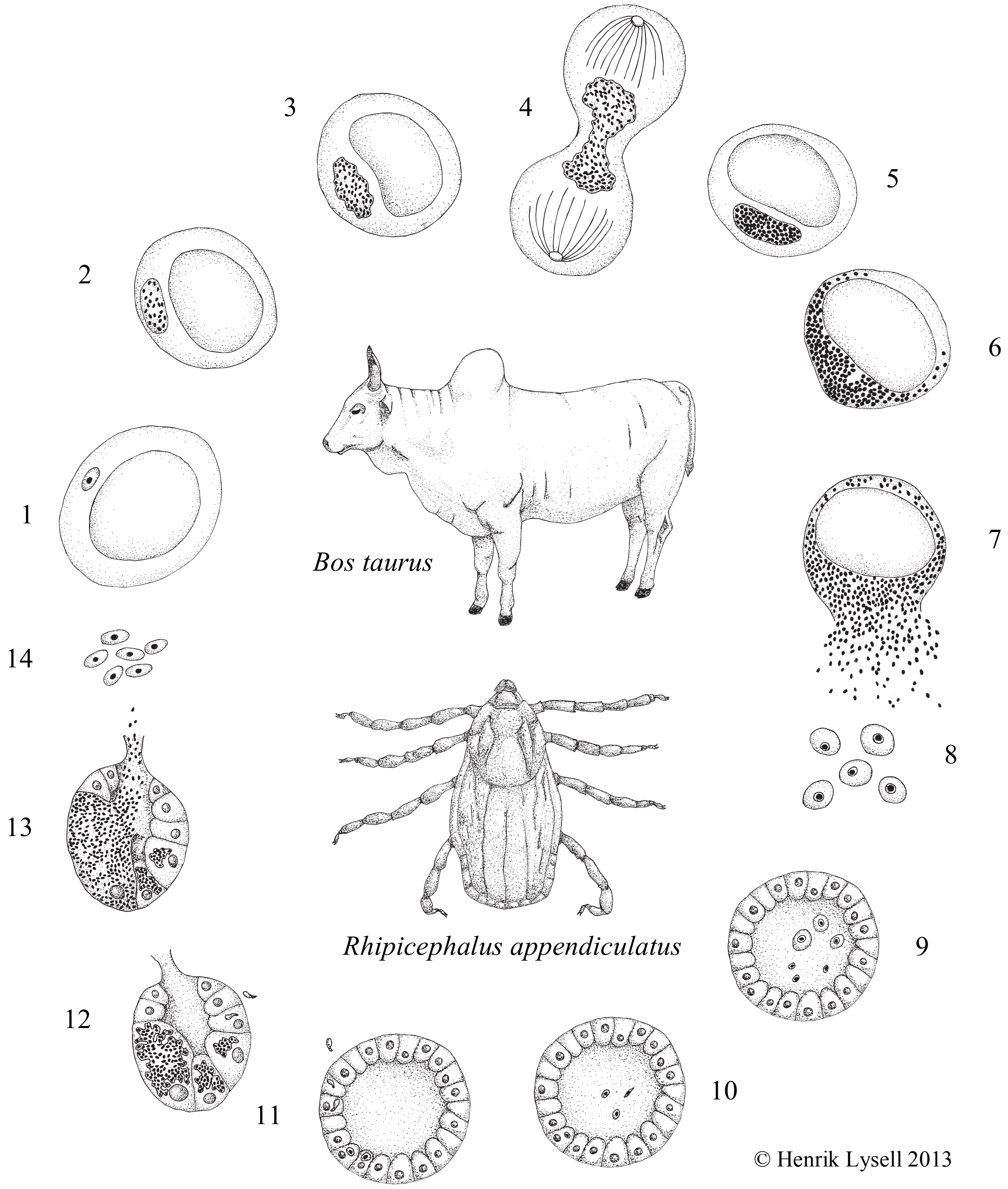
T.parva has a multi-species life cycle (summarized in *Figure 1*) including Cape buffalo (*Syncerus caffer*) or cattle (*Bos Taurus)*, and ixodid ticks, commonly *Rhipicephalus appendiculatus*. The parasite infects the mammalian host during tick feeding, where sporozoites from the tick saliva enter into the blood stream. The sporozoites bind to the host lymphocytes through a passive, orientation independent process, and are then internalized through "zippering" through the host membrane. While binding to the lymphocyte is a passive process, internalization is not. Once inside the lymphocyte, the parasite sheds the enclosing host membrane to lie freely in the cytoplasm (Shaw *et al.*, 1991).

Inside the host cytoplasm, the parasite quickly associates with the host microtubule system, which it associates strongly too. The parasite schizont then induces uncontrolled proliferation of the lymphocyte, multiplying by dividing in synchrony with the host cell.

In response to an unknown signal, infected lymphocytes stop dividing and enter merogony, causing the schizont to disassociate from the host microtubules, and start the production of merozoites. The merozoites enter the blood stream as the host cell is destroyed, where they infect erythrocytes. Merozoite entry into the erythrocyte is thought to use a similar mechanism as when the sporozoite enters the lymphocyte (Shaw, 2003).

Infected erythrocytes infect feeding ticks, and infect the tick gut epithelial cells, forming gametes, followed by the only sexual stage in the *T. parva* life cycle. This leads to the formation of motile kinetes, (Gauer *et al.*, 1995) which transfer to the tick salivary gland, inducing sporogony where new sporozoites are produced, allowing the tick to spreads the parasite to new mammalian hosts.

Understanding the different stages of the life cycle can be vital to a successful study. A study by Musoke and collaborators targeting sporozoite surface antigen p67 indicate that while this generate an immune response, an effective subunit vaccine will need to target multiple stages of the live cycle to guarantee immunity (Musoke *et al.*, 1993).



© Henrik Lysell 2013

Figure 1. Life cycle of *Theileria parva:* 1. Sporozoites infect the bovine lymphocyte. 2-3. The sporozoites replicate while interacting with the host microtubules. 4. The parasite initiates uncontrolled proliferation of host cells, replicating with the host cell. 5-7. Some cells start producing merozoites, destroying the host cell and spreading into the blood stream. 8. Merozoites infect the bovine erythrocyte, which is then ingested by ticks. 9-10. The parasite infects the gut lumen of the tick, and forms gametes, which undergo sexual reproduction. 11. Mobile gametes make their way into the tick salivary gland. 12-13. The infected salivary gland produces sporozoites. 14. The sporozoites are introduced to the bovine blood stream while the tick feeds (Shaw, 2003).

1.4 Bioinformatics

Bioinformatics can be defined as the application of computer science, mathematics, and engineering to study biological data. (Luscombe *et al.*, 2001) defines three broad aims for bioinformatics as i) Organize data in a way that allows researchers to access and contribute new data, ii) develop tools and resources to analyse the data, and iii) to use the tools to analyse and interpret the data in a biologically meaningful manner.

The history of bioinformatics, as described by (Attwood *et al.*, 2011) correspond well to this definition. With the advent of DNA sequencing technologies and protein structure prediction the need for computational methods to handle data became undeniable. The bioinformatics community quickly realized the need for proper storage and distribution of all the new data, and thus many publicly open databases, such as the Protein Data Bank (1972) and GenBank (1982) were founded. Around these databases grew sets of tools to extract, analyse, and refine the information, and the experience needed to convert this information to biological knowledge.

Today, bioinformatics is an important part of biological research, as it allows for the analysis of genomic sequence data, development of programs, databases, and storage systems, as well as systems for information distribution. Like statistics, bioinformatics can thus be seen as both a tool for biological scientists, and the research field in which these tools are developed, improved, and distributed.

1.4.1 Next Generation Sequencing

DNA sequencing by chain termination was first described by Sanger in 1977, effectively starting the era of genetic sequencing (Sanger *et al.*, 1977). This 'first generation' technology, nowadays commonly known as 'Sanger sequencing', allowed fragments of a few hundred base pairs of DNA to be sequenced. This sequencing technique became automated and parallelized in the 1990's, with the Human Genome project, culminating in the first draft Human genome in 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001). At this point, Sanger sequencing had developed to read lengths of ~1000 basepairs (bp), but still costing about $0.50 per kilo-base (kb) (Shendure & Ji, 2008).

In 2005, the first the so-called 'Next-Generation Sequencing' (NGS) or 'second generation' sequencing technology appeared. This was the 454 Genome Sequencer, soon followed by the Solexa/Illumina genome sequencing platform and the SOLiD by Applied Biosystems (van Dijk *et al.*, 2014). At release, the second generation produced short reads of 35bp (SOLiD/illumina) to 110bp (454) but they produced approximately 20M (454), 30M (SOLiD), and 100 million reads, at much lower costs than Sanger sequencing, allowing previously sequenced genomes to be re-sequenced using the previous version as a template (van Dijk *et al.*, 2014; Mardis, 2008).

This development allowed many more research projects to use DNA sequencing methods this drastically increased the need for bioinformatics. Since then, several new technologies have also been released, including the IonTorrent, and IonProton from Life Technologies, and all products have significantly increased in their output, read lengths, and quality, while prices have dropped. As an example, the Illumina HiSeq 2000 produce ~600Gbp data per run, producing approximately enough data for six human genomes in 11 days (Mardis, 2013).

Currently, the so-called 'third generation' sequencers are being tested. These include single molecule sequencing, allowing samples to be sequenced without first amplifying the included DNA. This reduces the preparation needed, and allows unknown/uncultivable microbes to be sequenced. These new technologies promise read-lengths of up to 10,000bp, but error rates are currently very high, and can be as high as 15% (Mardis, 2013).

1.4.2 Genomic Variants

A large part of comparative genomics is based on the study of variants. Sequencing and assembling genomes de novo (from fragments, without a template) is still a monumental task for everything but the simplest of organisms. Instead, re-sequencing a previously sequenced genome, or a close relative, the new sequenced is commonly mapped against the old one, and the differences are recorded, rather than assembling from scratch. This way, a map of single nucleotide polymorphisms (SNPs) as well as small insertions and deletions (indels) can be generated. These maps describe the differences between closely related organisms, but may not represent the entire variation between the true sequences, as the mapping introduces a bias towards similarity. Still – this method allows for analysis of closely related sequences with minimal time and monetary investment compared to re-doing the complete sequence.

