CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

UNIDAD IRAPUATO
UNIDAD DE GENÓMICA AVANZADA

Interacciones entre especies mediadas por RNAs pequeños

Tesis que presenta

MC José Roberto Bermúdez Barrientos

Para obtener el grado de

Doctor en Ciencias en Biología Integrativa

Director de tesis:
Dr. Cei Leander Gastón Abreu Goodger

Irapuato, Guanajuato
Agosto 2020

CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

UNIDAD IRAPUATO
UNIDAD DE GENÓMICA AVANZADA

Species interactions mediated by small RNAs

Presented by:

MSc José Roberto Bermúdez Barrientos

To attain the degree of

Doctor of Philosophy in Integrative Biology

Thesis advisor:
Dr. Cei Leander Gastón Abreu Goodger

Irapuato, Guanajuato August 2020

Este trabajo se realizó en el Centro de Investigación y de Estudios Avanzados del IPN - Unidad Irapuato.

Bajo la dirección del Dr. Cei Abreu Goodger en el Laboratorio de Genómica Computacional del RNA perteneciente a la Unidad de Genómica Avanzada.

En colaboración con la Dra. Amy Buck perteneciente a la Universidad de Edinmburgo.

Gracias al apoyo del Consejo Nacional de Ciencia y Tecnología (CONACyT) a través de los Proyectos CB-2016-01-284884, y la beca 434580. También al donativo RGY0069/2014 de la Human Frontiers Science Program.

Este trabajo se realizó del 1 de marzo de 2016 al 1 de junio de 2020.

Miembros del comité de asesores:
Dr. Alfredo Heriberto Herrera Estrella, Unidad de Genómica Avanzada, CINVESTAV Irapuato.
Dra. Laila Pamela Partida Martínez, Departamento de Ingeniería Genética, CINVESTAV Irapuato.
Dr. Rafael Montiel Duarte, Unidad de Genómica Avanzada, CINVESTAV Irapuato.
Dra. Selene Lizbeth Fernández Valverde, Unidad de Genómica Avanzada, CINVESTAV Irapuato.
Unidad de Genómica Avanzada, CINVESTAV Irapuato.

Miembros externos:
Dra. Paula Licona Limón, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México.

Fecha del examen de grado: 17 de Agosto de 2020.

Acknowledgements

To my advisor, Dr. Cei Abreu for his guidance and having such an impact in my academic and personal formation.

To my committee members: Dr. Laila P. Partida Martínez, Dr. Selene L. Fernández Valverde, Dr. Alfredo Herrera Estrella, Dr. Rafael Montiel Duarte and Dr. Paula Licona for their valuable comments, discussions and suggestions.

To Kyriaki Neophytou and Franklin Chow for reading and commenting this text.

To all current and past members of the RNAlab, especially to Obed Ramírez, Joel Rodriguez, Francisco Medina, Pablo González, Isaac Martínez and Cesaré Ovando.

To Franklin Chow, Kyriaki Neophytou, Amy Buck and her laboratory members for their collaborative spirit.

To my family for all their love and support.

To Paulina Guevara, thank you for everything.

To CONACYT for the funding.

Table of Contents

Acknowledgements ***4***

List of Abreviations ***8***

Abstract ***9***

Resumen ***10***

Preface ***12***

Introduction ***12***

- **RNA interference and small RNAs** **12**
- **Extracellular RNA and vesicles** **15**
- **Communication between different species via small RNAs** **18**
- **The immune system** **21**
  - Innate immune system 21
  - Adaptive immune system 22
- ***Heligmosomoides bakeri*** **24**
- **Immune response to *Heligmosomoides*** **26**
  - Intestinal epithelial cells (IECs) 26
  - Macrophages 28
- ***Heligmosomoides* excretory-secretory (HES) & immune modulation** **29**

Background ***30***

- ***H. bakeri* secretes EVs that have immune-regulatory capacity and harbor sRNAs** **30**
- ***H. bakeri* EVs suppress macrophage activation** **31**
- ***H. bakeri* secretes an Argonaute protein that loads 22Gs that originate from repetitive sequences** **32**

Chapter 1: Disentangling sRNA-Seq data to study RNA communication between species ***34***

- **Methods** **34**
  - Selected experiments and reference genomes 34
  - Processing of small RNA-Seq reads 34
  - Calculations of host, symbiont and ambiguous reads 34
  - Shared k-mers between genomes 35
  - Genome-guided sRNA assembly 35
  - De novo assembly of sRNA-Seq 36
  - Disambiguation of host-symbiont mixed samples 36
- **Results** **38**
  - Species systems that exchange small RNAs included in this study 38
  - Determining the amount of host, symbiont, and ambiguous reads in sRNA datasets 39

Ambiguity in host-symbiont sRNA-Seq reads is influenced by read length, genome size and phylogenetic distance 41
miRNAs, transfer and ribosomal RNAs are major contributors to ambiguous reads 44
Small RNA sequencing assembly 48
sRNA assembly decreased the number of ambiguous reads for all symbiosis pair datasets 49

Conclusions 51

Perspectives 51

Chapter 2: Detection of Heligmosomoides bakeri sRNAs in mouse cells 53

Methods 53

H. bakeri life cycle and EV isolation 53
Mouse cells uptake assays 53
Small RNA library preparation and sequencing 54
Genome-guided assembly 55
Differential expression analysis 55
Defining sRNA classes by length and fist nucleotide 55
Extracting individual sequences from differentially expressed *H. bakeri* clusters 56
Genomic origin for *H. bakeri* sRNAs 56
Expression comparison of *H. bakeri* detected clusters in mouse 56
Genome features visualization 56

Results and discussion 56

Our experimental design 57
Our clusters 57
Differential expression analysis aids detection of parasite sRNAs in mouse cells 59
Validation of differentially expressed parasitic sRNAs 64
Genomic origin of *H. bakeri* sRNAs detected in mouse cells 66
Comparison of *H. bakeri* detected clusters in mouse cells 70

Conclusions 77

Perspectives 77

Chapter 3: Predicted host targets for Heligmosomoides bakeri sRNAs 78

Methods 78

Non-redundant Hb-sRNAs 78
End-to-end Hb-sRNAs target predictions in host genes 78
Reward summed score 78

Results and discussion 79

Reducing redundant targets 79
End-to-end Hb-sRNAs target prediction in host 81

Conclusions 87

Perspectives 87

Chapter 4: RNA-Seq downregulation of predicted targets for Heligmosomoides sRNAs 89

Methods **89**
Read quality and adapter trimming 89
Gene expression quantification 89
Differential expression analysis 90
Test for downregulation of Hb-sRNAs target predictions in host 90
Results and discussion **91**
Effects of HES and EVs on mouse cells 91
Are *H. bakeri* small RNAs responsible for HES or EV effects? 98
Conclusions **103**
Perspectives **104**
References ***104***

List of Abreviations

Here I provide a list of commonly used abreviations in the present work. This is not an exhaustive list, and doesn't contain abreviations that were used only a couple times in the whole work.

22G. Reads that are 21-24 nt long whose first 5' nucleotide tends to be a guanine, these are RdRP products in *C. elegans* and close relatives.
AAM. Alternative macrophage activation.
Ago. Protein of the Argonaute family.
BMDM. Bone marrow-derived macrophages.
DE. Differential expression.
DEA. Differential expression analysis.
EV. Extracellular Vesicle.
FDR. False Discovery Rate.
Hb-sRNAs. *Heligmosomoides bakeri* small RNAs.
HES. *Heligmosomoides* excretory-secretory product.
IECs. Intestinal epithelial cells.
IL. Interleukin.
logFC. Logarithm base-2 of the fold change.
LPS. Bacterial lipopolysaccharide.
MDS. Multidimensional scaling.
MHC. Major histocompatibility complex.
miRNA. microRNA.
Mono-P. Small RNA samples that didn't received a phosphatase treatment prior to library construction and sequencing.
Poly-P. Small RNA samples that were treated with phosphatase prior to library construction and sequencing.
RdRp. RNA dependent RNA polymerase.
RNAi. RNA interference.
RNA-Seq. RNA sequencing.
rRNA. Ribosomal RNA.
sRNA. Small RNA.
sRNA-Seq. Small RNA sequencing.
TGF-β. Transforming growth factor beta.
T_h1. Type 1 T helper cell that promotes cell mediated immune responses.
T_h2. Type 2 T helper cell that promotes humoral immune responses.
tRNA. Transfer RNA.
WAGO. Worm-specific Argonaute protein.
yRNA. Are small non-coding RNAs that are part of the human Ro60 ribonucleoprotein particle.

Abstract

Small RNAs (sRNA) play a key role regulating target genes. Several studies have discovered that sRNAs may be used as communication molecules between organisms of different species. *Heligmosomoides bakeri* is a parasitic nematode that spends part of its lifecycle in the mouse intestine and can establish long-term infections. To do so, this nematode modulates its host immune response with a secretion called *Heligmosomoides* excretory-secretory products (HES). We have previously found extracellular vesicles (EVs) with immune modulatory capacity in HES. Both EVs and HES harbor sRNAs (Hb-sRNAs), but it's unclear if Hb-sRNAs are relevant for infection. We know that *H. bakeri* EVs are capable of getting into mouse cells, but it is unknown if these deliver their sRNA cargo.

Here we detect Hb-sRNAs in mouse cells and test whether Hb-sRNAs repress their predicted mouse mRNA targets. To achieve this, we first developed strategies to disentangle sRNA sequencing data (sRNA-Seq) from samples containing information of two interacting organisms. These strategies involved a combination of sRNA assembly and differential expression analysis. We applied the assembly approach to a diverse set of six pairs of organisms that communicate via sRNAs, representing animals, plants, fungi and bacteria. In all cases, sRNA assembly reduced the number of ambiguous sequences between host and symbionts compared to a baseline approach. Organism-of-origin determination is a challenging step for those reads that map perfectly to either genome.

We then applied our strategies to an *in vitro* experiment in which we incubated intestinal epithelial cells and macrophages with EVs or HES. The purified RNA was divided into two, in order to sequence sRNAs for detecting Hb-sRNAs in host cells; and to sequence mRNAs (RNA-Seq) to assess the repression capacity of Hb-sRNAs on host transcripts. I was able to detect more than 16,000 Hb-sRNAs in mouse cells, representing 0.18% of the 8.4 M unique sequences produced by the adult nematode. The condition with the strongest Hb-sRNA signal was the intestinal epithelial cells with EVs. Noteworthy, I found that 95.5%-99.6% of sRNAs reads that map to the parasite genome clearly behave as mouse sRNA genes according to our differential expression analysis. This indicates that mapping information alone is not sufficient to assign sRNA reads to an organism.

I predicted host targets for the Hb-sRNAs found inside mouse cells using an end-to-end alignment approach to get strict predictions. I then tested for an overall effect of Hb-sRNAs on mouse transcripts using RNA-Seq and I was able to detect a slight effect of Hb-sRNAs on their best 384 predicted targets in both analyzed cell types. I also analyzed the effect of *H. bakeri* secretion products on host transcripts and found differential responses for some genes due to EVs or HES. The transcripts for interleukin 33 (IL-33), a cytokine secreted by intestinal epithelial cells during *H. bakeri* infections, are repressed in intestinal epithelial cells with HES but not with the EV treatment. Interestingly, the IL-33 receptor is also repressed only with HES, but not EV treatment. We provide a valuable strategy for organism assignment for the sRNA species communication field. My results support an effect of Hb-sRNAs on host gene expression, although additional experiments are needed to test the relevance of Hb-sRNAs for the immune modulation activity of *H. bakeri* secretion products.

Resumen

Los RNAs pequeños (sRNAs) juegan un papel clave en la regulación de sus genes blanco. Recientemente se ha descubierto que los sRNAs pueden fungir como agentes de comunicación entre organismos de diferentes especies. *Heligmosomoides bakeri* es un nematodo parásito que pasa parte de su ciclo de vida en los intestinos de ratón y puede establecer infecciones a largo plazo. Para lograrlo dicho nematodo modula la respuesta inmune de su hospedero mediante lo que se conoce como productos de excreción-secreción de *Heligmosomoides* (HES). Hemos encontrado previamente vesículas extracelulares (EVs) con capacidad inmuno-modulatoria en HES. Tanto las vesículas como HES tienen sRNAs (Hb-sRNAs), pero no queda claro si los Hb-sRNAs son relevantes para la infección. Sabemos que las vesículas de *H. bakeri* son capaces de entrar a células de ratón, pero desconocemos si son capaces de liberar sus sRNAs.