With third generation sequencing technologies promising far longer sequencing reads, and single molecule sequencing, read mapping may soon be replaced by a dire need of an algorithm to efficiently align entire genomes instead, but in the end, the result will be the same, a map of variants and possible genomic realignments.

1.5 Metagenomics

Metagenomics, DNA sequencing directly from an environmental sample (or samples) in order to study the combined genomes therein, differs from conventional sequencing in a few ways. First of all, it bypasses the need for culturing, allowing the sequencing of uncultivable species. It may also produce hypotheses on community member interactions, and in some cases quantify proportions between species of the metagenomic community. (Kunin *et al.*, 2008)

Like all genomic studies, a metagenomic study starts with a sample, or set of samples. The details of how to sample an environment in order to get a complete picture of the metagenomic diversity is in itself not an easy task, and mostly relies on previous experience or from the statistics of preliminary studies. As the population characteristics of an environment can change drastically depending on multiple variables such as season, temperature, availability of nutrients, etc., the importance of sample metadata once again becomes apparent, in order to produce comparable data (Thomas *et al.*, 2012; Wooley *et al.*, 2010).

Direct sequencing of a sample without prior knowledge of content has a number of caveats though. While the recent years development in sequencing technologies has made sequencing several orders of magnitude cheaper compared to traditional Sanger sequencing, these new technologies produce shorter reads, of lower quality. This affects metagenomic projects a lot more than single genome projects, as metagenomic species cannot be assumed to be clonal, making it sometimes impossible to differ between species variants and sequencing errors (Wooley *et al.*, 2010; Wommack *et al.*, 2008). Modern sequencing technologies thus make quality filtering of raw sequence data incredibly important, and multiple methods to prune sequence data based on quality scores, repeats, etc. have been developed. If the sample comes from a host organism, be it human, animal or bacteria, the filtered reads are generally mapped towards the host genome as well, to leave only the non-host sequences in the metagenomic community.

Following pre-processing, the remaining sequences are commonly assembled into contiguous sequences (contigs). Metagenomic data assembly of is one of the most difficult assembly problems though, as the sequence assembly problem becomes increasingly difficult with higher number of community species, and especially as the number of species is generally unknown. The presence of genes conserved between species also introduces the risk of assembling the genes from the related species, creating interspecies chimera (Thomas *et al.*, 2012; Wooley *et al.*, 2010; Kunin *et al.*, 2008).

Ideally, the assembly contigs correspond to entire chromosomes and genomes. Generally though, this is very unlikely, and often the least numerous species in the metagenomic community might only be represented by a few non-overlapping reads (which in many cases cannot be connected).

If the goal is complete genomes, this preliminary assembly can be used to generate PCR primers, used in "flanking sequence walking", repeatedly extending the contigs by designing a sequencing primer close to the end of the contig, amplifying the source DNA using this primer, and sequencing. This can be done repeatedly to extend the contig until its end, or until a repeat region, or other region that cannot be traversed by a single read is found (Delwart, 2007).

Assembly is commonly followed by, taxonomic classification, or "binning", of the assembled contigs, and unassembled reads. There are two general classes of metagenomic classification algorithms, taxonomy-based and taxonomy-independent. Like the names suggest, taxonomy-based methods are based on patterns of known taxonomic groups. Taxonomy independent methods attempt to classify by similarity and sequence patterns, attempting to separate into species without attempting to discern which species, or their relationships. Taxonomy independent classification methods are thus closely related to unsupervised machine learning (Mande *et al.*, 2012)

Many methods perform quite well in classifying bacterial sequences using marker genes, such as the 16S and 18S ribosomal genes. The lack of such genes complicate classification of viral sequences though (Wooley *et al.*, 2010; Edwards & Rohwer, 2005). In addition, studies on viral metagenomics have found around 65% sequences with no closely related sequences in public databases, indicating that the viral population may be severely underrepresented (Edwards & Rohwer, 2005).

2 Aims of the Thesis

This thesis consists of four studies, tied together with a theme of metagenomics and scientific rigor in the field of metagenomic science.

This thesis is composed of the following studies:

1. Web-Based Biobank System Infrastructure Monitoring Using Python, Perl, and PHP.

2. Viral metagenomics demonstrates that domestic pigs are a potential reservoir for Ndumu virus.

3. Whole Genome Comparisons of the Three Major Stocks Comprising the Live Sporozoite Theileria parva Vaccine.

4. MetLab: An *In Silico* Experimental Design, Simulation and Validation Tool for Viral Metagenomics studies.

3 Web-Based Biobank System Infrastructure Monitoring Using Python, Perl, and PHP

3.1 Background

A biobank can be a veritable treasure trove of scientific information, just waiting to be unfrozen and refined into knowledge. A well-kept biobank can hold samples from rare disease outbreaks, uncultivable parasites, and a wide range of other hard-to-find cases.

Samples intended for biobanking have higher demands on both sampling and storage than samples collected for specific projects. As they must be prepared for planned studies, and for as many unknown, future studies as possible, biobanking often requiring the sample to be split, then prepared and stored in multiple ways for multiple kinds of analyses (De Paoli, 2005). The biobank concept range from small institution repositories consisting of a freezer and a notebook, to large-scale commercial enterprises with millions of samples in liquid nitrogen, advanced database integration and robotics for sample handling. Biobanks generally have a specific purpose as well, as examples, many countries, such as Iceland, maintain population banks, and many hospitals maintain banks of clinical samples (De Souza & Greenspan, 2013).

The need for proper monitoring and logging of samples in large biobanks is obvious, with tens of thousands of samples needing to be maintained for decades, losing information, or thawing the samples, can destroy vast amounts of potential knowledge and monetary investments (Riegman *et al.*, 2008).

The need for small labs to run similar systems for their regular freezers and samples is less obvious, but no less important. A freezer door left open overnight can render samples for multiple projects unusable, and cause vast amounts of trouble for the sample owners. If such an event is not noticed, erroneous information from flawed samples may result in false conclusions from a study.

Demands for a small-scale lab monitoring system differs quite a lot from large-scale projects though. Smaller system can't rely on having infrastructure designed specifically for their installation, or investing in specially outfitted hardware. Thus the system needs to be modular, adaptable, and rely only on affordable standard components.

3.2 Methods

To allow a system to be constructed, maintained, and continuously updated with the smallest possible investment in time and money, an open source system was designed, based on standard hardware components. The total hardware is a Raspberry Pi, LabJack automation devices, two servers, temperature and humidity probes, serial to Ethernet adapters, web cameras, a USB-modem, and network switches. A conceptual schematic of the system can be seen in *Figure 2*.

The system uses EI1022 and EI1034 temperature probes to monitor refrigerators, freezers, and incubators, as well as the stand-by generator, server room and liquid nitrogen biobank. These probes differ in measuring range and sensitivity, with the EI1022 measuring between -40°C and 100°C with a sensitivity of ±1.5°C, and the EI1034 measuring between -17°C and 110°C with a sensitivity of ±0.3°C. A combined temperature and humidity-measuring probe, EI1050, is being evaluated for use in greenhouses and server rooms as well.

All measuring probes are connected to a LabJack, model UE9, UE9-pro, or U6. The UE9 versions have 16 analog inputs, 23 digital I/O ports, as well as Ethernet and USB interfaces and an internal temperature sensor used to measure ambient temperatures. The UE9-pro also has a high-resolution analog to digital converter. The UE6 lacks Ethernet connection, but is capable of measuring very low voltages, making it ideal for reading information from the liquid nitrogen system. As the UE6 has no Ethernet connection, it is instead connected to a Raspberry Pi mini computer via USB. This Raspberry Pi runs the program that monitor the liquid nitrogen bulk tank, as well as monitoring the door and the three cameras that are part of the biobank access control system.

The liquid nitrogen biobank is made up of five Taylor-Warton LAB80K freezers, holding approximately 80.000 samples each, and a smaller liquid nitrogen freezer. A 3000L liquid nitrogen holding-tank monitored by the LabJack U6 supplies the liquid nitrogen freezers. All these freezers have built-in temperature measuring and alerting systems, which are interfaced using an RS232 serial interface connected to a serial to Ethernet converter.

The biobank is located in a single entry room, monitored by three Ethernet-enabled web-cameras. Two of these cameras record the inside of the room, and the third record outside the entry. These cameras, along with the door switch, document anyone entering or leaving the room. The system also logs liquid nitrogen level, liquid nitrogen filling, lids opening and closing, and any alarms. All events, along with camera images, are stored in a MySQL database log, as well as sending an SMS picture to responsible personnel whenever a freezer lid is opened.

The systems has two servers, one runs a MySQL database, storing all logs, temperatures, and images from the logging systems, while the other runs the actual logging scripts and the web application. All records are continuously compared against previous values, making sure that no values are outside set limits. Values are stored at a minimum every 30 minutes, but values are checked every few seconds, and if they diverge too much from the last logged value, they are stored. This system ensures high detail during changes, and efficient storage of stable values. SMS alerts are sent out if any values exceed their limits, alerting responsible personnel.

The system is built with two servers to allow it to be self-monitoring as well. The application server and MySQL server continuously check each other as well as the logging systems, ensuring that silent failure is almost impossible. All temperature and other logging data are connected to the Laboratory Information Management System (LIMS), connecting each sample to its temperature history.

The last part of the system is the web interface. The interface continuously updates graphs and tables to give an easy overview of the temperature history, and any events that's happened recently. The system is openly available at http://azizi.ilri.cgiar.org/.

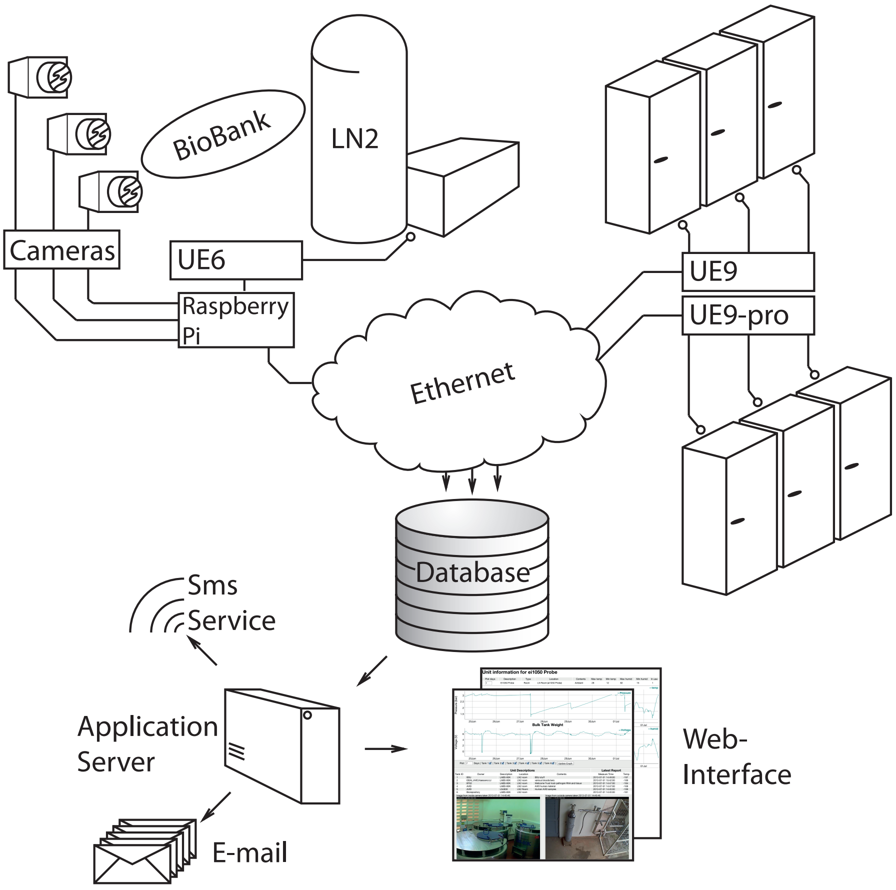


Figure 2. General design of the biobank surveillance system setup at the International Livestock Research Institute in Nairobi, Kenya.

3.3 Results, Discussion and Future Prospects

The results from a system like this are of course immeasurable, as the successful deployment of the system has the same effect as having no accidents without the system. That said – the system did on several occasions alert personnel to freezers that were not properly closed, or had lost power, which may not have been detected without proper monitoring.

Since the users design the system, it can simultaneously be designed to make the system safer and easier to use. Making sure that interfaces present the information that scientists require in their everyday work, creates a second layer of security, as the system makes work easier, making staff happy to use it. Building your own system can be time consuming though, and requires rigorous testing to make sure nothing can fail without an alert being sent.

4 Viral metagenomics demonstrates that domestic pigs are a potential reservoir for Ndumu virus

4.1 Background

As the demand for livestock products increases in Africa, so has the domestic pic production increased and intensified across the continent. Their relatively rapid growth rate and large litter sizes, as well as the wide range of possible feed make pigs ideal for meat production in many locations. In Uganda, the domestic pig production industry has developed quickly, but as the production increases, so does the economic dependency on this production system increase. Greater pig numbers mean that disease outbreaks have the potential of faster spread and more devastating consequences. Considering the wide range of zoonotic diseases able to spread between human and pig, this is clearly a matter of concern (Smith *et al.*, 2011; Blomstrom *et al.*, 2009; Myers *et al.*, 2007; Banks *et al.*, 2004; Parashar *et al.*, 2000; Philbey *et al.*, 1998).

Since resources are lacking in many developing countries, public health surveillance and development of early warning systems are often lacking. Where possible, viral metagenomics using high-throughput sequencing is therefore being increasingly applied in veterinary epidemiology.

4.2 Methods

Serum samples were collected from 16 domestic pigs (*Sus scrofa*) from five sites in Uganda as part of a study on African Swine Fever (ASF), which is endemic in the region. The samples were pooled based on their sampling location, DNA and RNA were extracted, and libraries were prepared for GS FLX 454 sequencing. The sequencing reads had primers removed, and reads shorter than 50 base pairs were discarded. Repeats were masked using RepeatMasker[1], and reads with more than 50% repeats were discarded.

Each pool of filtered sequences was assembled using the Newbler assembler[2] using default settings. All assembled contiguous sequences (contigs), and unassembled reads, were then classified using BLASTN and BLASTX (Altschul *et al.*, 1990) against the non-redundant nucleotide and amino acid databases from NCBI[3] followed by taxonomic classification with MEGAN 4.0 (Huson *et al.*, 2007).

4.3 Results

The sequencing generated a total of 289,038 reads (average length 175 nucleotides), of which 190,706 remained after filtering. The assembly sorted 77% of the reads into contigs, which were then analyzed with BLASTX and BLASTN. As BLASTX and BLASTN gave similar results, only BLASTN results are reported.

The analysis resulted in 38% of all sequences having a GenBank match. Of these, 52% (of both DNA and RNA) mapped to the domestic pig and 36% mapped to other mammalian genomes (these matches may be, at least in part, due to the incomplete status of the porcine genome). A total of 6.2% of DNA and RNA sequences mapped to viruses. Of the DNA, African swine fever virus (ASFV) was identified in all pools. Torque teno viruses (TTVs) were identified in two pools. These are species-specific viruses currently considered non-pathogenic, and which are known to infect pigs. Among the RNA pools, three pools showed significant sequence similarity with porcine endogenous retroviruses. Lastly, two pools showed significant identity (98%) to Ndumu Virus (NDUV).

To validate the NDUV findings, all reads that were part of these two pools were re-classified by BLASTN, detecting 2% and 5% viruses. Notably, no reads mapped to class *Insecta*, indicating that this finding was not the result of insect contamination. To validate Ndumu virus, a mapping assembly was made against the public Ndumu virus sequence from GenBank (NC_016959) using Roche gsMapper, resulting in two contigs with an average coverage of 10x.

PCR-amplification using NDUV specific primers further verified the presence of NDUV. The PCR product was sequenced on an ABI Prism 3700 DNA analyzer and the resulting sequences were compared against other alphaviruses published in Genbank (AF069903, U73745, EF536323,

[1] RepeatMasker Open-3.0.1996-2010 http://www.repeatmasker.org

[2] Newbler 2.5.3

[3] version June 2011

AB032553, AY702913, GQ433358, X04129, AF369024, AF079456, NC_016962, NC_016959, M69205, and HM147992). The protein coding regions of all these viruses, as well as Ndumu virus and Semliki forest virus, were aligned using Clustalw (Larkin *et al.*, 2007), and a phylogenetic tree was constructed using MEGA 5.0 (Tamura *et al.*, 2011). This analysis show the virus found in this study clustering far closer to Ndumu virus than any other in the group. This confirms that the detected virus is an Ndumu virus, found for the first time in a vertebrate host, the domestic pig.

4.4 Discussion and Future Prospects

This study used metagenomic sequencing of domestic pig (*sus scrofa*) serum in order to characterize the viral community. In addition to finding ASFV and TTVs, Ndumu virus, which has never before been found in pigs, was detected.

Ndumu virus is a single stranded RNA arbovirus of the *Togaviridae* family transmitted by mosquitoes. Infected mice do not survive infection, but although antibodies to the virus have been identified in humans, no human mortality has been attributed to infection with Ndumu virus.

This study shows the importance of the metagenomic approach in veterinary medicine, as it allows for detection of pathogens in unknown hosts, as well as pathogens that are completely unknown.

5 Whole genome comparisons of the three major stocks comprising the live sporozoite *Theileria parva* sporozoite vaccine

5.1 Background

Theileria parva is an intercellular parasite, which infects Cape buffalo (*Syncerus caffer)* and cattle (*Bos taurus*). It is native to eastern, central, and southern Africa, where it. The infection spreads between buffalo and cattle via ticks, primarily *Rhipicephalus appendiculatus*, it cannot spread directly between infected animals.

Infection with *T.parva* is asymptomatic in buffalo, which is believed to be the ancestral host of the parasite. In cattle, infection causes an acute lymphoproliferative disease called East Coast Fever (ECF). Untreated cattle develop high fever, disease of the lymph nodes, and a reduction in white blood cells. Infected, susceptible cattle generally die within 14-21 days of the original infection as a result of the parasite spreading throughout the tissues (McKeever & Morrison, 1990).

Currently immunization against ECF is done by controlled infection with a live *T.parva* stabilate along with a long-acting formulation of oxytetracycline. This is known as the "infection and treatment method", or ITM (Di Giulio *et al.*, 2009; Bishop *et al.*, 2001). Lasting protection against a variety of strains is obtained with this method using a mix of three parasite stocks – Muguga, Kiambu5, and Serengeti-transformed, commonly known as the "Muguga cocktail". It has been shown that the induced ECF immunity lasts at least 43 months without further challenge, but it is believed that given regular natural exposure to the parasite, immunity may be maintained indefinitely (Burridge *et al.*, 1972).

The Muguga cocktail vaccine has been widely deployed among Tanzanian Maasai pastoralists, with over 400,000 calves vaccinated. It has also been used in pilot studies in the Ugandan dairy sector, and elsewhere in east Africa. The success of these vaccinations has created an increased demand in Southern Sudan and Kenya as well (Morzaria *et al.*, 1999).

There has been reports of incorrectly used vaccine causing severe ECF infection, but increasing the dose of oxytetracycline has reduced this problem significantly (Di Giulio *et al.*, 2009). ITM has several other drawbacks though. It requires a cold chain for delivery, and the cost of production and verification of new vaccine batches make the vaccine expensive. In addition, vaccine production requires infecting cattle, which may result in genetic recombination between strains (Henson *et al.*, 2012).

In this study we studied the genomic sequence of all Muguga cocktail stocks, with the intent of characterizing the genomic diversity in the antigen genes.