En este trabajo detectamos Hb-sRNAs en células de ratón y ponemos a prueba si los Hb-sRNAs reprimen sus blancos predichos en mRNAs de ratón. Para ello, primero desarrollamos estrategias para separar datos de secuenciación de sRNAs (sRNA-Seq) de muestras que contengan información de dos organismos interactuando. Nuestras estrategias involucran una combinación de ensamblaje de sRNAs y un análisis de expresión diferencial. Nosotros evaluamos la aproximación de ensamblaje de sRNAs en un grupo diverso de seis pares de organismos que se comunican mediate sRNAs. En todos los casos, las técnicas de ensamblaje de sRNAs reducen el número de secuencias ambiguas entre el hospedero y su simbionte en comparación con una aproximación sin ensamblaje. La determinación del organismo de origen es particularmente difícil para aquellas lecturas que alinean perfectamente a ambos genomas.

Posteriormente, apliqué nuestras estrategias a un experimento *in vitro* de células de epitelio intestinal o macrófagos incubadas con EVs o HES. El RNA usado para este experimento se separó en dos para realizar la secuenciación de sRNAs (sRNA-Seq) con el objetivo de detectar Hb-sRNAs en las células del hospedero, y para llevar a cabo la secuenciación de mRNAs (RNA-Seq) con el fin de evaluar la capacidad de represión de los Hb-sRNAs en transcritos del hospedero. Encontré más de 16,000 Hb-sRNAs en células de ratón, lo cual representa el 0.18% de las 8.4 M de secuencias únicas producidas por el nematodo adulto. La condición con la señal más fuerte de Hb-sRNAs fue las células de epitelio intestinal con EVs. Notablemente, encontré que del 95.5%-99.6% de los sRNAs que alinean perfectamente al genoma del parásito se comportan como genes de ratón de acuerdo a un análisis de expresión diferencial. Esto nos indica que la información de alineamientos no es suficiente para asignar lecturas de sRNAs a uno u otro organismo.

Realicé predicciones de blancos en el hospedero para los Hb-sRNAs encontrados dentro de las células de ratón usando una aproximación de extremo a extremo para obtener prediciones estrictas. Posteriormente puse a prueba si los Hb-sRNAs tienen un efecto global en transcritos de ratón usando RNA-Seq. De esta manera pude detectar un efecto sutil de los Hb-sRNAs en sus mejores 384 blancos predichos para ambos tipos celulares. También analicé el efecto de los productos secretados de *H. bakeri* en transcritos del hospedero y encontré respuestas diferentes para algunos genes debido a EVs o HES. Los transcritos de la interleucina 33 (IL-33), una citoquina secretada por las células de epitelio intestinal durante las infecciones de *H. bakeri*, se reprime debido al tratamiento con HES, pero no así con EVs. En este trabajo aportamos una estrategia valiosa para asignación de organismos para el campo

de comunicación entre especies mediada por sRNAs. Mis resultados apoyan la hipótesis de un efecto de los Hb-sRNAs en la expresión del hospedero, aunque se requieren experimentos adicionales para poner a prueba la relevancia de los Hb-sRNAs para la actividad inmuno-moduladora de los productos secretados de *H. bakeri*.

Preface

In the Introduction I describe basic concepts required to understand my PhD work, such as RNAi, small RNAs (sRNAs), extracellular vesicles, communication between organisms of different species, and the immune system. I also introduce the parasitic nematode *Heligmosomoides bakeri*, our model system for sRNA-mediated communication. In Chapter 1 I describe strategies we developed to disentangle sRNA-Seq data of host and pathogen interactions. In Chapter 2 I describe the application of these strategies to the detection of sRNAs of a nematode parasite in an *in vitro* mouse cell-line experiment. Chapter 1 and part of Chapter 2 were published in *Nucleic Acids Research* journal under the title “Disentangling sRNA-Seq data to study RNA communication between species” (Bermúdez-Barrientos et al. 2020). In Chapter 3 I predict targets for the nematode small RNAs detected in mouse cells and discuss some of the difficulties associated to small RNA target prediction. In Chapter 4 I assess the effect of *H. bakeri* secretion products on mouse cells and test for an effect of detected nematode small RNAs on host gene expression.

Introduction

RNA interference and small RNAs

RNA interference (RNAi) is a biological process where double stranded RNA triggers the production of small RNAs that target and regulate the expression of complementary RNA transcripts. Small RNAs are 18-30 nt RNA fragments that do not code for protein and are used as guides by an Argonaute protein to target a transcript in a sequence-specific manner. Other processes that involve RNAi and have been used in the literature include co-suppression (plants) (Napoli, Lemieux, and Jorgensen 1990), post-transcriptional gene silencing (plants and animals such as *Drosophila*) (Hamilton and Baulcombe 1999) and quelling (fungi), a name derived from quell which refers to supression (Villalobos-Escobedo, Carreras-Villaseñor, and Herrera-Estrella 2016). All these processes (co-supression, post-transcriptional gene silencing an quelling) involve the production of sRNAs from a double stranded RNA and the silencing of homologous sequences. The RNA interference pathway (RNAi) has a widespread distribution across the eukaryote domain, although some clades such as *Saccharomyces* sensu stricto complex (fungi, Ascomycota), *Ustilago* (fungi, Basidiomycota), *Leishmania* (protozoa, trypanosomatid) and *Plasmodium falciparum* (protozoa, Apicomplexa) have lost most or all of the RNAi machinery (Nicolas, Torres-Martinez, and Ruiz-Vazquez 2013).

The ancestral function of the RNAi machinery was likely to protect the germline from selfish ribonucleic elements such as viruses and transposable elements (Obbard et al. 2009). The RNAi pathway was later adapted to process endogenous dsRNAs, e.g. microRNAs, and help drive organism development, by exaptation. Exaptation is the usage of a given structure or process for a different function or need that it originally evolved for, the classical example are the usage of feathers to fly when they probably originated to provide thermal regulation (Gould, Stephen Jay; Vrba 1982).

The core enzymatic players for the RNAi pathway are RNAse III Dicer-like proteins, piwi and PAZ domain-containing Argonaute (Ago) proteins and RNA-dependent RNA Polymerases (RdRPs) (Shabalina and

Koonin 2008). Argonaute and Dicer homologs show a wider distribution across eukaryotes including animals, plants, fungi and some protozoa. On the other hand, RdRPs are more phylogenetically scattered, mostly occurring in plants, but also in some nematodes, insects and fungi (Tomoyasu et al. 2008).

I will describe the RNAi pathway as it functions in animals. In RNAi, Dicer recognizes double-stranded RNA molecules and cleaves them into ~20 nt products that are loaded into Ago proteins. These dsRNA trigger molecules may originate from exogenous or endogenous sources.

MicroRNAs (miRNAs) are the most famous source for endogenous dsRNA that triggers RNAi. These miRNAs are encoded in the genome and are transcribed by RNA polymerase II. The primary transcript (pri-miRNA) folds itself forming one or more hairpin structures (He and Hannon 2004). Animal miRNAs are first processed by Drosha, which cuts the hairpin from the pri-miRNA, which is then called the pre-miRNA (He and Hannon 2004). Drosha processing occurs in the nucleus, and the pre-miRNA leaves the nucleus via Exportin 5 (He and Hannon 2004). In the cytoplasm, Dicer cleaves the pre-miRNA hairpin to produce a mature (guide strand) miRNA and a complementary sequence (passenger strand) (He and Hannon 2004). The guide strand is then loaded into an Argonaute protein (ALG-1 in *C. elegans*) and this complex is called the RNA-induced silencing complex (RISC) (He and Hannon 2004). The seed sequence of a miRNA corresponds to positions 2 to 8 from the 5' end. This region is key for the RISC complex-target transcript interaction, especially in animals. Groups or families of miRNAs can be defined according to their seed sequence, and members of a family share many of their targets. Animal miRNAs also tend to bind to the 3' UTR regions of target mRNAs. The most prevalent miRNA repression mechanism in animals is mRNA destabilization via deadenylation and subsequent mRNA decapping (66-90% of repressive effect), while translation inhibition also contributes but to a lesser extent (Eichhorn et al. 2010). Destabilization or translation inhibition of the mRNA can happen even with low levels of sequence complementarity. In some cases, perfect sequence complementarity in the seed region may be enough to exert inhibition.

In 2006 Craig Mello and Andrew Fire received the Nobel prize in physiology or medicine for their work elucidating the components of RNAi biogenesis in the model nematode *Caenorhabditis elegans* (Fire et al. 1998). I will describe some of the RNAi features for *C. elegans* as this nematode shares a great deal of its sRNA biology with *Heligmosomoides bakeri,* the main focus of my thesis.

C. elegans is an organism that has remarkable capabilities in terms of its RNAi biology(Youngman and Claycomb 2014). This nematode is capable of uptaking dsRNA from its environment and silencing endogenous genes that are complementary to the internalized molecules (environmental RNAi) (W. M. Winston et al. 2007). This silencing effect may spread through different parts of the nematode body, a phenomenon termed systemic RNAi (W. M. Winston, Molodowitch, and Hunter 2002). Additionally, small initial amounts of dsRNA may trigger an organimal silencing effect, which corresponds to RNAi signal amplification (Fire et al. 1998). Finally, *C. elegans* has the potential to transfer silencing effects, guided by sRNAs, to its offspring through a phenomenon named transgenerational inheritance (K. C. Brown and Montgomery 2017). This suite of capacities can be attributed to a multitude of protein components of the RNAi machinery that is accompanied by a complex endogenous sRNA landscape (Kim et al. 2005).

Nematodes experienced an expansion of Ago proteins. For example, *Caenorhabditis elegans* has 25 Argonaute proteins while humans have only 8. Out of these 25 Argonaute proteins, 18 form a distinct clade exclusive of nematodes, which are collectively named WAGO (worm-Argonaute) (Claycomb 2014). The great diversity in Argonaute proteins in nematodes has been associated with an increased diversity of functions for the RNAi pathway (**Table 1**) (Buck and Blaxter 2013). In the Background section, I will introduce a WAGO that is secreted by the parasitic nematode *Heligmosomoides bakeri* and we present evidence about its possible role in parasitism.

ALG-1 (Argonaute plant-Like Gene) homologs are widely conserved across organisms containing the RNAi pathway (Tops, Plasterk, and Keiting 2006). These proteins use miRNAs as guides and participate in several endogenous processes such as embryonic development, metabolism and cell fate. PRG-1 (Piwi Related Gene) binds to piRNAs and these complexes are responsible for protecting the germ line from transposable elements (Batista et al. 2008). HRDE-1 (Heritable RNAi Deficient) is an Argonaute protein that displays nuclear localization and drives transgenerational epigenetic memory (Buckley et al. 2012). ERGO-1 (Endogenous-RNAi deficient arGOnaute) is expressed in the oogenic gonad that binds 26G sRNAs (26 nucleotides long, starting with a G), its targets are depleted from conserved genes and this Ago protein presumably buffers the expression of new genes (Gent et al. 2010). CSR-1 (Chromosome-Segregation and RNAi deficient 1) binds 22G sRNAs (described below) and its targets are enriched in germline-expressed genes such as those required for early developmental stages. CSR-1 is a very interesting Ago protein as it promotes gene expression instead of repression, and it also participates in chromosome segregation. A current model suggests that CSR-1 and PRG-1 surveil the germline expression with opposite outcomes: expression for CSR-1 targets and repression for PRG-1 targets (Almeida and Andrade-navarro 2019).

Table 1. Selected examples of different functions performed by different Argonaute proteins in *C. elegans.*

Argonaute	sRNA size and 1st nucleotide	Function
ALG-1	22U	miRNAs binding, endogenous process regulation
PRG-1	21U	piRNAs binding, silencing of transposable elements in germ line
HRDE-1	22G	Transgenerational inheritance of germline RNAi
ERGO-1	26G	Putative buffering of expression of newly acquired genes
CSR-1	22G	Licensing of expression in germline and chromosome segregation

C. elegans displays RNAi signal amplification. This phenomenon is initiated by primary siRNAs that are derived from double stranded RNA (dsRNA) and Dicer activity. Secondary siRNAs contribute to a signal amplification of repression (**Figure 1**). Secondary siRNAs are synthetized by RNA-dependent RNA polymerases (RdRPs) (Sijen et al. 2007). RdRPs in nematodes such as *C. elegans* are non-processive and produce short ~22 nt sRNAs that begin with a guanine (so they are known as 22G), that are loaded into WAGOs. As their name suggests, RdRPs synthetize RNA from another template RNA. Since this synthesis is primer-independent, the first nucleotide gets incorporated directly as a triphosphate nucleotide, resulting in small RNAs with a 5’ tri-phosphate. Small RNA-Seq libraries that were phosphatase-treated will be referred as poly-P, libraries that didn’t received this phosphatase treatment will be referred as mono-P through this work.

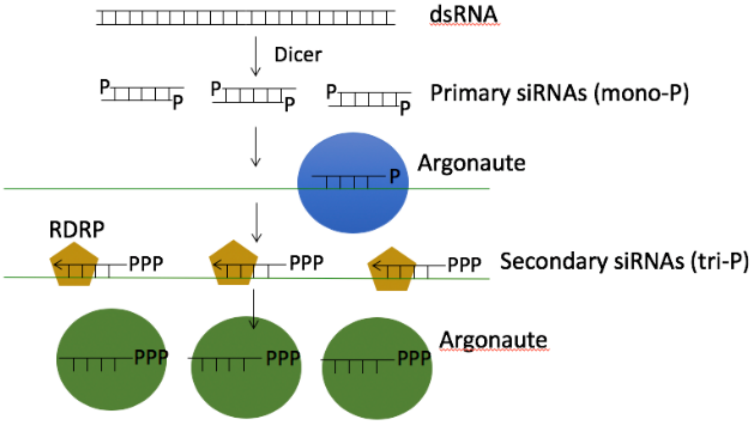


Figure 1. Secondary siRNAs in *C. elegans* contribute to the amplification of the repression signal.

Some readers may be more familiar with plant RdRPs, and there are differences between RdRPs in plants and nematodes. In plants, RdRPs such as RDR6 produce a long double stranded product that is later processed by DCL4 (Dicer-like 4) or DCL5 into individual siRNAs. As a result of cutting a long dsRNA into smaller 21-24 chunks, these will have 5’ mono-phosphate ends (Z. Xie et al. 2005).

In *C. elegans*, SID-2 (systemic RNAi defective-2) is an RNA transporter that is required for environmental RNAi and is expressed in the intestinal epithelium cells (W. M. Winston et al. 2007). The phylogenetic distribution of SID-2 is limited: it is not even present in *C. briggsae* (*C. briggsae* and *C. remanei* are the closest relatives to *C. elegans*) (Kiontke, Karin; Fitch 2005). SID-1 is more widely distributed, with homologs found in *C. elegans* (nematode), *Diabrotica virgifera* (arthropoda) (Ivashuta et al. 2015), and the mammalian genome (chordata) suggesting that SID-1 may have been present in the metazoan last common ancestor. SID-1 is an RNA transporter required for RNA movement between cells and for a systemic RNAi response in nematodes (W. M. Winston, Molodowitch, and Hunter 2002). *H. bakeri* lacks a SID-2 homolog, suggesting that it is unable to perform environmental RNAi, but it does contain a SID-1 homolog (Chow et al. 2019).

Extracellular RNA and vesicles

Many people who have worked with RNA in the laboratory may be under the impression that this molecule is rather unstable, and thus unsuitable for extracellular communication. However, extracellular RNA is common: miRNAs have been found in all human fluids analyzed so far (Yuana, Sturk, and Nieuwland 2013). This RNA is protected from RNAse activity by encapsulation in extracellular vesicles (EVs) and/or association with Argonaute proteins. In 2007, Valadi and collaborators (Valadi et al. 2007) showed that EVs can transfer nucleic acids between mast cells. They observed transfer of miRNAs and mRNAs, and the transferred mRNAs could even be translated in recipient cells.

Extracellular vesicles (EVs) are lipid bilayer-delimited particles that are naturally released from many kinds of cells but, unlike cells, cannot replicate (Witwer and Théry 2019). The study of EVs dates back to the early 1980's. In 1983 Pan and Johnstone observed multivesicular bodies (MVB), structures involved in exosome biogenesis (Pan and Johnstone 1983). Since then, EVs have become a whole research field: the International Society for Extracellular Vesicles (ISEV) was founded in 2011 and the Journal of Extracellular Vesicles was founded in 2012. EVs have attracted so much attention that in 2012 the National Institutes of Health (NIH) announced a program funding EV and extracellular RNA studies: the Extracellular RNA communication Consortium (ERCC).

There are three main classes of EVs: microvesicles, exosomes and apoptotic bodies (**Table 2**). Additional classes considered in the literature include oncosomes, exophers, exomeres, etc. These additional categories won't be described or discussed further, as they are not related to my thesis.

Table 2. Main extracellular vesicles properties.

	Exosomes	Ectosomes	Apoptotic bodies
Size of vesicles/ shape	30 – 100 nm, regular	100 – 1,000 nm, irregular	50 - 5,000 nm, irregular
Markers	LAMP-1, tetraspanins, Alix, MHC-I, MHC-II, HSP70, TSG100	Selectins, integrins, tissues factor and cell-specific markers	Histones, organelles
Origin	Endosomal compartments of cells	Cell surface plasma membrane	Cells which undergo apoptosis
sedimentation	100,000 – 130,000 x g	16,000 – 25,000 x g	5,000 -16,000 x g

Exosomes have an endosomal origin. For exosomes to exists, endosomes should form first (Kowal, Tkach, and Thery 2014). Endosomes are formed by plasma membrane invagination; this gives rise to an early endosome that will mature into multivesicular bodies (MVB) **Figure 2A**. MVBs typically have two possible fates, they either end up being degraded in a lysosome or their contents are released to the extracellular space by exocytosis (fusion of the MVB with the plasma membrane). Exosomes are the smallest EV class: their diameter ranges from 30 to 100 nm. When seen by electron microscopy, exosomes have a characteristic 'saucer-like' morphology, a flattened sphere that is limited by a lipid bilayer (Théry et al. 2002). Characteristic surface markers for exosomes include several members of the tetraspanin family such as CD9, CD63, CD81 and CD82, the major histocompatibility complexes I and II, integrins, the chaperone heat shock protein 70, etc (Théry et al. 2002).

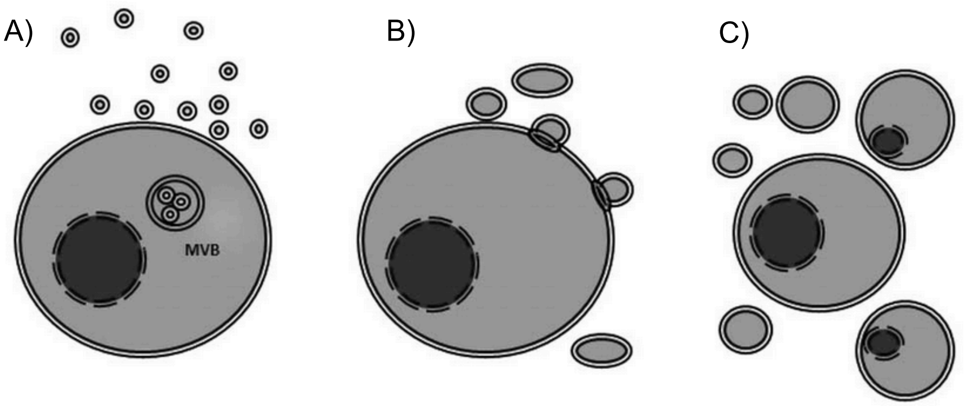


Figure 2. Cartoons of extracellular vesicles classification. A) Exosomes, B) Ectosomes and C) Apoptotic bodies. MVB stands for multivesicular bodies.

Ectosomes, also known as microvesicles or microparticles, have a plasma membrane origin. The plasma membrane starts budding and finally sheds into 100-1000 nm vesicles (Cocucci and Meldolesi 2015). Ectosomes are typically larger than exosomes, but smaller than apoptotic bodies (**Figure 2B)**. Under certain stimuli, some cells release high amounts of ectosomes in a short period of time, some examples of these sudden releases include tissue factor secretion by platelets during coagulation as well as PC-12 adrenal gland cell line ectosomes upon exposure to ATP (Cocucci, Racchetti, and Meldolesi 2009). These ectosome bursts may lead to a diminished cell size due to membrane loss, that is compensated by the subsequent incorporation of intracellular membrane elements to the cell membrane (Cocucci and Meldolesi 2015). Characteristic surface markers for microvesicles are selectins, integrins and cell-specific markers (Pegtel and Gould 2019).

Apoptotic bodies form as part of membrane protrusions (or blebs) associated to programmed cell death. Apoptotic bodies are the largest of EVs with a size range between 1-5 µm; they are so big that they can even contain organelles **Figure 2C**. Other contents include genomic DNA chunks and histones, due to nuclear fragmentation (Atkin-smith and Poon 2016).

The first evidence for the role of EVs in cell-to-cell communication comes from exosomes from antigen presenting cells. In 1996 a seminal study by Raposo *et al.* showed that EVs can be used to present antigens and to stimulate an immune response *in vivo* (G Raposo, H W Nijman, W Stoorvogel, R Liejendekker, C V Harding, C J Melief 1996). Additional cases for cell-to-cell communication with EVs include human placental trophoblasts. Here, exosomes transfer miRNAs that protect non-placental cells from infection by inducing autophagy in virus-infected cells (Delorme-axford et al. 2013).

Melanoma tumors evade T cell surveillance by releasing exosomes that contain PD-L1 (programmed death-ligand 1). Exosome PD-L1 interacts with PD-1 in CD8 T cells, hindering tumor cell killing (G. Chen et al. 2018). This study also revealed that exosome PD-L1 levels increased with increased levels of IFN-γ, which reflects how exosome contents may be altered with different environmental cues.

EVs have several potential biomedical applications. For instance, Coakley *et al.* applied nematode EVs as a vaccination that reduced worm burden in successive infections (Coakley et al. 2017) (more details in Background). EVs are also promising biomarkers for several diseases. Just to mention an example, a 2015 study showed that exosomes containing the cell surface peptidoglycan glypican-1 are good biomarkers to detect early pancreatic cancer (Melo et al. 2015). The discussion of EV biomedical applications are beyond the scope of this work and won’t be commented further.

Communication between different species via small RNAs

In a perspective paper, Sarkies and Miska argued about RNA being transferred between organisms as a means of communication or “social RNA" (Sarkies and Miska 2013). The authors elaborated on this idea based on the model nematode *C. elegans*, the reason being that this organism is capable of taking up environmental dsRNA and silencing its own genes (see Introduction: RNAi and sRNAs). This capacity makes *C. elegans* an excellent model to study the function of almost any of its genes. However, it is intriguing that an organism would allow its own genes to become susceptible to environmentally available RNAs. Such susceptibility raises the question of whether this ability provides a benefit to this nematode in its natural environment. The authors proposed that this capacity could provide herd immunity to a nematode population against possible threats such as RNA viruses. The proposed model would work like this: if an individual *C. elegans* encounters a possible viral threat and triggers its own RNAi response, the nematode could then secrete an RNA that may initiate the same protective response across its neighbors. If this model does happen in nature, this would represent a case of RNAi transferred between organisms of the same species.

Host-induced gene silencing (HIGS) consists of transgenic plants that express exogenous RNAi triggers that silence essential genes in pathogens. HIGS is a promising technology to protect crops against a wide variety of pests such as fungi, oomycetes, insects, nematodes and even parasitic plants (Nunes and Dean 2012). A former term to describe HIGS is Parasite-derived resistance (PDR) and its origin dates back to the late 1980s. The first evidence for PDR was published in 1988 and comes from tobacco plants expressing the capsid protein for the tobacco mosaic virus. These plants displayed a delay in viral disease development (Register and Beachy 1988). The fact that HIGS confers protection to plants against pests suggests that inter-species RNAi may occur in nature.