5.2 Methods

Three stocks of *T.parva* were sequenced in this study. Muguga was derived from piroplasms purified from animal BM256 infected with cloned parasite stabilate 3968. Kiambu5 was derived from stabilate 4137, which is directly derived from stabilate KV 68 which was used in the first batch of the Muguga cocktail. The Serengeti-transformed was derived from Serengeti-transformed seed stabilate 69, which is a precursor to the FAO1 ITM vaccine.

All experiments on cattle were performed according to the ILRI institutional animal care and use committee, which complies with the United Kingdom government rules and guidelines for animal experimentation. Before infection, all animals, Friesian calves aged 6 months, were screened using ELISA with a *T.parva*-specific p104 PCR assay, as well as an 18S rRNA gene PCR assay to verify that they had no previous exposure to the parasite.

The Muguga and Serengeti-transformed stocks were sequenced on a 454 GS FLX sequencing platform, and the Kiambu5 stock was sequenced using an IonTorrent 416 chip. All sequences were mapped against the Muguga reference strain (GenBank AAGK01000001- AAGK01000009) (Gardner *et al.*, 2005). The 454 sequences were mapped using the GS Reference Mapper (454 Life Science), and the Kiambu5 sequences were mapped using the Burrows Wheeler aligner (Li & Durbin, 2009) using default parameters.

Variants were called for all strains using the Genome Analysis Toolkit form the Broad Institute (McKenna *et al.*, 2010), and the variants were filtered using SAMtools (Li *et al.*, 2009). These variants were then used to create ‘pseudosequences', sequences identical to the reference except for the variant locations, from all known coding sequences in the *T. parva* genome. These were aligned back to the reference, and dN/dS values, the ratio of non-synonymous to synonymous SNPs compared to random mutations, were calculated using the LPB93 algorithm of the yn00 program in the PAML package (Yang, 2007).

To visualize the genetic relationship between the isolates, a dendrogram was calculated using the stocks in this study, as well as the genomes published by Hayashida and collaborators (Hayashida *et al.*, 2013). 100 coding sequences were randomly selected, and concatenated into pseudo-contigs, which were then align and a dendrogram calculated using clustalw2 (Larkin *et al.*, 2007), and plotted using the APE library of R (Team, 2013; Paradis *et al.*, 2004). The finished dendrogram can be seen in *Figure 3*.

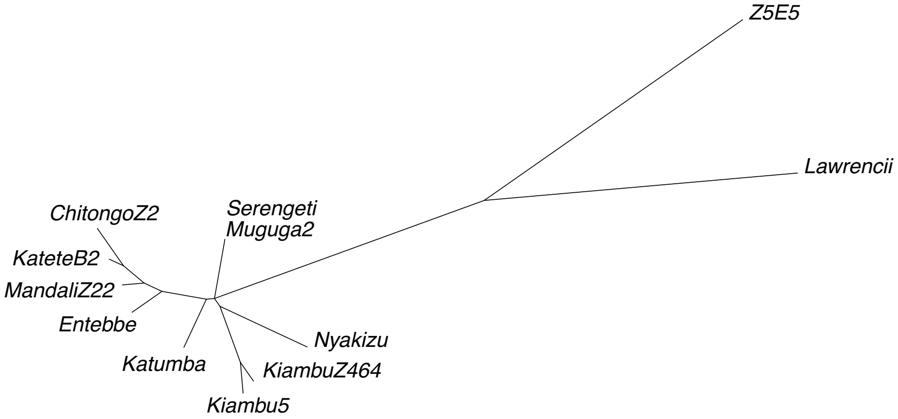


Figure 3. Dendrogram of multiple *T. parva* stocks, based on 50 randomly selected coding sequences. The *Z5E5* and *Lawrencii* are buffalo derived parasite strains while the remainder are cattle derived. Note that for these randomly selected sequences, there were no differences between Muguga and Serengeti-transformed.

5.3 Results

In this study, three *T. parva* isolates were sequenced using two different sequencing platforms, and mapped against the previously sequenced Muguga reference genome. The mapping resulted in a 94% reference coverage for the Kiambu5 stock, and 98% and 99% for the Muguga2 and Serengeti-transformed stocks respectively. The three stocks show very high sequence similarity, with SNP densities from 0.10 SNPs/Kbp for Serengeti-transformed to 4.70SNPs/Kbp for Kiambu5. Muguga2, a re-sequencing of the Muguga reference strain, was included to measure in-strain variation, and it was proven to be almost identical with only 586 variants compared to the reference. The Kiambu5 isolate was the most divergent, with 39,296 variants relative the reference. Unexpectedly, Serengeti-transformed, thought to be the most divergent strain as it was originally derived from buffalo, only showed 957 variants compared to the reference. A summary of the variant calling can be seen in Table 1.

Table 1. Summary of genomic variants of the *T. parva* stocks compared to the Muguga reference sequence.

.Strain	Total number of SNPs	Synonymous	Non-synonymous	Nonsense	Indels	Intron
Serengeti-transformed	957	291	420	1	0	121
Kiambu5	39 296	20 086	8 587	16	748	7 615
Muguga2	586	206	264	1	0	48

The high similarity between Serengeti-transformed and Muguga allowed detailed study of all variants. Most interestingly, non-synonymous modifications were discovered in seven ABC-transporter genes, one of them being TpABC2, described in (Kibe *et al.*, 2001), and five of its paralogs.

As a total of 2,233 CDSs has sequence variants between Kiambu5 and the Muguga refrence, including 1,708 with non-synonymous changes, not all variation could be manually evalutad. Instead, a set of genes was selected using a number of criteria.

The first subset of genes to be evaluated was previously known antigens, namely *T. parva* antigens Tp1-5, Tp7-9, p32, p67, p104, p150, PIM, gp34, X88, and 11E (Xue *et al.*, 2010; Graham *et al.*, 2006; Skilton *et al.*, 2000; Skilton *et al.*, 1998; Daubenberger *et al.*, 1997; Ebel *et al.*, 1997; Iams *et al.*, 1990a; Iams *et al.*, 1990b). This set was combined with the genes from the dN/dS analysis that had dN/dS ratios over 4 standard deviations above the mean, or a dN value more than 4 standard deviations above mean.

As a mapping assembly relies on similarity to the reference, we also made a list of the 100 genes with the lowest mapping scores, as these may indicate regions with high levels of sequence variation or genomic rearrangements outside of the range of mapping assembly.

5.4 Discussion and Future Prospects

The Muguga cocktail includes three parasite stocks, each one believed to add to the effectiveness of the vaccine. A notable result of this study is the extreme similarity between the Muguga and Serengeti-transformed stocks, with Serengeti-transformed having 41 times less variants than Kiambu5 compared to the Muguga reference. This leads to one of two possible solutions: i) the Serengeti-transformed strain is closely related to Muguga, but within the 52 CDSs with non-synonymous SNPs is the genetic reason for its contribution to the vaccines effect, or ii) the included Serengeti-transformed is the result of a previous mix-up with Muguga, and thus only represent a second Muguga stock. As the need to re-investigate the contribution of the Serengeti-transformed stock in the Muguga Cocktail was discussed over ten years ago in (Bishop *et al.*, 2001), we are inclined to lean towards the latter explanation. In either case, if more of the Muguga cocktail is to be produced, the inclusion of the Serengeti-transformed stock should be thoroughly evaluated.

Going further with this research, the results from this study should be experimentally evaluated in order to reduce or verify the list of potential antigen candidates identified in this study. Further, the Kiambu5 stock, since it may be the main contributor to the success of the Muguga cocktail along with Muguga, should be the focus of a proper genomic study. This is to resolve whether there are genomic rearrangements, or unique genomic features which may contribute to the live vaccine.

6 MetLab: An *In Silico* Experimental Design and Validation Tool for Metagenomics

6.1 Background

Where metagenomics is the study of all genomes within an environment, viral metagenomics is the study of the virome, the set of viral genomes, within the same environment (Hugenholtz & Tyson, 2008; Edwards & Rohwer, 2005). Samples are prepared and treated in order to reduce the host genome (if any), this way the virome of a sample can be studied without the need of culturing, allowing for unbiased study of the included viruses. With the reduction in cost and increased availability of sequencing technologies, combined with the development of better analysis tools, the usefulness of this methodology has increased dramatically (Chen & Pachter, 2005).

Metagenomic investigation is a complicated problem, and viral metagenomics has a number of unique complications associated with it. First of all, viruses are small. Even small bacterial genomes are huge compared to common viruses, meaning that the bacterial sequences will drown out the viral ones (Edwards & Rohwer, 2005). This can be solved either by enriching the viral DNA, or by increasing sequencing depth (Rosseel *et al.*, 2013; Willner & Hugenholtz, 2013), but still problems remain. Some studies estimate that 95% of the viral diversity is unknown (Delwart, 2007; Edwards & Rohwer, 2005), meaning that taxonomic identification may be impossible.

A core problem of metagenomics is the experimental design, it you don't know what you're looking for, how do you know if you've found it? With enough coverage all species within the community will be found, but how do you estimate the needed sequencing? Coverage theories for metagenomics attempt to solve this problem with statistics and probability theory, enabling the needed sequencing depth to be estimated. This estimation is dependent on good estimates of the genome size range and relative species abundance in the samples though (Rodriguez & Konstantinidis, 2014; Wendl *et al.*, 2013; Willner & Hugenholtz, 2013).

Metagenomic studies can be divided into four general parts: quality control, assembly, taxonomic classification/binning, and functional annotation of results (Stanhope, 2010). While single genome assemblies from clonal DNA can handle lower-quality sequences to some extent, by error correcting based on overlapping reads, metagenomic studies have no way of differentiating errors from variation in the final result. Because of this, quality control pre-processing is extremely important as it reduces the chance of miss-assembly, and thus misestimation of sample diversity (Gomez-Alvarez *et al.*, 2009; Hoff, 2009). The complexity of the assembly increases as well. Metagenomic assembly is the opposite of traditional assembly in that is focuses on low mean coverage and a highly diverse pool of genomic sequences (Kunin *et al.*, 2008). This problem can be simplified somewhat for viral metagenomics based on animal or human samples by removing all sequences that map towards the host genome.