In a seminal work led by Dr. Jin in 2013, Weiberg and collaborators reported the first case of cross-species communication mediated by small RNAs (Weiberg et al. 2013). *Botrytis cinerea* is a necrotrophic fungus that affects around 200 plant species. Small RNAs encoded in the fungal genome reach *Arabidopsis thaliana* cells during infection, and down-regulate genes related to defense against pathogens. Mechanistically, *B. cinerea* sRNAs get loaded into the *A. thaliana* AGO1 protein and target MAP kinase genes relevant for the signaling response against infection (Weiberg et al. 2013). The three *Botrytis* sRNAs that were characterized in this study, out of potentially hundreds, were derived form fungal LTRs (Long Terminal Repeat), which originate from the action of transposable elements. This suggests that *Botrytis* RNAi machinery may have been exapted from an original mobile element defense system to a weapon used during parasitism. Repetitive elements, such as transposable elements, have been proposed to have an influence on the genomic architecture of fungal plant pathogens, and may associate with virulence genes (Möller and Stukenbrock 2017). It would be interesting to investigate the frequency of LTR-derived sRNAs production within the fungal kingdom, with special attention for fungal pathogens.

Dr. Jin's group later reported that *A. thaliana* responds to *Botrytis* infection by producing its own EV-loaded sRNAs (Cai et al. 2018). These *A. thaliana* sRNAs reach *Botrytis* and inhibit the expression of fungal genes that are involved in pathogenicity, with a bias toward vesicle trafficking pathways (7 out of 32 target genes). Small RNA sequencing from fungal protoplasts isolated from infected tissue revealed 42 *A. thaliana* sRNAs reaching fungal cells, some of them being produced from a trans-acting small interfering RNA (tasiRNA) plant locus. *A. thaliana* triple mutant dcl2/3/4 (Dicer-like proteins 2, 3 and 4) and single mutant rdr6 (a RdRP relevant for siRNA signal amplification) was more susceptible to *B. cinerea* infection, suggesting that the tasiRNA pathway is relevant for anti-fungal RNAi response (Cai et al. 2018). This is the first example of a natural case of HIGS. This case also exemplifies an evolutionary RNA arms-race between a pathogen and its host.

In the literature, the phenomenon of a sRNA produced by an organism and repressing the expression of a gene of another organism has been frequently termed cross-kingdom RNAi (Weiberg, Bellinger, and Jin 2015). I think this term imposes a restriction in the definition, implying a long phylogenetic distance between the interacting organisms. I propose the term inter-species RNAi, this term describes more accurately some recently described interactions such as *Cuscuta campestris* with *A. thaliana* and parasitic nematodes infecting mammalian hosts (described below). These cases involve interactions within kingdoms, but between organisms belonging to different species.

The 2010s witnessed the rise of the inter-species RNAi field (Weiberg et al. 2013) and the posterior accumulation of examples of this phenomenon. We now know that sRNA-mediated communication is a frequent strategy among parasites to take advantage of their hosts. In the following paragraphs, I will introduce some of these cases that are relevant for my PhD thesis.

Cuscuta campestris is an obligate parasitic plant incapable of performing photosynthesis. In order to survive, *C. campestris* infects other plants to obtain water and nutrients using a root-like organ named haustorium to penetrate the xylem of its host (Kaiser et al. 2015). The haustorium allows molecular exchange between host and parasite, including viruses, proteins and mRNAs. Shahid and collaborators explored whether haustorium allows for sRNA movement between *C. campestris* and the model plant *A. thaliana* (Shahid et al. 2018). The authors used sRNA-Seq to characterize the sRNA populations present in the parasitic stem and interaction site (including the haustorium). A total of 43 miRNAs were upregulated in haustoria relative to parasitic stem, of these, 26 were 22 nt long miRNAs which is a less common size for plant miRNAs compared to 21 nt. These 22 nt miRNAs are associated with the production of phased secondary siRNAs (phasiRNAs). Production of phasiRNAs is triggered by a 22 nt miRNA guiding Argonaute-mediated slicing of a primary precursor transcript. This cleavage event recruits RDR6 (an RdRP) that makes the sliced mRNA double stranded, which is later processed by Dicer proteins into 21 (DCL4) or 24 nt (DCL5) siRNAs. The phasiRNAs receive their name after their characteristic 21 or 24 nt interval phased piling up when aligned to the producing locus (Fei, Xia, and Meyers 2013). The authors discovered that the parasitic plant miRNAs initiate a repression cascade dependent on host's DCL-4 and RDR6, showing that *C. campestris* hijacks its host's RNAi machinery. Finally, a search for potential binding sites revealed that some of the targets for *C. campestris* miRNAs are conserved in other dicots.

In a follow up paper, Johnson and collaborators analyzed the sRNA expression profiles of seven members of the *Cuscuta* genus in their interaction with *A. thaliana* (Johnson, de Pamphilis, and Axtell 2019). The majority of haustorium-induced sRNAs were not present in more than one *Cuscuta* species. Nearly half of the *C. campestris* haustorium-induced sRNAs (208/408) originate from miRNA hairpins, and a large proportion of these *C. campestris* miRNAs were 22 nt long. As mentioned before, in plants 22 nt miRNAs trigger the production of secondary siRNA accumulation which amplifies degradation of the targeted mRNA, analogously to what happens in *C. elegans* with the 22G RNAs (Fei, Xia, and Meyers 2013). The authors then used two different techniques to detect the interaction between *C. campestris* sRNAs and *A. thaliana* mRNAs: secondary siRNA accumulation and degradome analysis. The degradome analysis recovers 5' ends of both capped and uncapped mRNAs, and is useful to detect Ago slicing activity. By combining these techniques, the authors identified novel targets for *C. campestris* sRNAs, most of them in conserved protein coding genes (Johnson, de Pamphilis, and Axtell 2019). *C. campestris* sRNAs can be grouped into superfamilies that display variation in a three-nucleotide period. The authors propose that this nucleotide variation compensates for synonymous substitutions in host targets, hindering host targeting avoidance.

Inter-species RNAi may also occur in bacterial-eukaryote interactions as well as mutualistic relationships. Ren and collaborators showed that *Bradyrhizobium japonicum* produces tRNA-derived sRNAs that influence *Glycine max* nodulation (Ren et al. 2019). These rhizobial tRNA-derived fragments are loaded into soybean AGO1, as shown by AGO1 pull down and stem-loop PCR. Some of the predicted targets for these bacterial sRNAs include orthologs for ROOT HAIR DIRECTIVE 3, HAIRY MERISTEM 4, and LEUCINE-RICH REPEAT EXTENSION-LIKE 5, which are important for root hair and plant development in *A. thaliana.* Silencing either of three selected individual tRNA-derived fragments with short tandem mimics resulted in reduced nodulation numbers as well as an aberrant early-stage infection in root hairs (Ren et al. 2019). Knockouts for the mentioned predicted targets for rhizobial tRNA-derived fragments promote nodulation, while overexpression of these plant mRNA targets results in nodulation inhibition. This is the first case of sRNA mediated communication in bacterial-legume interactions.

The filarial nematode *Litomosoides sigmodontis* generates excretion-secretion (ES) products that can be collected and studied in laboratory conditions. Quintana *et al* discovered that this ES contains sRNAs that are protected from degradation, although it is unclear if this protection is due to EVs or to RNA-protein interactions (Quintana et al. 2019). They discovered that ES is rich in tRNAs, rRNAs and contains miRNAs to a lesser extent, in contrast, miRNAs are more abundant than tRNAs in adult worms. Interestingly, 22G sRNAs are found in adult worms, but are absent from ES; this is different to ES products from clade V parasitic nematodes, where 22G sRNAs are the most abundant RNA class (Chow et al. 2019). The authors then looked for *L. sigmodontis* sRNAs in serum and macrophages of infected mongolian gerbils and found that miRNAs become the predominant RNA class detected *in vivo*. In particular, miR-92-3p and miR-71-5p are robustly detected in serum and macrophages (Quintana et al. 2019). Let-7 is highly expressed in macrophages, but it is hard to tell if it is being produced by the rodent or the nematode as they have identical sequences. Further work is needed to test if *L. sigmodontis* miRNAs are functional inside macrophages.

The immune system

The immune system represents all the different biological processes, structures and mechanisms to protect an organism from pathogens. This includes physical barriers such as skin, innate system components, as well as vertebrate's sophisticated adaptive system (Kindt et al. 2007). Here, I will present a general introduction to the mammalian innate and adaptive immune systems in order to understand how mice usually deal with nematode infections, and the mechanisms used by the parasite *Heligmosomoides* to subvert this response.

Cytokines are small proteins relevant for cellular signaling and responsible for most of the biological effects in the immune system (Berger 2000). The immune system response and effector cell populations are controlled by which combination of cytokines are present in a given tissue location.

Innate immune system

Macrophages are a type of leukocyte (white blood cell) that eliminate pathogens, cellular debris and anything that doesn't have the characteristic proteins of a healthy cell on its surface by phagocytosis. Macrophage classical activation (M1) is stimulated by a cytokine called interferon gamma (IFN-γ) and lipopolysaccharides (LPS) that are found in the outer membrane of bacteria. Classical activation results in microbicidal activities and secretion of pro-inflammatory cytokines such as IL-6, TNF and IL-1ß (Gordon 2003). These cytokines favor activation of cellular immunity (see adaptive immune system). Classical activation aids fighting bacterial and viral infections. Alternative activation of macrophages (M2 or AAM) is stimulated by IL-4 and IL-13 cytokines (Gordon 2003). AAM is the preferable way to deal with allergens and parasites such as gastrointestinal nematodes.

Granulocytes are cells of the innate immune system that receive its name for their granules found in their cytoplasm. These granules harbor nitric oxide, reactive oxygen species as well as proteins relevant for defense such as antimicrobial peptides, acid hydrolases and lysozymes (Kindt et al. 2007). The contents of granules are released upon threat encounters by a process known as degranulation, the release of granules contents harms invading pathogens in granulocyte surroundings. Granulocytes include neutrophils, eosinophils and basophils (Kindt et al. 2007). Neutrophils are the most abundant granulocyte, representing 50-70% of the white blood cell population, these are capable of performing phagocytosis, degranulation and use extracellular traps to immobilize bacteria (Kolaczkowska and Kubes 2013). Neutrophils are the first cell to arrive to inflammation sites. Eosinophils are phagocytic cells that play a relevant role during multicellular parasites infections, allergic reactions and asthma. Basophils are non-phagocytic cells that play a role during helminth infection, but can also cause pathologies during allergic reactions (Voehringer 2013).

Mast cells are functionaly related to basophils, both cell types release histamine and cytokines such as IL-2, IL-3, IL-4, IL-5, IL-6, IL-9, IL-13, IL-15 and TSLP to drive immune reactions (Voehringer 2013). Mast cells generally have fixed locations such as connective tissue or intestinal mucosa, while basophils are motile.

Natural killer cells (NK) are derived from a common lymphoid progenitor, this is the same cell that gives rise to T lymphocyte, however, NK cells do not express the T cell receptor (see below). NK are fast cell responders to viral infections or tumor cells, they cause the death of targeted cells by lysis or inducing apoptosis (Vivier et al. 2011).

Dendritic cells (DC) guard and sense foreign elements and if needed, they warn T cells about possible infections and the nature of the threat (viral, nematode, bacterial, etc.) (Coombes and Powrie 2008). DCs are named after how their branched projections resemble a neuron's dendrites. Both macrophages and DCs are antigen presenting cells. Macrophages and DCs express a range of receptors that are capable of recognizing pathogen associated molecular patterns (PAMPs). Toll-like receptors (TLRs) are a famous type of receptors that recognize PAMPs, a typical example is the binding of bacterial lipopolysaccharide (LPS) to TLR4 (Neill, Golenbock, and Bowie 2013). There are additional categories of receptors that can recognize PAMPs such as C-type lectin receptors. C-type lectins are a superfamily of proteins that recognize a wide spectrum ligands (G. D. Brown, Willment, and Whitehead 2018). They have one or more C-type lectin domain, and were originally named because of its capacity to bind carbohydrates in a Ca^{2+}-dependent manner. However, some C-type lectin domains can recognize a wider repertoire of ligands, including proteins and lipids (G. D. Brown, Willment, and Whitehead 2018).