The final general parts of a metagenomic study are taxonomic classification and functional annotation. Taxonomic classification is generally one of the main objectives of a metagenomic study (Edwards & Rohwer, 2005). This problem is related to assembly in that they are heavily dependent on each other. Being able to classify all the reads correctly would significantly improve assembly, and a correct assembly will significantly improve classification. Many classification methods are based on comparison to known genomes, so called taxonomy-based methods, while others are based on sequence composition, codon usage, patterns, and motifs, so called taxonomy-independent methods (Mande *et al.*, 2012). As some studies indicate that as much as 95% of the viral diversity might be unknown (Delwart, 2007; Edwards & Rohwer, 2005), taxonomy-based methods have a disadvantage in classification, but the advantage that classification also identifies species and relations.

Finally, annotation is the task of identifying genes and functions in a genome. This is an important task, but as it is highly dependent on the type of genome, and focus of the study, it will not be part of this general framework.

In this study we aim to simplify the problems of metagenomics, quality pre-processing, assembly, and taxonomic binning by implementing a number of useful tools into an easy to use system called “MetLab”.

6.2 Methods

MetLab is modularly designed, with the computational modules distinct from the graphical user interface (GUI). The computational modules can be used independently as stand-alone command line applications, while the GUI provides a user-friendly way of accessing the tools. The entire application is written using Python2.7[4], with the GUI being a standard python TkInter interface[5]. The application is open source, released under the GPLv3 license. Using python gives the application platform independence, while the open source license ensures that the system can be freely extended by whoever wishes to do so.

The system has three modules: an experimental design module that helps with metagenomic sequencing estimation, a viral metagenomic dataset simulator, and a metagenomic analysis pipeline module including three pipelines.

Metagenomic sequencing estimation is done using the algorithms published by Wendl *et al.* (Wendl *et al.*, 2013), which are an adaptation of Stevens' theorem. This algorithm has some caveats though, as it is based on a large sum of alternatingly very large and small numbers. Standard precision arithmetic in most programming and scripting languages will, given such a sum, repeatedly hit their limits, introducing a growing rounding error that may heavily distort the results. To combat this problem, and speed up the computation of this computationally intensive algorithm, the experimental design module is written as a Python C-extension using the GNU MPFR library (Fousse *et al.*, 2007), an arbitrary-precision floating point library. This module includes implementations of theorem 1 (gap consensus) and its first corollary for calculating the complete coverage probability, presented in Equation 1.

$$P(B=k) = \binom{R}{k} \sum_{\beta=k}^{\eta} \binom{R-k}{\beta-k} (-1)^{\beta-k} \alpha^{\beta} (1-\beta\varphi)^{\beta-1} (1-\beta\varphi\alpha)^{R-\beta}$$

$$P(B=0) = \sum_{\beta=0}^{\eta} \binom{R}{\beta} (-\alpha)^{\beta} (1-\beta\varphi)^{\beta-1} (1-\beta\varphi\alpha)^{R-\beta}$$

Equation 1. Metagenomic gap consensus probability, and complete coverage probability (B=0), where B is the target number of sequence gaps in an assembly of k gaps, R is the number of sequence reads, φ is the probability of a position being covered, α is the species abundance in the community and η is the smaller of R and int($1/\varphi$).

[4] http://www.python.org

[5] http://tkinter.unpythonic.net/wiki/

Both functions are available in the module and can be used from the command line, while the GUI simplifies use by estimating full coverage from MetaMaker profiles.

MetaMaker is the second module of the system, which is the metagenomic sequencing simulator. It has two main functions; first it can calculate a statistical profile from real world sequencing data, secondly it can use these profiles, along with random downloaded viral genomes, to generate simulated viral metagenomic data sets. The script will download a user defined number of random viral genomes from the National Center for Biotechnology Information (NCBI), and use the statistical profile to generate reads with the same length and error profiles as the original data. The module uses BioPython[6] and NumPy[7] for bioinformatic functions and efficient numerical calculations.

The final module is the analysis pipelines module. This module runs configurable pipelines for metagenomic analysis. The default version has three built-in pipelines, which all require single-end or paired-end sequencing data as input. Common to all pipelines is quality pre-processing with Prinseq-lite (Schmieder & Edwards, 2011), and taxonomic classification with kraken (Wood & Salzberg, 2014).

The first two pipelines are named "Environmental sample", which runs kraken directly on the pre-processed sequence reads, and "Environmental sample with Assembly", which does de novo assembly of the pre-processed reads using MIRA4 (Chevreux *et al.*, 1999) before classification. The final pipeline is called "Animal Sample", and differs from the other in having an extra pre-processing step; using Bowtie2 (Langmead & Salzberg, 2012) to map reads towards a user provided host genome, extracting the unmapped reads using SAMtools (Li *et al.*, 2009), and finally classifying the unmapped data.

6.3 Results, Discussion and Future Prospects

Validating the experimental design implementation proved difficult, as we couldn't find a previous implementation to compare towards. Instead, a naïve implementation was written in python, and compared using small numbers to avoid the accumulation of rounding errors.

To select binning tools, the MetaMaker module was used to create a simulated IonTorrent viral metagenomic dataset, with sequences from 58 different viruses, exponentially distributed in the sample. Quality pre-

[6] http://biopython.org

[7] http://numpy.org

processing and classification with kraken resulted in 80,59% correct classification.

As most classification algorithms are too slow to work directly on short-read datasets, the dataset was assembled into 366 contigs (N50 18.071). Only 11.3% of the reads were assembled, but as the dataset contains 58 species, many with very low abundances, this is not unexpected. Validation of contig classification is more difficult than read classification, as the taxonomic information is not easily available after the assembly. Comparing the classification results proportions to the known content of the sample gives a strong indication to the quality of classification though. All contigs were classified with five selected tools: Kraken (Wood & Salzberg, 2014), MEGAN5 (Huson *et al.*, 2007), ProViDE (Ghosh *et al.*, 2011), BLAST-LCA from the Fragment Classification Package (Parks *et al.*, 2011), and NBC (Naïve Bayesian Classification tool) (Rosen *et al.*, 2011), using two different training sets for the classifiers, one with only viruses, and one with viruses, bacteria, and archea.

Summarily, kraken gives very similar results to how it classified the raw reads, indicating that the method is consistent. MEGAN produced very similar results to kraken, while NBC classified over 60% of the set as bacterial and Blast-LCA and ProViDE gave 54% and 40% unclassified sequences respectively. As kraken does not rely on a resource consuming BLASTX alignment, it classifies much faster than MEGAN, which resulted in us using kraken as the systems default classifier.

Testing on simulated data is good for comparing accuracy of classification, but using real world data is of course better. To further test the system two sets of published metagenomic data was classified. The first dataset is from (Dacheux *et al.*, 2014), "*A preliminary study of viral metagenomics of french bat species in contact with humans: Identification of new mammalian viruses*", and the second from (Granberg *et al.*, 2013) "*Metagenomic Detection of Viral Pathogens in Spanish Honeybees: Co-Infection by Aphid Lethal Paralysis, Israel Acute Paralysis and Lake Sinai Viruses*". Comparing our pipeline to these published datasets show some differences in viral detection, notably, kraken fails to detect some viruses found in these studies with BLAST based methods. This is likely due to the stricter alignments used by kraken, as both papers indicate that the alignment hits for these detections where not very exact. MetLab is intended to be a first step in metagenomic analysis though, and if the intention is to identify novel viruses or distantly related viruses to published genomes, the kraken classification can still be used as a cleaning step to remove known sequences from the pool.

7 Conclusions

There are many conclusions to be drawn from this thesis; first of all, they show the strength of metagenomic methodologies. As genomic sequencing move towards higher output and lower prices, and the third generation sequencing promises longer read lengths and less input material needed, we may very well be facing an explosion of metagenomics, similar to the recent years rapid expansion of single genome sequencing projects. As more genomes are identified and studied using unbiased methods, the higher the value of future studies, as classification becomes more and more precise, making metagenomics more and more viable as a method. This is especially true for viral metagenomics, which is currently hampered by lack of tools and classification databases. The unexpected finding of Ndumu virus in Ugandan pigs underline how little we currently know about the viruses carried in even our most common domestic animals, and with the potential for zoonotic spread, there is much to gain by further study.

The study on *Theileria parva* also shows the importance of comparative genomics, and the need for proper, detailed study of known pathogens. Understanding pathogen biology requires understanding the genomic variability as well, both in finding targets for possible vaccine studies, to help understand cellular biology in general, and to understand how pathogens evolve and adapt to new hosts.

Lastly, with biological research producing more and more genomics projects, a greater need for bioinformatic analysis of results emerges as well, leading to a need to automate and optimize as many steps as possible. There are several benefits to this, in that it ensures that researchers time is spent as efficiently as possible, as well as simplifying the repeatability of studies.

Research is very resource consuming, and the further new projects go beyond the current knowledge, the less knowledge is left to inform decisions, making it more and more complex to go forward. This complexity is necessary for research; it can by very definition never be reduced. But there are several things that can be optimized in many research endeavours. This ‘unnecessary complexity’ comes from all the things that are known, but which still needs to be done, like routine analyses, sampling, storage, logging, etc. Investing in research infrastructure can reduce these problems, but can become a problem in itself. Having invested in certain systems mean that you are more tied to using them, having essentially traded flexibility for efficiency. These trade-offs should always come into consideration when planning the future of a research group or institution, and the importance of flexibility and adaptability should be stressed at all levels of a research institution, including infrastructure.

8 Populärvetenskaplig Sammanfattning

8.1 DNA-sekvensering, Bioinformatik och Metagenomik

När det mänskliga genomet, samlingen av allt en människas DNA, kartlades mellan 1990 och 2003, var det en långsam, arbetsam och dyr process. Projektet för att ta fram det första mänskliga genomet kostade nästan 3 miljarder dollar. Sedan dess har DNA-sekvenseringstekniker gått framåt med en otrolig hastighet. Idag kan ett motsvarande projekt slutföras på någon månad till en kostnad av några tusen dollar.