Antigens are exposed portions of a molecule or epitopes that can be recognized by several components of the immune system such as an antibody or a T cell receptor (M. Sela 1998). Antigen presenting cells (APCs) take pathogen antigens and display them for recognition by T cells. Antigen presenting cells include dendritic cells, macrophages and B cells. These antigens are displayed in the major histocompatibility complex class II (MHC class II). There's a second histocompatibility complex class I (MHC I), which is expressed by all nucleated cells of the body and is used to distinguish healthy from unhealthy cells such as cancerous or virus-infected cells (Kindt et al. 2007).

Innate lymphoid cells (ILCs) are a recently discovered cell type that is particularly present below mucosal-like tissues such as the intestinal epithelium, lungs, etc. They have some characteristics of T cells (see Adaptive immune system) such as being a great source of cytokines, but they lack the T cell receptor. ILCs can be further subdivided according to the cytokines they receive and how they act to different types of immune responses (Koyasu et al. 2018).

Adaptive immune system

Classically, immunity was categorized into cellular immunity and humoral immunity. These broad effects can be attributed to two types of lymphocytes: thymus-derived lymphocytes (T cells) and Bone marrow-derived lymphocytes (B cells) respectively (Kindt et al. 2007). B cells are responsible for producing antibodies and provide humoral immunity, this protection receives its name given classical experiment where cell-free serum could provide protection from a pathogen. Antibodies or immunoglobulins are Y-shaped proteins that can specifically recognize a portion of a foreign molecule (antigen) and facilitate recognition by other immune cells or can inactivate a potential threat by completely surrounding or trapping threats by gluing them to each other (opsonization). T cells express a cell surface receptor named T cell receptor (TCR) that is setup to recognize an antigen loaded into the MHC. T cells can be divided into two broad categories: CD8+ (Cellular differentiation factor 8) and CD4+ T cells. CD8+ cells are responsible of killing tumor cells, virus or intracellular bacteria-infected cells (cellular immunity) and are called cytotoxic T cells (Tc), on the other hand CD4+ T cells are also named T helper cells (Th) (Kindt et al. 2007).

Helper T cells (Th) are the most prolific cytokine producers. Th cells drive the immune response to contend with different threats by producing different cytokines, Th cells are classified according to the cytokines they produce into Th1, Th2, Th17 and Treg (**Table 3**) (Schmidt-Weber 2008). When we suffer an infection of intracellular pathogens such as viruses or bacteria, Th cells produce INF-gamma driving a Th1 response (classically termed cellular immunity). When extracellular pathogens signals warn Th cells, these produce IL-4, IL-5 and IL-13 cytokines that will favor a Th2 response (classicaly referred as humoral response). Two additional T-helper cell populations were recently discovered: Th17 and Treg cells, that produce tissue inflammation and immune suppression respectively. Th17 response is also triggered by extracellular pathogens and it is associated with the production of IL-17, IL-22 and IL-21 (Schmidt-Weber 2008). Treg (T-regulatory) cells are associated with tolerance to self-antigens and harmless foreign antigens (such as food) and produce IL-10 and TGFB1 cytokines (Schmidt-Weber 2008). Tregs actively suppress immune system activation and prevent self-reactivity, their importance is evidenced by the severe autoimmune syndrome immunodysregulation polyendocrinopathy enteropathy X-linked (IPEX), which is caused by a deficiency in Tregs production (Chatila et al. 2000). Th cells are so important that human immunodeficiency virus (HIV) targeting these cells results in immunodeficiency and renders acquired immunodeficiency syndrome (AIDS) patients susceptible to otherwise trivial infections.

Table 3. Immune response categories, based on helper T cells, associated cytokines and functions.

	Th1	Th2	Th17	Treg
Effective against	Intracellular pathogens (viruses, bacteria)	Extracellular pathogens (nematodes)	Extracellular pathogens (bacteria, fungi)	Tolerance
cytokines	Interferon gamma IFN-γ, IL-12	IL-4, IL-5, IL-9, IL-13	IL-17, IL-22, IL-21	IL-10, TGFB1
Effects	Cellular immunity	Humoral immunity, allergic reactions, eosinophilia	Tissue inflammaton	Immune supression

When a DC processes and presents its antigen on its MHC II, it migrates to the closest lymph node in search for a T cell that has a TCR that matches its displayed epitope (**Figure 3**). T cells reside within lymph nodes, in particular for the intestinal immune system these are named mesenteric lymph nodes (MLN). When a DC finds such a naive T cell, the naive T cell becomes activated and starts replicating several times in a process known as clonal expansion. Clonal expansion produces memory T cells and effector T cells (Th and Tc). Both Th and Tc cells migrate to the site of infection to help the other immune system elements. Some of the activated Th cells look for a B cell that had previously recognized the same antigen of interest via its B cell receptor (BCR, the membranal from of antibodies), processed that antigen and displayed it on its surface on the MHC II (**Figure 3**). If the antigen on display by the B cell is recognized by the Th cell TCR, the interaction is further strengthened by the binding of CD40L protein (on the Th membrane) with CD40 protein (on the B cell surface) (Kindt et al. 2007). As a result of this interaction, the B cell will clonally expand and mostly differentiate to plasma cells producing the antibody that recognizes the antigen of the current pathogen (Kindt et al. 2007).

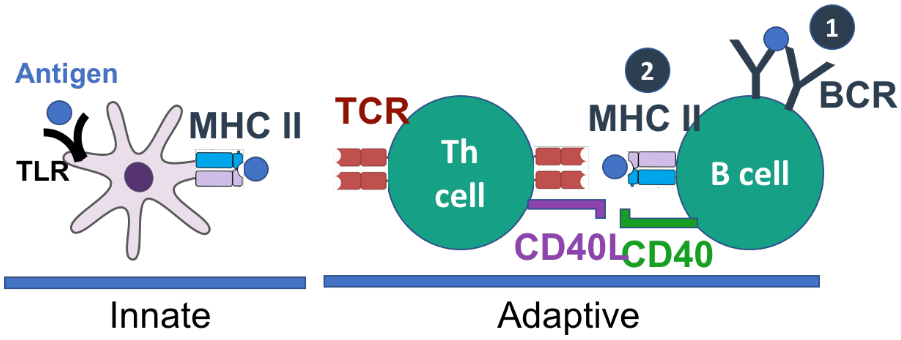


Figure 3. Simplified signaling interactions at the cell surface of immune system cells. TLR Toll-like receptor, MHC II Major Histocompatibility Complex II, TCR T cell receptor, BCR B cell receptor. CD40 cluster of differentiation 40.

After an immune response action, a subset of T and B cells remain as memory cells, resulting in a faster response in case the same threat occurs again. This is what makes people who have overcome an illness, or received a vaccine, immune to further exposure to the same pathogen.

Heligmosomoides bakeri

Heligmosomoides bakeri is a parasitic nematode that infects mice and represents a well-established model to study chronic nematode infections, since adult nematodes can be detected up to 46 weeks in the first infection (Behnke, Menge, and Noyes 2009). The parasitic nematode *Nippostrongylus brasiliensis* naturally infects rats, however, it is able to infect mice. *N. brasiliensis* is used as a model to study acute and transitive infections lasting 6-8 days post-infection (dpi) (Gerbe et al. 2016), or require low worm numbers to survive longer in the host (Behnke, Menge, and Noyes 2009).

It is estimated that 1 out of 8 people around the world are affected by nematode intestinal parasites. Examples of human parasites include the filarial nematodes *Onchocerca volvulus,* the causal agent on river blindness, and *Wuchereria bancrofti,* one of the causal agents of elephantiasis (Paily, Hoti, and Das 2009). These illnesses are non-lethal, however they cause a serious burden that may compromise the work capacity and independence of the affected individual, such as blindness or swelling of limbs. Additional nematodes relevant to human health include *Ascaris lumbricoides*, *Trichinella spiralis*, *Enterobius vermicularis*, *Ancylostoma duodenale*, *Necator americanus*, *Strongyloides stercoralis*, *Trichuris trichiura*, among others (Stepek et al. 2006).

Parasitic nematodes pose a threat not only to humans, but also to livestock and plants. *Haemonchus*, *Cooperia* and *Trichostrongylus* are some of the genera found in cattle, where gastrointestinal nematode infections can result in reduced milk production and reduced body weight compared to antiparasitic-treated individuals (Rodríguez-Vivas, Roger Ivan; Grisi, Laerte; Perez de León 2017). In plants, the root-knot nematode *Meloidogyne incognita* has a wide range of potential hosts and causes crop losses for potato, sweet potato, tomato as well as other cultivars (Abad and Williamson 2010).

Heligmosomoides bakeri is widely used in the laboratory as a model to study chronic gastrointestinal worm infections such as that of the hookworm *Necator americanus*. A Californian strain, *H. polygyrus bakeri* is particularly attractive due to its high numbers in the laboratory model *Mus musculus*. *Heligmosomoides polygyrus polygyrus* is a European strain that reaches lower worm burdens than *H. polygyrus bakeri*. *Heligmosomoides bakeri* was previously named *Nematospiroides dubius* and a number of older papers use this nomenclature. Cable and collaborators argued that *H. polygyrus bakeri* and *H. polygyrus polygyrus* are different species based on comparisons of ribosomal internal spacer (ITS) and the mitochondrial cytochrome c oxidase I molecular markers (Cable et al. 2006). In addition, undergoing comparative genomic analyzes suggest that these are two different species (Blaxter, unpublished results), therefore, we shall refer to *H. polygyrus bakeri* as *H. bakeri* throughout the rest of this work.

H. bakeri infects mice and spends part of its life cycle in its host intestine (**Figure 4**) (Reynolds, Filbey, and Maizels 2012). The eggs are excreted and hatch outside of the host, where larvae develop from L1 until reaching infective stage L3. These L3 larvae are eaten by mice and penetrate the intestinal walls reaching the serosa layer within 24 hrs, where they develop to L4 and adults. Once they have reached adulthood by day 10 post-infection (dpi), nematodes emerge back into the intestinal lumen where they feed on host intestinal tissue (Bansemir and Sukhdeo 2016). Adult worms coil around the small intestine villi to attach themselves and mate, producing eggs that will get excreted along with the feces.

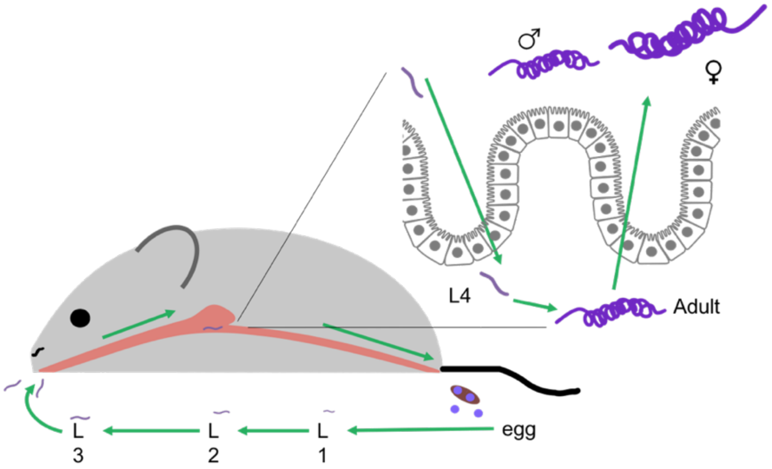


Figure 4. *Heligmosomoides* lifecycle.

Heligmosomoides has evolved mechanisms to evade the mammalian immune response (discussed below), this has resulted in primary infections being typically non-resolving, and secondary infections lasting months (Maizels et al. 2012). However, there’s variation in secondary infection resolution time depending on the mouse strain. Slow responsive mouse strains take more than 20 weeks to expel worms (CBA, C3H, SL and A/J strains), intermediate responsive ones take 8 to 20 weeks (C57BL/6, C57BL/10 and 129/J strains), fast responsive strains take 6 to 8 weeks (BALB/c, DBA/2 and NIH strains), and rapid responsive strains take 4 to 6 weeks to get rid of worms (SJL and SWR strains) (Reynolds, Filbey, and Maizels 2012).

Immune response to *Heligmosomoides*

The immune response to *Heligmosomoides* involves many different cell types, but I will describe the role of intestinal epithelial cells and macrophages, as these are the cell types that we studied during this thesis.