DNA-sekvensering är inte en enkel process dock. Det mänskliga genomet är 3.2 miljarder baspar långt, och vid DNA-sekvensering får man kortare fragment som behöver sättas ihop. Sanger-sekvensering, den tidigaste typen av sekvensering ger fragment på ungefär 1000 baspar med hög kvalitét, men de är dyra att producera. Så kallade ”andra generationens” sekvensering producerar istället fragment på 35-400 baspar, men producerar stora mängder data, billigt, på kort tid. Den bioinformatiska processen för att sätta samman stora mängder små fragment är dock mycket mer komplicerad än att sätta samman längre fragment.

Bioinformatik kan beskrivas som applikationen av datavetenskap, matematik och statistik för att lösa biologiska forskningsproblem. Stora delar av den bioinformatiska processen går ut på att hantera stora mängder data, göra den tillgänglig för andra forskare, samt utveckla verktyg och metoder för att analysera dessa data. En stor del av bioinformatisk forskning relaterar till sekvensanalys, då DNA-sekvenseringsprojekt producerar enorma mängder data som kräver mycket datorkraft för att hantera och analysera.

Metagenomik är en metod där man, istället för att sekvensera en enda organism, sekvenserar allt genetiskt material i ett prov. Ett problem för sekvensering av många organismer, så som bakterier och virus, är att man behöver kunna odla dem för att producera DNA för sekvensering.

Metagenomisk sekvensering har många fördelar; det kan användas för att kartlägga vilka arter som finns i t.ex. en viss jordmån, ett djurs tarmflora, sjuka djur, eller liknande. Det finns också problem. Att sätta samman *ett* genom är krångligt, men att sätta samman ett okänt antal genom från okända arter är extremt mycket svårare.

8.2 Biobankövervakning över internet

Insamling av biologiska prover är grunden för biologisk forskning. Provtagning är dock både dyrt och komplicerat, samt kan vara beroende av särskilda omständigheter såsom sjukdomsutbrott, klimatförhållanden, eller andra sällsynta faktorer. Ett sätt att hantera detta problem är att ta de prover man kan när de finns tillgängliga, och sedan lagra dem i en biobank till dess de kan undersökas.

För att sådana system ska kunna fungera krävs dock noggrann hantering av proverna. Det krävs att all metadata om provtagningen finns tillgänglig, såsom plats, tid, väderförhållanden, sättet proverna togs, sjukdomssymptom för kliniska prover, etc. Det krävs även att proverna är säkra medan de lagras och inte riskerar att tinas upp, slarvas bort, eller blandas ihop.

På ILRI, International Livestock Research Institute, i Nairobi, Kenya, finns en biobank – och för att garantera kvaliteten på proverna däri utvecklades ett system för övervakning av samtliga delar av biobankssystemet.

Systemet övervakar labb-frysar, inkubatorer, flytande kväve-frysarna som utgör biobanken, serverrummet, reservgeneratorn samt temperaturen i samtliga rum. Temperatursensorer kopplas till LabJack-mätningsmoduler via Ethernet, och avläses från en applikationsserver som kör egenprogrammerade script för att läsa av mätutrustningen. Mätningarna lagras sedan på en databasserver och jämförs med gränsvärden för varje system. Om något mätvärde går utanför gränsvärdena skickas Sms-meddelanden till berörd personal som sedan kan åtgärda felet innan några prover tar skada. Systemet har även webb-kameror som övervakar och sparar en logg av samtliga händelser i biobanken. Alla mätningar i databasen finns tillgängliga i ett webb-interface som kan nås på http://azizi.ilri.cgiar.org för statistik och uppdateringarna. Ett schema över systemet finns i *Figure 2* på sidan 26.

8.3 Ndumu-virus hittat i tamsvin i Uganda

Vid en projektundersökning i Uganda testades tamsvin i ett projekt om Afrikansk Svinfeber (ASF). ASF är en mycket dödlig sjukdom för tamsvin, som sprids av olika typer av afrikanska vildsvin vilka smittas, men överlever sjukdomen. I detta projekt användes metagenomisk DNA-sekvensering för att utvärdera vilka virus som fanns i proverna utöver det förväntade afrikansk svinfeberviruset.

Prover från tamsvin samlades in från flera områden i Uganda och DNA-sekvenserades. Sekvenserna undersöktes sedan och jämfördes mot kända sekvenser från publika databaser. Förutom förväntade sekvenser från gris, återfanns även ungefär 6% DNA från virus. I detta DNA hittade man från alla platser spår av afrikanskt svinfebervirus, samt från två provtagningsplatser Torque teno virus, en virusfamilj som tidigare återfunnits i grisar och som allmänt anses icke-sjukdomsframkallande. Från två provtagningsplatser återfanns även ett virus som såg ut att ha stor likhet till Ndumu Virus (NDUV), ett virus som tidigare bara hittats i insekter.

För att bekräfta att det var detta virus, återklassificerades alla DNA-fragment för att bekräfta att inga fragment var från insekts-DNA, vilket skulle tyda på att provet kunde vara kontaminerat. Inga sådana sekvenser återfanns dock. För att bekräfta ytterligare skapades sekvenseringsmarkörer från det DNA man hittat, och originalprovet sekvenserades från dessa markörer.

Resultaten bekräftade att det var ett Ndumu-virus i provet. Mänskliga antikroppar mot detta virus har påträffats i studier tidigare, men inga dödsfall har rapporterats. Möjligheten för virus att spridas mellan gris och människa gör dock att rapporter som denna alltid är viktiga på grund av potentialen att utvecklas mot ett sjukdomsframkallande virus.

8.4 Undersökning av parasitstammarna i ett levande vaccin

Theileria parva är en protozo-parasit (urdjur) som infekterar afrikansk buffel samt nötkreatur i östra, mellersta, och södra Afrika, där den årligen orsakar stora ekonomiska förluster för nötkreatursuppfödare. Parasiten sprids via fästingar, som infekteras då de dricker infekterat blod, och parasiten sprids sedan till nya djur genom fästingarnas saliv när de äter från nya djur. Parasiten har en förhållandevis invecklad livscykel som beskrivs närmare i *Figure 1* på sidan 15.

På 70-talet utvecklades ett levande vaccin för att kontrollera sjukdomen, kallat ”Muguga cocktailen”. Detta vaccin är en mix av infekterade fästingar som innehåller en potentiellt dödlig dos parasiter. Vaccinet åtföljs sedan av en stor dos långsamt verkande oxytetracyclin som begränsar sjukdomsförloppet och ger djuret möjlighet att återhämta sig och utveckla immunitet.

Levande vaccin har klara nackdelar dock. Bl.a. kräver det frysmöjligheter, det är dyrt att framställa, det måste hanteras av kunnig personal, och hanteras det fel kan det sprida parasiten istället för att skydda mot den. För att kompensera för detta gjorde vi en sekvenseringsstudie där de tre parasitstammar som ingår i vaccinet – Muguga, Serengeti-transformerad, och Kiambu5.

DNA från odlade parasiter sekvenserades och skillnaderna mellan Serengeti-transformerad- och Kiambu5-stammarna kartlades. Tanken är att om en mix av dessa tre parasitstammar ger särskilt bra skydd mot parasiten, så borde man kunna återskapa samma skydd, men bara använda de delar av parasiten som immunförsvaret reagerar på utan att behöva använda levande parasiter.

Den största upptäckten från denna studie var att Serengeti-transformerad-stammen, om skall vara en buffelparasit som anpassats till att infektera nötkreatur, och därmed bör vara förhållandevis olik Muguga-referensen, visade sig vara nästan exakt identisk. Detta tyder på att denna parasitstam någon gång blandats ihop med Muguga, och att fel stam har fått ingå i vaccinet. Misstankar om detta har framkommit även tidigare, men detta blir ett starkt bevis för denna slutsats. Vaccinet är dock fortfarande verkningsfullt, så endera finns något mycket viktigt bland de skillnader som faktiskt fanns, eller också är alla skillnader i Kiambu5-stammen.

Kiambu5-stammen hade tusentals skillnader mot Muguga, så det fanns inte möjlighet att kartlägga alla, men med statistiska metoder tog vi fram en lista på möjliga vaccinkandidat-gener att utvärdera i en möjlig framtida studie. En visuell representation av släktskapet mellan de olika *T. parva*-stammarna i denna studie finns i *Figure 3* på sidan 33.

8.5 MetLab, ett verktyg för metagenomiska undersökningar

Metagenomik är som tidigare nämnts ett bioinformatiskt avancerat, och förhållandevis nytt forskningsämne. Detta gör att många grupper som har frågeställningar som kan besvaras med metagenomiska undersökningar kan ha problem att planera sina studier. De flesta verktyg som har publicerats är även inriktade på bakteriell metagenomik. Bakteriell metagenomik är på vissa sätt enklare än viral metagenomik då bakterier har vissa så kallade ”markörgener”, som kan användas för att klassificera dem. Några sådana finns inte i virus. Det finns dock verktyg som kan arbeta med virala data.

Vad som började som en serie verktyg som utvecklades inom gruppen för att uppfylla våra krav, visade sig snart så pass användbara att vi valde att programmera ihop dem till ett samlat verktyg för att planera, simulera och klassificera virala metagenomiska data, som vi valde att kalla ”MetLab”, skrivet i script-språket Python2.7.

Programmet har tre delar - den första är en modul för att beräkna hur mycket sekvenseringsmaterial man behöver för att få en viss mängd data från alla organismer i ett metagenomiskt prov. Sannolikhetsläran för sådana beräkningar är avancerad och beräkningskrävande, så detta fick byggas som en särskilt modul i programmeringsspråket C, som är snabbare och minneseffektivare för beräkningar, samt med hjälp av ett särskilt kodbibliotek för extra stor beräkningsnoggrannhet.

Den andra modulen används för att simulera metagenomikförsök. Denna modul har två verktyg: först kan det användas för att beräkna en statistisk profil från riktiga sekvenseringsdata, sedan kan det använda dessa statistiska profiler för att simulera data med samma statistiska profil, men med kända virus som laddas ner från en genomdatabas på NCBI. Dessa simulerade data kan sedan användas för att testa bioinformatiska verktyg. Eftersom simulerade data kommer från kända organismer kan man därför testa ungefär hur bra resultat man kan förvänta sig från en riktig sekvensering.