Mammals usually contend with helminth infections through the Type 2 immune response, characterized by the production of IL-4, IL-5, IL-9 and IL-13 cytokines (Reynolds, Filbey, and Maizels 2012). In mouse strains that naturally do not expel their parasites, a Type 1 response is favored, characterized by production of IFN-γ and IL-12 cytokines (Artis and Grencis 2008).

Intestinal epithelial cells (IECs)

Intestinal epithelial cells (IECs) perform a variety of functions in response to *Heligmosomoides* infection. These range from detecting helminths, alerting immune system cells and actively participating in worm expulsion (Artis and Grencis 2008).

It is still unclear how IECs detect the presence of nematodes in the intestinal lumen, since no nematode-specific pattern recognition receptors have been identified to date. IECs may sense physical damage through damage-associated molecular patterns. Remarkably, IECs express MHC class I and II, as well as all the required machinery for antigen processing and presentation. However, the capacity for IECs to present antigens is still controversial (Artis and Grencis 2008).

IECs send alerts about nematode presence by secreting IL-33, IL-25 and Thymic stromal lymphopoietin (TSLP) cytokines (**Figure 5**). IECs are a potent source for IL-33, which is considered a Type 2 immune response accelerator. IL-33 is constitutively expressed and can be found in the nucleus of epithelial cells in barrier tissues and in endothelial cells in blood vessels (Scott et al. 2018). IL-33 is released upon IEC necrosis and its presence is considered an alaram signal of cellular damage. In the context of the intestinal epithelium, IL-25 is produced by tuft cells, a subclass of IECs that display similar signaling pathways to those of taste buds (Nadjsombati et al. 2018). Nadjsombati and collaborators proved that the stimulation of the receptor SUCNR1 by succinate in tuft cells its enough to produce IL-25 (Nadjsombati et al. 2018). TSLP synthesis is dependent on the NF-kB signaling pathway. The NF-kB signaling pathway is relevant to contend with nematodes, since its disruption exclusively within IECs resulted in defective development of Th2 cytokine production and an increased susceptibility to a nematode infection by *Trichuris muris* (Artis and Grencis 2008).

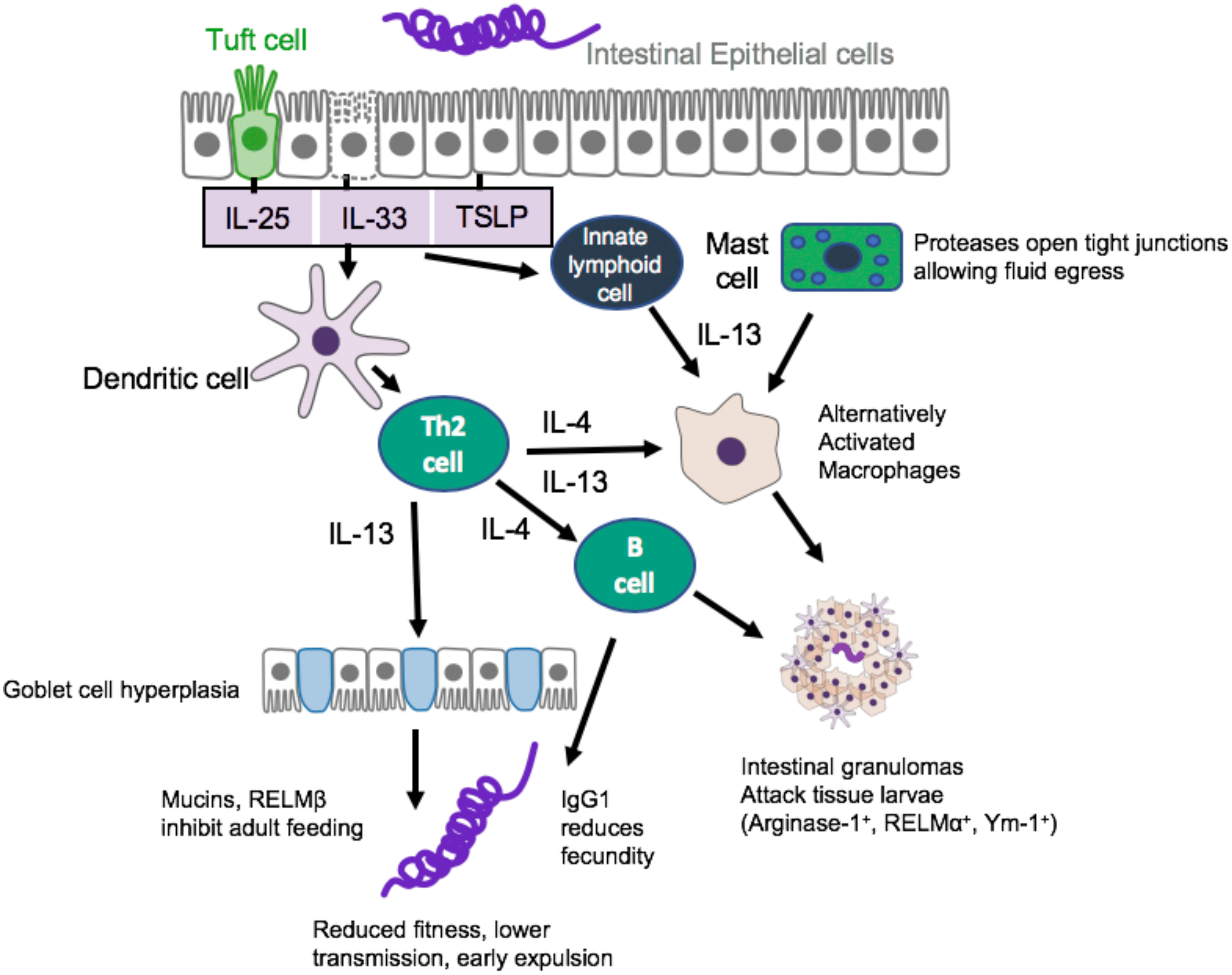


Figure 5. Pathways of immunity to *Heligmosomoides*. Adapted from (Maizels et al. 2012).

Tuft cells, a low-proportion cell type in the intestinal epithelium, are key to triggering a type 2 immune response by sensing and alerting about infection via IL-25 cytokine production and secretion (Gerbe et al. 2016). *N. brasiliensis* infection produces an increase in tuft cell numbers measured as the number of tuft cells per crypt-villus. This tuft cell hyperplasia is also observed during *H. bakeri* infections starting at 1.9% at a basal level, up to 6.3% at 3 dpi and even more at 10 dpi representing 8.5% of intestinal epithelial cells (Haber et al. 2017). A mutant mouse strain lacking tuft cells displays more *N. brasiliensis* worm burden than its wild type counterpart (Gerbe et al. 2016).

IEC changes in response to nematode infection include an increased permeability due to protease degradation of tight junctions (Artis and Grencis 2008). Additional changes in IECs include an increase of goblet cell numbers termed “goblet cell hyperplasia”, which correlates with *H. bakeri* resistance. This increase in goblet cells by 10 dpi was also detected in a single-cell RNA-Seq study (Haber et al. 2017). Goblet cells secrete mucus and are a source for the RELM-ß cytokine, a member of the resistin-like family that directly affects *Heligmosomoides* worm feeding (Herbert et al. 2009). Normal IEC to goblet cell differentiation is initiated by IL-13 sensing (Artis and Grencis 2008). IECs also produce a phospholipase A_2, PLA_2g1B, that plays a role in *H. bakeri* expulsion and reduces the number of luminal nematodes in a dose-dependent manner when externally supplied (Entwistle et al. 2017).

Macrophages

Macrophages are key players for an effective immune response to a second infection. This is evident as macrophage depletion via clodronate treatment compromises the ability to expel worms. Alternatively activated macrophages (AAMs) are responsible for fighting nematodes while classically activated macrophages deal with bacterial infections. IL-4 and IL-13 promote macrophage alternatively activation (M2 or AAM), while IFN-γ triggers classical macrophage activation (M1 or CAM) (Gordon 2003). AAM also results in increased expression of the MHC II and antigen presenting capacities for macrophages, which will enhance the humoral immunity.

Developing nematodes (L3 to L4 stages) display low mobility, and this is the point when worms are more susceptible to an attack by macrophages. AAMs attack developing nematodes found in the serosa layer resulting in granulomas, which are foci for macrophage alternative activation and nematode killing. The more granulomas observed, the fewer adult worms recovered from mice. The adult worms that make it to the intestinal lumen are out of reach for AAMs. Granulomas form during second infections because AAM differentiation is induced by memory Th2 cells, resulting from a first encounter with the nematodes (Anthony et al. 2006). Noteworthy, the most resistant mouse strains (such as SJL) develop granulomas during first infections (Filbey et al. 2014).

Classically activated macrophages (CAMs) express Nitric oxide synthase (iNOS), which confers macrophages the capacity to kill virus or bacterial infected cells with reactive oxygen species. AAMs, on the other hand express Arginase-1 (Arg-1). iNOS and Arg-1 compete for the same substrate L-Arginine. Arg-1 converts L-Arginine to L-ornithine, polyamines and urea. Arginase-1 (Arg-1) is a key enzyme to cope with *H. bakeri* challenges, as a treatment with S-(2-bronoethyl)-1-cysteine, an Arg-1 inhibitor, renders mice incapable of clearing secondary infections (Anthony et al. 2006). Arg-1 products may directly harm nematodes as worms isolated from a secondary infection have higher cytochrome oxidase activity, a stress-response marker, relative to primary infection worms. Importantly, this difference in cytochrome oxidase activity is lost with the arginase inhibitor treatment. It has also been shown that L-ornithine or polyamines diminish larval motility in *in vitro* macrophages larval co-culture experiments, and this effect was sensitive to the Arg-1 inhibitor (Bieren et al. 2013).

AAMs also display high expression of RELM-alpha and Ym-1 (Coakley et al. 2017). The RELM-alpha (encoded by Retnla) cytokine produced by AAMs, seems to down-regulate an exacerbated Th2 response, as suggested by studies using Retnla−/− mice. These mice show a graver inflammation pathology (in liver, intestine and lungs) associated to the Th2 response than wild type mice (Nair et al. 2009). Ym-1 is a member of a chitinase-like family but lacks any detectable chitinase activity. Some authors suggest a role for Ym-1 in damage repair caused by nematodes moving through intestinal tissue (Maizels et al. 2012).

Heligmosomoides excretory-secretion (HES) & immune modulation

Heligmosomoides owes most of its immune evading capacities to secreted products that are collectively named *Heligmosomoides* excretion-secretion (HES). HES can be harvested and studied by removing adult nematodes from infected mice 14 days post infection (dpi). The collected nematodes are washed and maintained in serum-free medium and the first 24 hours of culture are discarded to avoid mouse contamination. HES is then collected from day 2 to 14 days post-harvest, with subsequent centrifugation and filtering to remove eggs (Johnston et al. 2015).

Segura and collaborators studied the effects of HES on Dendritic Cells (DCs) *in vitro*. Dendritic cells are major contributors to presenting antigens to T cells and heavily influence the decision of the type of immune response to trigger. Experiments using co-cultured DCs and naive T cells revealed that HES-treated DCs resulted in reduced production of IFN-γ (Th1 cytokine) or IL-4 (Th2 cytokine) by T cells, but increased IL-10 levels (Treg cytokine). This suggests that DC exposure to HES would favor Treg production instead of Th1 or Th2 cells and their corresponding immune responses, thus having immunomodulatory capacity (Segura et al. 2007).

Treg cells attenuate the immune response and are part of the intestine homeostasis during food antigen response. Grainger and collaborators explored whether *H. bakeri* may exploit this T cell population in order to avoid expulsion. They discovered that HES can drive Th0 cells to differentiate to regulatory T cells (Treg) which repress Th2 response (Grainger et al. 2010). In mammals, TGF-β promotes the differentiation of naive T helper cells to Tregs via activation of Foxp3 transcription factor. Interestingly, Grainger *et al.* also discovered that *H. bakeri* secretes a TGF-β analog that drives Treg differentiation. They also found that *Teladorsagia circumcincta*, a parasite closely related to *H. bakeri*, also secretes a TGF-β mimic, but *Haemonchus contortus* excretion-secretion lacks such an analogous cytokine activity.

Coakley *et al.* (Coakley et al. 2017) studied how HES inhibits both classical (Th1-driven) and alternative activation (Th2-driven). I will present more details of these findings in the Background section of this thesis, due to the particular relevance for my project.

HES contains several protein components (at least 374), the most abundant of them are venom allergen-like (VAL) family proteins with 25 representatives found in HES (Buck et al. 2014). The function of these proteins remains elusive although they are also present in the adult nematode surface. Hewitson et al. (Hewitson et al. 2013) collected L4 excretion-secretion (ES) form tissue stage larvae and compared them with regular HES from intestinal lumen adult nematode stage. They found 135 shared proteins between L4 ES, 229 adult HES exclusive proteins and 79 L4 ES exclusive proteins. They found SXC-like domain-containing proteins, VAL family proteins and Sushi-like domain proteins (typically involved in protein-protein interactions) among L4 enriched amino acid sequences. HES also contain sRNAs, which I will further discuss in the following Background section.

Background

All three papers described in this section refer to work done in the laboratory of Dr. Amy Buck at the University of Edinburgh, UK, while part of the bioinformatic analyses described below were done in our group.

H. bakeri secretes EVs that have immune-regulatory capacity and harbor sRNAs

In 2014 Buck and collaborators (Buck et al. 2014) reported that *Heligmosomoides* HES products contain RNA. The most abundant sRNA class found was miRNA, with other RNA classes such as yRNAs present to a lesser extent. yRNAs are small non-coding RNAs that are part of the human Ro60 ribonucleoprotein particle. They were discovered as this complex is targeted by antibodies in some Lupus autoimmune syndromes. Studies later revealed that homologs for yRNAs exist in nematodes as well as in some bacteria (Kowalski and Krude 2015). Several secreted parasitic miRNAs such as miR-100-5p, lin-4-5p, miR-83a-3p, miR-263-5p, let-7a-5p, miR-79a-3p and miR-63 have identical seeds to mouse miRNAs, which raised the question of nematode miRNAs targeting host genes through existing target sites.

They also discovered that HES contains extracellular vesicles (EVs) with a size range of 50 – 100 nm and have a 'saucer-like' morphology described previously for exosomes (Théry et al. 2002). In Buck *et al* (Buck et al. 2014) these EVs are referred to as exosomes, however additional experiments would be needed to prove that these EVs have an endosomal origin. As a precaution, I will refer to *Heligmosomoides* vesicles as EVs throughout this thesis. These EVs contain a WAGO protein (see Introduction: Small RNAs and RNAi), from now on referred to as exWAGO (excreted worm-argonaute protein).

EVs provide protection from degradation to secreted miRNAs by encapsulation, as shown by RNase assays, detergent application and qPCR quantification. These EVs may originate from the nematode intestinal lumen as suggested by TEM microscopy images (Buck et al. 2014). Additionally, a proteome analysis also suggests that EVs have an intestinal origin given the presence of intestinal acid phosphatase family member, P-GlycoProtein related family member, Vacuolar H ATPase family members 15 and 16, as well as other 13 proteins that are typically present in nematode intestinal epithelial cells.

Interestingly, *H. bakeri* EVs can suppress a type 2 innate immune response *in vivo*. To test the immune modulatory capacity of *H. bakeri* EVs, Buck *et al* used an asthma model in mice (Havaux et al. 2005) and measured the number of eosinophils in bronchoalveolar lavage upon fungal (*Alternaria* extract) antigen stimulation. A *H. bakeri* HES pre-treatment resulted in reduced numbers of eosinophils relative to those mice treated with buffer.

The authors then showed that *H bakeri* EVs have the capacity to be internalized by mouse intestinal epithelial cells (a MODE-K cell line, from now on referred as IECs) by staining EVs with PKH67 cell membrane labelling and assessing uptake with fluorescence microscopy. They also tested if sRNAs are also internalized by measuring yRNAs and miRNA-16 levels inside IECs with PCR quantifications. These results show that EVs are internalized by host cells and that they do deliver some of their sRNA cargo.

Buck and collaborators (Buck et al. 2014) showed that IL-33R and Dusp1 are down-regulated in IECs by *H. bakeri* EV incubation. IL-33R is part of the receptor for IL-33, an alarmin cytokine involved in protection against multicellular parasites. Dusp1 is a regulator of MAPK signaling associated with diminishing the type 1 pro-inflammatory reaction to Toll like receptor (TLR) ligands. In order to associate the repression of these genes with parasitic miRNAs activity, the authors designed the following experiment: They used a vector with luciferase as a reporter that has the 3’ UTR region of the gene of interest, later cells of interest are transformed with a synthetic parasitic miRNA together with the reporter vector. If the synthetic miRNA does hinder the luciferase activity this provides evidence that the tested 3’UTR region can be regulated by the transformed miRNA. The authors found reduced luciferase activity for Dusp1 reporter, which possess 7mer binding sites for miR-200, let-7 and miR-425. However, no reduction in luciferase activity was observed for IL-33R, even though it contains two 7mer sites for miRNA-71. The difference in repression was attributed to higher conservation of Dusp1 3’UTR region relative to that of IL-33R.

Finally, the authors found that there are shared parasitic miRNAs detected in intestinal cells treated with *H. bakeri* EVs and those found in mouse serum infected with *Litomosoides sigmodontis*. Shared miRNAs include miR-100, miR-71 and Bantam (miR-58), and these are implicated in pharyngeal development, regulation of lifespan, and developmental apoptosis processes (Buck et al. 2014).

Overall, Buck *et al* (Buck et al. 2014) suggested that parasitic nematodes may use EV-loaded RNAs to modulate the mammalian immune response.

H. bakeri EVs suppress macrophage activation

In 2017, Coakley and collaborators (Coakley et al. 2017) reported a study on the effects of *H. bakeri* EVs on macrophages. They found higher EV internalization by bone marrow-derived macrophages (BMDM) than by intestinal epithelial cells. This was shown by marking vesicles with the fluorescent dye PKH67 and measuring EV uptake using flow cytometry. Macrophage EV uptake is an active process that can be blocked by applying cytochalasin D, an actin remodeling inhibitor that has been used to block endocytosis and phagocytosis in other studies. The application of cytochalasin D greatly diminished EVs uptake by BMDM, revealing that this internalization requires phagocytosis (Coakley et al. 2017).

EV uptake is favored when BMDM are pre-treated with IL-4/IL-13 (Th2 cytokines) and hampered with lipopolysaccharide pre-treatment, a trigger molecule for classical macrophage activation. Applying anti-EV antibodies increases BMDM EV uptake and promotes EV and lysosome co-localization which suggests EV degradation. Importantly, in absence of EV antibodies, EV and lysosomes do not co-localize as much as with anti-EV treatment. This suggests that EV could potentially escape lysosome degradation in wild type conditions (without EV antibodies) (Coakley et al. 2017).

H. bakeri EVs suppress the activation of both type 1 (classical) and type 2 (AAM) activation of macrophages as revealed by quantifying hallmark genes for either pathway. Chosen markers for classical activation include tumor necrosis factor (TNF), inducible nitric oxide synthase (iNOS), IL-6 and IL-12.

Tested markers for alternative activation of macrophages (AAM) include Retnla, Ym1 and arginase-1. All these markers are repressed upon incubation with *H. bakeri* HES, EVs or even supernatant (EV-depleted HES) relative to control conditions (Coakley et al. 2017).

AAM is also associated with the expression of the ST2 subunit of the IL-33 receptor. In accordance with Buck *et al* (Buck et al. 2014), Coakley found that ST2 is repressed upon EV incubation. The repression of ST2 is dependent on EV internalization, as this effect was disrupted by applying cytochalasin D or EV antisera. Notably, type 2 activation is still suppressed via EV treatment in ST2 deficient mice, which suggests that there still may be other receptors implicated in type 2 response activation (Coakley et al. 2017).

Interestingly, vaccination experiments with any of the nematode secretion products, HES, EVs or supernatant with an adjuvant (a substance that the immune response to a vaccine) before nematode challenge stimulates immunity to *H. bakeri* (Coakley et al. 2017). This nematode secretion-based vaccination may have potential applications for protecting susceptible human populations from parasitic nematodes.

Finally, Coakley and co-workers also showed that ST2-/- mice are more susceptible to *Heligmosomoides* infection, even after EV vaccination (Coakley et al. 2017). Mutant ST2-/- mice showed higher egg burden in feces and adult worm counts, as well as less granulomas relative to wild type mice. IL-33R defective mice also displayed reduced numbers or macrophages, ILC2 and Th cells in mesenteric lymph nodes (MLNs).

It is yet to be determined if the sRNAs present in *Heligmosomoides* secretion products play a role in repressing any of the EV-downregulated genes: Retnla, Ym1 and Arg1, ST2 subunit of IL-33R, tumor necrosis factor (TNF), inos, IL-6 and IL-12.

H. bakeri secretes an Argonaute protein that loads 22Gs that originate from repetitive sequences

In a recent paper, we showed that exWAGO is truly vesicular and present at 3.4 (+/- 1.1) copies per EV. This was shown by sucrose gradient isolation of EVs, a proteinase K protection assay followed by a subsequent western blot assays using an exWAGO antibody. In these assays exWAGO co-purifies with EV fractions (Chow et al. 2019).

Our collaborators in Edinburgh sequenced and assembled a new version of the *H. bakeri* genome. This allowed comparative genomics with *C. elegans* and other clade V nematodes. The *H. bakeri* genome contains a much larger repertoire of repetitive regions than *C. elegans* (Chow et al. 2019). More than half (58.3%) of the *H. bakeri* genome contains some type of repeated element, while in *C. elegans* it is less than 20%.

A comparison of RNAi machinery components in clade V nematodes suggests that an expansion of WAGOs occurred in the *Caenorhabditis* clade that didn’t happen in other rhabditida members, such as the strongylidae family. Nearly all of the machinery for miRNA and piRNA pathways is conserved across

Clade V nematodes, while other siRNA pathways such as CSR-1, as well as nuclear Argonautes NRDE-3 and HRDE-1 are absent in parasitic nematodes, including *H. bakeri* (Chow et al. 2019).

Phylogenetic analysis of Argonautes revealed that exWAGO is highly conserved in parasite nematodes belonging to clade V. The gene structure (number of exons) for exWAGO is conserved across clade V parasites. The phylogeny structure of Argonaute proteins suggests that *Caenorhabditis* SAGO-1, SAGO-2 and PPW-1 are co-orthologues of exWAGO, as these *C. elegans* WAGOs branch from within exWAGO. Interestingly, *C. monodelphis*, a basal species to the *Caenorhabditis* genus, has a clear exWAGO homolog (Chow et al. 2019). Further characterizations of *C. monodelphis* exWAGO and comparisons to clade V parasite homologs may reveal interesting insights about exWAGO evolution.

Analysis of available RNA-Seq data revealed that exWAGO is the most expressed Argonaute protein in the majority of clade V parasites (Chow et al. 2019). Chow and collaborators also identified exWAGO in the excretory-secretory (ES) product of *Nippostrongylus brasiliensis* (a rat parasite affecting lungs and intestine). There was no evidence for any other secreted Argonaute protein in ES of *H. bakeri* or *N. brasiliensis* (Chow et al. 2019).

Secondary 22G siRNAs are the major contributors to adult *C. elegans* and *H. bakeri*, as well as the EV sRNA profiles. These are produced by RNA-dependent RNA polymerases (RdRP, see Introduction), and have a 5' tri-phosphate due to direct ribonucleotide triphosphate incorporation as the first building block during synthesis. These 22Gs were not discovered previously in *Heligmosmomoides* as the tri-phosphated 5' end impedes detection of these sRNAs in classical sRNA-Seq since only those with a 5' mono-phosphate can be ligated during library preparation. To detect 22Gs a prior phosphatase treatment is needed that will turn 5' polyphosphate to monophosphate and will enable adapter ligation. Adult RdRP siRNAs seem to have similar endogenous functions in both *C. elegans* and *H. bakeri*, as they map antisense to many mRNA and retrotransposons in both species. Interestingly, EV-enriched 22G tend to originate from repeated sequences that are *Heligmosomoides* specific or "novel repeats", as well as transposons (1.7 and 1.9-fold enrichment, respectively, relative to adult *H. bakeri* 22Gs) (Chow et al. 2019).

Finally, exWAGO immunoprecipitation from both adult worms and EV show enrichment for 22G sequences and depletion for miRNAs and Y-RNAs. For example, miRNA-100 is depleted from the adult and the EV-loaded exWAGO.