Den sista delen av programmet används för att köra ”pipelines”, serier av program för metagenomisk bioinformatik. Programmet har i grunden tre av dessa pipelines, där alla börjar med kvalitetskontroll av in-data, och slutar med taxonomisk klassificering. De tre pipelines som finns skiljer sig en aning i mellanstegen dock. Pipelinen ”Animal sample” (prov från djur) filtrerar bort DNA från värddjuret innan klassificering, ”Environmental sample” (miljöprov) kör klassificering direkt, och ”Environmental sample with assembly” (miljöprov med assembly) försöker passa ihop DNA-fragmenten till större sammanhängande delas innan klassificering.

Systemet testades på simulerad och riktig data, och har för- och nackdelar, men vi tror att då det var användbart för oss kan det också vara det för andra grupper. Hela programmet är ”open source” (öppen källkod), så vem som helst får bygga vidare på det för att göra de förbättringar de önskar. Programmet och dess källkod finns att tillgå på https://github.com/norling/metlab.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol,* 215(3), pp. 403-10.

Attwood, T.K., Gisel, A., Eriksson, N.E. & Bongcam-Rudloff, E. (2011). Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective. *Bioinformatics-Trends and Methodologies. Rijeka: Intech Online Publishers*.

Banks, M., Bendall, R., Grierson, S., Heath, G., Mitchell, J. & Dalton, H. (2004). Human and porcine hepatitis E virus strains, United Kingdom. *Emerg Infect Dis,* 10(5), pp. 953-5.

Bishop, R., Geysen, D., Spooner, P., Skilton, R., Nene, V., Dolan, T. & Morzaria, S. (2001). Molecular and immunological characterisation of Theileria parva stocks which are components of the 'Muguga cocktail' used for vaccination against East Coast fever in cattle. *Vet Parasitol,* 94(4), pp. 227-37.

Blomstrom, A.L., Belak, S., Fossum, C., McKillen, J., Allan, G., Wallgren, P. & Berg, M. (2009). Detection of a novel porcine boca-like virus in the background of porcine circovirus type 2 induced postweaning multisystemic wasting syndrome. *Virus Res,* 146(1-2), pp. 125-9.

Burridge, M.J., Morzaria, S.P., Cunningham, M.P. & Brown, C.G. (1972). Duration of immunity to East Coast fever (Theileria parva infection of cattle). *Parasitology,* 64(3), pp. 511-5.

Chen, K. & Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol,* 1(2), pp. 106-12.

Chevreux, B., Wetter, T. & Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. In: *Proceedings of Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99*1999, pp. 45-56.

Dacheux, L., Cervantes-Gonzalez, M., Guigon, G., Thiberge, J.M., Vandenbogaert, M., Maufrais, C., Caro, V. & Bourhy, H. (2014). A preliminary study of viral metagenomics of French bat species in contact with humans: identification of new mammalian viruses. *PLoS One,* 9(1), p. e87194.

Daubenberger, C., Heussler, V., Gobright, E., Wijngaard, P., Clevers, H.C., Wells, C., Tsuji, N., Musoke, A. & McKeever, D. (1997). Molecular characterisation of a cognate 70 kDa heat shock protein of the protozoan Theileria parva. *Mol Biochem Parasitol,* 85(2), pp. 265-9.

De Paoli, P. (2005). Biobanking in microbiology: From sample collection to epidemiology, diagnosis and research. *FEMS Microbiology Reviews,* 29(5).

De Souza, Y.G. & Greenspan, J.S. (2013). Biobanking past, present and future: responsibilities and benefits. *AIDS,* 27(3), pp. 303-12.

Delwart, E.L. (2007). Viral metagenomics. *Rev Med Virol,* 17(2), pp. 115-31.

Di Giulio, G., Lynen, G., Morzaria, S., Oura, C. & Bishop, R. (2009). Live immunization against East Coast fever--current status. *Trends Parasitol,* 25(2), pp. 85-92.

Ebel, T., Middleton, J.F., Frisch, A. & Lipp, J. (1997). Characterization of a secretory type Theileria parva glutaredoxin homologue identified by novel screening procedure. *J Biol Chem,* 272(5), pp. 3042-8.

Edwards, R.A. & Rohwer, F. (2005). Viral metagenomics. *Nat Rev Microbiol,* 3(6), pp. 504-10.

Fousse, L., Hanrot, G., Lefévre, V., Pélissier, P. & Zimmermann, P. (2007). MPFR: A Multiple-Precision Binary Floating-Point Library with Correct Rounding. *ACM Transactions on Mathematical Software,* 33(2), pp. 13:1-13:15.

Gardner, M.J., Bishop, R., Shah, T., de Villiers, E.P., Carlton, J.M., Hall, N., Ren, Q., Paulsen, I.T., Pain, A., Berriman, M., Wilson, R.J., Sato, S., Ralph, S.A., Mann, D.J., Xiong, Z., Shallom, S.J., Weidman, J., Jiang, L., Lynn, J., Weaver, B., Shoaibi, A., Domingo, A.R., Wasawo, D., Crabtree, J., Wortman, J.R., Haas, B., Angiuoli, S.V., Creasy, T.H., Lu, C., Suh, B., Silva, J.C., Utterback, T.R., Feldblyum, T.V., Pertea, M., Allen, J., Nierman, W.C., Taracha, E.L., Salzberg, S.L., White, O.R., Fitzhugh, H.A., Morzaria, S., Venter, J.C., Fraser, C.M. & Nene, V. (2005). Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. *Science,* 309(5731), pp. 134-7.

Gauer, M., Mackenstedt, U., Mehlhorn, H., Schein, E., Zapf, F., Njenga, E., Young, A. & Morzaria, S. (1995). DNA measurements and ploidy determination of developmental stages in the life cycles of Theileria annulata and T. parva. *Parasitol Res,* 81(7), pp. 565-74.

George, J.E., Pound, J.M. & Davey, R.B. (2004). Chemical control of ticks on cattle and the resistance of these parasites to acaricides. *Parasitology,* 129 Suppl, pp. S353-66.

Ghosh, T.S., Mohammed, M.H., Komanduri, D. & Mande, S.S. (2011). ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformation,* 6(2), pp. 91-4.

Gomez-Alvarez, V., Teal, T.K. & Schmidt, T.M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J,* 3(11), pp. 1314-7.

Graham, S.P., Pelle, R., Honda, Y., Mwangi, D.M., Tonukari, N.J., Yamage, M., Glew, E.J., de Villiers, E.P., Shah, T., Bishop, R., Abuya, E., Awino, E., Gachanja, J., Luyai, A.E., Mbwika, F., Muthiani, A.M., Ndegwa, D.M., Njahira, M., Nyanjui, J.K., Onono, F.O., Osaso, J., Saya, R.M., Wildmann, C., Fraser, C.M., Maudlin, I., Gardner, M.J., Morzaria, S.P., Loosmore, S., Gilbert, S.C., Audonnet, J.C., van der Bruggen, P., Nene, V. & Taracha, E.L. (2006). Theileria parva candidate vaccine antigens recognized by immune bovine cytotoxic T lymphocytes. *Proc Natl Acad Sci U S A,* 103(9), pp. 3286-91.

Granberg, F., Vicente-Rubiano, M., Rubio-Guerri, C., Karlsson, O.E., Kukielka, D., Belak, S. & Sanchez-Vizcaino, J.M. (2013). Metagenomic detection of viral pathogens in Spanish honeybees: co-infection by Aphid Lethal Paralysis, Israel Acute Paralysis and Lake Sinai Viruses. *PLoS One,* 8(2), p. e57459.

Hayashida, K., Abe, T., Weir, W., Nakao, R., Ito, K., Kajino, K., Suzuki, Y., Jongejan, F., Geysen, D. & Sugimoto, C. (2013). Whole-genome sequencing of Theileria parva strains provides insight into parasite migration and diversification in the African continent. *DNA Res,* 20(3), pp. 209-20.

Henson, S., Bishop, R.P., Morzaria, S., Spooner, P.R., Pelle, R., Poveda, L., Ebeling, M., Kung, E., Certa, U., Daubenberger, C.A. & Qi, W. (2012). High-resolution genotyping and mapping of recombination and gene conversion in the protozoan Theileria parva using whole genome sequencing. *BMC Genomics,* 13, p. 503.

Hoff, K.J. (2009). The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics,* 10, p. 520.

Hugenholtz, P. & Tyson, G.W. (2008). Microbiology: metagenomics. *Nature,* 455(7212), pp. 481-3.

Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. (2007). MEGAN analysis of metagenomic data. *Genome Res,* 17(3), pp. 377-86.

Iams, K.P., Hall, R., Webster, P. & Musoke, A.J. (1990a). Identification of lambda gt11 clones encoding the major antigenic determinants expressed by Theileria parva sporozoites. *Infect Immun,* 58(6), pp. 1828-34.

Iams, K.P., Young, J.R., Nene, V., Desai, J., Webster, P., ole-MoiYoi, O.K. & Musoke, A.J. (1990b). Characterisation of the gene encoding a 104-kilodalton microneme-rhoptry protein of Theileria parva. *Mol Biochem Parasitol,* 39(1), pp. 47-60.

Kibe, M.K., Macklin, M., Gobright, E., Bishop, R., Urakawa, T. & ole-MoiYoi, O.K. (2001). Characterisation of single domain ATP-binding cassette protien homologues of Theileria parva. *Parasitol Res,* 87(9), pp. 741-50.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev,* 72(4), pp. 557-78, Table of Contents.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J. & International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature,* 409(6822), pp. 860-921.

Langmead, B. & Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods,* 9(4), pp. 357-9.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. & Higgins, D.G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics,* 23(21), pp. 2947-8.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25(14), pp. 1754-60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25(16), pp. 2078-9.

Luscombe, N.M., Greenbaum, D. & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med,* 40(4), pp. 346-58.

Mande, S.S., Mohammed, M.H. & Ghosh, T.S. (2012). Classification of metagenomic sequences: methods and challenges. *Briefings in Bioinformatics*.

Marcellino, W.L., Salih, D.A., Julla, I.I. & Hussein, A.M.E. (2011). Economic impact of east coast fever in central equatorial state of south Sudan. *International Research Journal of Agricultural Science and Soil Science,* 1(6), pp. 218-220.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet,* 24(3), pp. 133-41.

Mardis, E.R. (2013). Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif),* 6, pp. 287-303.

McKeever, D.J. & Morrison, W.I. (1990). Theileria parva: the nature of the immune response and its significance for immunoprophylaxis. *Rev Sci Tech,* 9(2), pp. 405-21.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* 20(9), pp. 1297-303.

Morzaria, S., Spooner, P., Bishop, R. & Mwaura, S. Live Vaccines for Theileria parva: Deployment in Eastern, Central and Southern Africa. In: *Proceedings of International Livestock Research Institute Proceedings*1999, pp. 56-61.

Mukhebi, A.W., Perry, B.D. & Kruska, R. (1992a). Estimated economics of theileriosis control in Africa. *Preventive Veterinary Medicine,* 12(1–2), pp. 73-85.

Mukhebi, A.W., Perry, B.D. & Kruska, R. (1992b). Estimated economics of theileriosis control in Africa. *Preventive Veterinary Medicine,* 12, pp. 73-85.

Musoke, A., Nene, V. & Morzaria, S.P. (1993). A Sporozoite-based vaccine for Theileria parva. *Parasitol Today,* 9(10), pp. 385-8.

Myers, K.P., Olsen, C.W. & Gray, G.C. (2007). Cases of swine influenza in humans: a review of the literature. *Clin Infect Dis,* 44(8), pp. 1084-8.

Pappas, G., Roussos, N. & Falagas, M.E. (2009). Toxoplasmosis snapshots: global status of Toxoplasma gondii seroprevalence and implications for pregnancy and congenital toxoplasmosis. *Int J Parasitol,* 39(12), pp. 1385-94.

Paradis, E., Claude, J. & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics,* 20(2), pp. 289-290.

Parashar, U.D., Sunn, L.M., Ong, F., Mounts, A.W., Arif, M.T., Ksiazek, T.G., Kamaluddin, M.A., Mustafa, A.N., Kaur, H., Ding, L.M., Othman, G., Radzi, H.M., Kitsutani, P.T., Stockton, P.C., Arokiasamy, J., Gary, H.E., Jr. & Anderson, L.J. (2000). Case-control study of risk factors for human infection with a new zoonotic paramyxovirus, Nipah virus, during a 1998-1999 outbreak of severe encephalitis in Malaysia. *J Infect Dis,* 181(5), pp. 1755-9.

Parks, D.H., MacDonald, N.J. & Beiko, R.G. (2011). Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics,* 12, p. 328.

Philbey, A.W., Kirkland, P.D., Ross, A.D., Davis, R.J., Gleeson, A.B., Love, R.J., Daniels, P.W., Gould, A.R. & Hyatt, A.D. (1998). An apparently new virus (family Paramyxoviridae) infectious for pigs, humans, and fruit bats. *Emerg Infect Dis,* 4(2), pp. 269-71.

Radley, D.E., Brown, C.G.D., Burridge, M.J., Cunningham, M.P., Kirimi, I.M., Purnell, R.E. & Young, A.S. (1975a). East coast fever: 1. Chemoprophylactic immunization of cattle against Theileria parva (Muguga) and five theilerial strains. *Veterinary Parasitology,* 1(1), pp. 35-41.

Radley, D.E., Brown, C.G.D., Cunningham, M.P., Kimber, C.D., Musisi, F.L., Payne, R.C., Purnell, R.E., Stagg, S.M. & Young, A.S. (1975b). East coast fever: 3. Chemoprophylactic immunization of cattle using oxytetracycline and a combination of theilerial strains. *Veterinary Parasitology,* 1(1), pp. 51-60.

Radley, D.E., Young, A.S., Brown, C.G.D., Burridge, M.J., Cunningham, M.P., Musisi, F.L. & Purnell, R.E. (1975c). East coast fever: 2. Cross-immunity trials with a Kenya strain of Theileria lawrencei. *Veterinary Parasitology,* 1(1), pp. 43-50.

Riegman, P.H., Morente, M.M., Betsou, F., de Blasio, P., Geary, P. & Marble Arch International Working Group on Biobanking for Biomedical, R. (2008). Biobanking for better healthcare. *Mol Oncol,* 2(3), pp. 213-22.

Rodriguez, R.L. & Konstantinidis, K.T. (2014). Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics,* 30(5), pp. 629-35.

Rosen, G.L., Reichenberger, E.R. & Rosenfeld, A.M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics,* 27(1), pp. 127-9.

Rosseel, T., Van Borm, S., Vandenbussche, F., Hoffmann, B., van den Berg, T., Beer, M. & Hoper, D. (2013). The origin of biased sequence depth in sequence-independent nucleic acid amplification and optimization for efficient massive parallel sequencing. *PLoS One,* 8(9), p. e76144.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74(12), pp. 5463-7.

Schmieder, R. & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics,* 27(6), pp. 863-4.

Shaw, M.K. (2003). Cell invasion by Theileria sporozoites. *Trends Parasitol,* 19(1), pp. 2-6.

Shaw, M.K., Tilney, L.G. & Musoke, A.J. (1991). The entry of Theileria parva sporozoites into bovine lymphocytes: evidence for MHC class I involvement. *J Cell Biol,* 113(1), pp. 87-101.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology,* 26(10), pp. 1135-1145 %@ 1087-0156.

Skilton, R.A., Bishop, R.P., Wells, C.W., Spooner, P.R., Gobright, E., Nkonge, C., Musoke, A.J., Macklin, M. & Iams, K.P. (1998). Cloning and characterization of a 150 kDa microsphere antigen of Theileria parva that is immunologically cross-reactive with the polymorphic immunodominant molecule (PIM). *Parasitology,* 117 (Pt 4), pp. 321-30.

Skilton, R.A., Musoke, A.J., Wells, C.W., Yagi, Y., Nene, V., Spooner, P.R., Gachanja, J., Osaso, J., Bishop, R.P. & Morzaria, S.P. (2000). A 32 kDa surface antigen of Theileria parva: characterization and immunization studies. *Parasitology,* 120 (Pt 6), pp. 553-64.

Smith, T.C., Harper, A.L., Nair, R., Wardyn, S.E., Hanson, B.M., Ferguson, D.D. & Dressler, A.E. (2011). Emerging swine zoonoses. *Vector Borne Zoonotic Dis,* 11(9), pp. 1225-34.

Stanhope, S.A. (2010). Occupancy modeling, maximum contig size probabilities and designing metagenomics experiments. *PLoS One,* 5(7), p. e11652.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol,* 28(10), pp. 2731-9.

Team, R.C. (2013). *R: A Language and Envronment for Statistical Computing*. [Computer Program]. Available from: http://www.R-project.org/.

Thomas, T., Gilbert, J. & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp,* 2(1), p. 3.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y. & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet,* 30(9), pp. 418-426.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. (2001). The sequence of the human genome. *Science,* 291(5507), pp. 1304-51.

Wendl, M.C., Kota, K., Weinstock, G.M. & Mitreva, M. (2013). Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. *J Math Biol,* 67(5), pp. 1141-61.

Willner, D. & Hugenholtz, P. (2013). From deep sequencing to viral tagging: recent advances in viral metagenomics. *Bioessays,* 35(5), pp. 436-42.

Wommack, K.E., Bhavsar, J. & Ravel, J. (2008). Metagenomics: read length matters. *Appl Environ Microbiol,* 74(5), pp. 1453-63.

Wood, D.E. & Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol,* 15(3), p. R46.

Wooley, J.C., Godzik, A. & Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput Biol,* 6(2), p. e1000667.

Xue, G., von Schubert, C., Hermann, P., Peyer, M., Maushagen, R., Schmuckli-Maurer, J., Butikofer, P., Langsley, G. & Dobbelaere, D.A. (2010). Characterisation of gp34, a GPI-anchored protein expressed by schizonts of Theileria parva and T. annulata. *Mol Biochem Parasitol,* 172(2), pp. 113-20.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol,* 24(8), pp. 1586-91.

Acknowledgements

Absolomon Kihara, for teaching me a lot about PHP, and for being as excited about databases as I am.

Alan Orth, for teaching me how to deal with server administration, and how to be an mjanja.

Emelie Zonabend, for helping me find my way out of the jungle.

Erik Bongcam-Rudloff for being my supervisor, and teaching me about photography.

Etienne deVilliers, for being my co-supervisor in Kenya, and for showing me the Mangrove swamps.

Göran Andersson, for being my co-supervisor, and for great lectures.

Hans-Henrik Fuxelius, for showing me what bioinformatics meant, and what it means to be a Doctor.

Harry Noyes, for first making me feel welcome at ILRI and introducing me to the group.

Juliette Hayer, for staying at work till 3am to help me meet my deadlines.

Karl Ståhl, for his incredible kindness and enthusiasm – you make me feel like science is *fun*.

Katarina Truvé, for support and making me understand the importance of making tools available to the public.

Margaret MacDonald, for being my neighbor at ILRI, and keeping me alive through disease and bureaucracy.

Oscar Eriksson, for all the help with technical things that needed a 'real' engineer.

Oskar Karlsson, for reading way too many papers, being way too excited, and way too nice to me.

Richard Bishop, for a constantly surprising knowledge about *T. parva*, and for making me feel like I went to Woodstock.

Steve Kemp for teaching me the importance of rigor, monitoring, and meta-data.

Sara Norling, for being my constant support, and constant source of inspiration.

This work was funded by the following grants:

SIDA: SWE-2011-154, "Using next generation sequencing to support the development of infection and treatment vaccination methods in East Africa". 2012-2013

FORMAS: 2012-586, "Viral Metagenomics and Bioinformatics as powerful novel tools in veterinary infection biology". 2013-2015