These recent studies provide evidence for a role of sRNAs in the immune modulation capacity of *H. bakeri* HES and EVs (Chow et al. 2019). However, these studies do not prove that sRNAs are directly implicated in immune suppression. Buck *et al.* (Buck et al. 2014) proposed miRNAs as the main characters driving an immunomodulatory effect, but we recently revealed that 22Gs are the main components of *H. bakeri* secreted products (Chow et al. 2019). The role for 22Gs as potential immunomodulatory molecules remains completely unexplored, until the present work. Here we want to relate the presence of *H. bakeri* 22G sRNAs found in it's secretion products (HES and EVs) with down-regulation of their potential targets in an *in vitro* experiments with mouse cells.

Chapter 1: Disentangling sRNA-Seq data to study RNA communication between species

Most of the results presented in this chapter are part of the paper “Disentangling sRNA-Seq data to study RNA communication between species” published in *Nucleic Acids Reseach* (Bermúdez-Barrientos et al. 2020).

We will refer to pairs of interacting organisms as host-symbiont species pairs. The symbiont organism lives interacting with a host species in a favorable (mutualism) or unfavorable (for the host) manner (parasitism). We chose the term symbiont as we have representatives for both favorable and unfavorable interactions.

Methods

Selected experiments and reference genomes

All pairs of host-symbiont species used in this work are shown in **Table 5**. Additional information of the sRNA-Seq data processing from these experiments is included in **Supplementary Table 1**. The reference genomes used are found in **Supplementary Table 2**. To ease finding ambiguous reads across both genomes, a combined reference was produced by concatenating the sequences from both genomes for each host-symbiont pair. Ribosomal sequences were included as an extra contig in those cases where they were missing. A two-word label was added to all fasta headers to readily distinguish symbiont from host genome sequences. All combined reference files were indexed using Bowtie-1.2.2 (B Langmead et al. 2009).

Processing of small RNA-Seq reads

We used FastQC to inspect the quality of sRNA-Seq reads from each library. We then used reaper (Davis et al. 2013) to trim the 3’ adapter sequence and remove low quality nucleotides with the following parameters: -geom no-bc, -mr-tabu 14/2/1, -3p-global 12/2/1, -3p-prefix 8/2/1, -3p-head-to-tail 1, -nnn-check 3/5, -polya 5 -qqq-check 35/10, -tri 35. Remaining sequences shorter than 18 nt were discarded. When needed, reads were collapsed to unique individual sequences with counts using tally (Davis et al. 2013). One replicate for IECs control cells (incubated for 24 hours without any nematode treatment) was an outlier according to PCA analysis, and it lacked a clear peak for mouse miRNAs, and was thus excluded from further analyses. All the other 17 IECs libraries showed a clear miRNA peak and were considered for further analyses.

Calculations of host, symbiont and ambiguous reads

All libraries were mapped to the separate host and symbiont genomes using Bowtie-1.2.2 (B Langmead et al. 2009) and requiring perfect end-to-end hits (-v 0). Each read was classified as: host if it only mapped to the host genome, symbiont if it only mapped to the symbiont reference or ambiguous if it mapped to both genomes. Read length distributions showing this categorization for all symbiosis sets are shown in **Supplementary Figure 1**.

Shared k-mers between genomes

The fraction of shared k-mers (s) of length 12-30 in two random genomes of fixed sizes was calculated with the following equation:

$$S = \frac{Nab}{Na+Nb-Nab}$$

Where Na and Nb represents the number of k-mers in two sets of random k-mers or genomes a and b. Nab is the number of shared k-mers in both genomes. The values of Na, Nb and Nab were calculated using the theoretical approach given by (Fofanov et al. 2004).

The fraction of shared k-mers between sizes 12-30 that are shared between each pair of real genomes was calculated using Jellyfish 2.2.10 (Marçais and Kingsford 2011).

Genome-guided sRNA assembly

We used ShortStack/3.8.5 (Shahid and Axtell 2014) to perform the genome-guided sRNA assembly. We used parameters that favor small clusters (

Table 4): a minimum coverage of one read, requiring 0 mismatches, using unique-mapping reads as a guide to assign multi-mapping reads (mmap: u), a padding value of 1, reporting all bowtie alignments (bowtie_m: ‘all’), and a ranmax value of 50000 to avoid losing reads mapping to multiple sites. The default bowtie cores and sorting memory values were also increased to improve processing time. Reads were aligned to the concatenated host and symbiont reference genomes mentioned above.

Table 4. ShortStack parameters used for genome-guided assembly

Parameter	Value	Default	Explanation
ranmax	50000	3	reads with more than this number of mapping positions and no guidance is possible (with uniquely mapping reads) will be dropped and marked as unmapped
bowtie_m	'all'	50	Maximum number of mapping sites to be considered for each read
dicermin	20	21	Minimum read size range for a cluster to be considered as Dicer producing locus
dicermax	24	24	Maximum read size range for a cluster to be considered as Dicer producing locus
mincov	1	5	Minimum read coverage for a locus to be considered for cluster definition. Can also be expressed as reads per million (rpm)
pad	1	75	Maximum nucleotide distance between two clusters to be merged into a single cluster
sort_mem	40G		RAM memory used during alignment file sorting
v	0	1	Number of maximum mismatches to allow between read and alignment site
cpu	8	1	Number of cpus to use by bowtie.

De novo assembly of sRNA-Seq

Six popular RNA-Seq *de novo* assemblers were used to evaluate the *de novo* assembly of sRNA reads: Oases (Schulz et al. 2012), rnaSpades (Bushmanova et al. 2019), SOAPdeNovo-Trans (Y. Xie et al. 2014), Tadpole, TransAbyss (Robertson et al. 2010) and Trinity (Grabherr et al. 2011). These assemblers were also tested using their first "k-mer extension" step: a) rnaSpades "--only-assembler", Trans-AbySS "--stage contigs" and Trinity "--no_run_chrysalis"; b) the equivalent for Oases was to use contigs generated by velvetg, while for SOAPdenovo-Trans the -contig output file was used; c) Tadpole is a simple assembler that only performs k-mer extension. All the generated contigs were post-processed as follows: 1) all reads used to generate the assembly were aligned back to the contigs using Bowtie-1.2.2 (-v 0), and 2) using the BAM files from these alignments, contig edges that did not have any reads mapping to them were trimmed. All contigs were then mapped to the concatenated reference genomes to decide if they were host or symbiont.

Disambiguation of host-symbiont mixed samples

To help determine the origin of ambiguous reads (that map equally well to both genomes) we used *de novo* assembled contigs or genome-guided clusters. Clusters are defined directly on a specific genome; therefore, they are non-ambiguous by definition. Contigs are assembled in absence of a genome, but as they are longer than reads, they should be less ambiguous. We mapped contigs to genomes, first with Bowtie-1.2.2 (B Langmead et al. 2009) to find perfect hits, and unmapped contigs were then aligned with a more relaxed search with Bowtie2-2.3.3 (Ben Langmead and Salzberg 2012) allowing for a small number insertions, deletions and mismatches in end-to-end mode with parameters -D 15 -R 2 -N 0 -L 22 -i S,1,1.15. For those contigs that mapped imperfectly to both genomes, the alignment with less mismatches was chosen (XM:i<N> SAM optional field). We then mapped all reads to contigs or clusters. The following procedure was applied to both contigs and clusters. Here we will refer to them as assemblies.

We classified all mapping reads into three sets: those that map to multiple assemblies (multi-mapping reads), reads that map to a single assembly (unique-mapping reads) and reads that do not align to any sequence. With these conditions this is a similar problem as that of assigning multi-mapping reads to transcript isoforms. Tools such as ERANGE (Mortazavi et al. 2008), a method developed for CAGE (Faulkner et al. 2008), RSEM (B. Li et al. 2010) and ShortStack (Shahid and Axtell 2014) use unique-mapping reads as "guide" reads to assign reads that map to multiple transcripts. The core idea is that the proportion of   unreads is a good estimate of the proportion of multi-mapping reads produced by each transcript. In our implementation we first sum the counts of all uniquely mapping reads for each assembly across all libraries, getting global unique-mapping read counts. To use only the most informative assembly with global uniquely mapping read counts, we filtered the top 0.2% (we tested different thresholds and ~90% of multi-mapping reads can be assigned with this threshold). We later distribute multi-mapping reads that align to these assemblies proportionally to global uniquely mapping counts. Reads that map to other assemblies, as well as reads that map to ambiguous assemblies or do not map at all, remain ambiguous.

Results

Species systems that exchange small RNAs included in this study

For this study we considered most of the reported models to date for sRNA-mediated communication, listed in **Table 5**. These include the model plant *A. thaliana* infected with the fungus *Botrytis cinerea* or the parasitic plant *C. campestris*. The rodent *Meriones unguiculatus* infected with the filarial nematode *Litomosoides sigmodontis*. Also, an experiment designed through our collaboration with Amy Buck: *Mus musculus* cells treated with EVs isolated from the parasite model nematode *Heligmosomoides bakeri*. We also included one mutualism representative with the legume *Glycine max* nodules containing *Bradyrhizobium japonicum*.

Table 5. Small RNA sequencing datasets of interacting organisms

Host	Symbiont	Tissue or condition	Data availability	Reference
A. thaliana	*Botrytis cinerea*	Rosette leaves: 24, 48 and 72 hours after infection	Sequence Read Archive: SRP019801. Samples: SRX252403, SRX252404, SRX252405	(Weiberg et al. 2013)
A. thaliana	*C. campestris*	*A. thaliana* stems 4cm above a *C. campestris* haustorium	Sequence Read Archive: SRP118832. Samples: SRX3214812, SRX3214813	(Shahid et al. 2018)
Meriones unguiculatus	*Litomosoides sigmodontis*	Serum from infected gerbils	GEO: GSE112949. Samples: GSM3091975, GSM3091976, GSM3091977, GSM3091978, GSM3091979	(Quintana et al. 2019)
Mus musculus	*Heligmosomoides bakeri*	MODE-K and BMDM cell lines: 4 and 24 hours after adding EVs or total HES	GEO: GSE124506. Samples: GSM3535462, GSM3535463, GSM3535464, GSM3535468, GSM3535469, GSM3535470	This work
Glycine max	*Bradyrhizobium japonicum*	10 and 20 days nodule	Sequence Read Archive: SRP164711. Samples: SRR7986783, SRR7986788	(Ren et al. 2019)

Each of these symbiosis systems presents different properties. A major difference is that there are samples where the actual parasite/symbiont was present during the RNA extraction procedure. In these samples we expect to find both the endogenous signal from each organism, as well as secreted RNAs. In contrast, there are samples where we expect only secretion products (excretory-secretory products) such as those of *Litosomoides* and *Heligmosomoides*.

In the *A. thaliana-Botrytis* system leaves were inoculated with fungal spores, this means that the fungus, a potential sRNA source, is found in the leaves. Here we expect a dominant host signal and perhaps a increasing parasite signal as the infection progresses.

Axtell and collaborators *C. campestris* interaction libraries consists of the portion in which *C. campestris* holds and penetrates *A. thaliana* stems (Shahid et al. 2018). While negative controls consist of *A. thaliana* stems 4 cm above the *C. campestris* haustorium (root-like organs that penetrate the host plant tissue and may kill it) and parasite libraries consist of the *C. campestris* stem 4 cm below the haustorium. We will consider *A. thaliana* stems as the interaction library of interest, as those *C. campestris* sequences found here may be considered actual parasite sequences that reached this far inside *A. thaliana* stems.

M. unguiculatus samples consist of host serum infected with *L. sigmodontis*. Here the filarial nematode is absent from samples as these filarial nematodes remain in the host's pleural cavity and we expect to detect mostly excretion-secretion sRNAs. For *Heligmosomoides bakeri-Mus musculus*, mouse cells were incubated with *H. bakeri* EVs or HES. For both the *L. sigmodontis* and *H. bakeri* libraries the producing nematodes are absent from the samples. We expect the rodent samples to be similar to a needle in a haystack situation, where small amounts of parasitic reads are outnumbered by those of the host.

The *Glycine-Bradyrhizobium* samples consists of nodules harboring bacteria at two different time points. Here, the symbiont cells are included and we expect to find some bacterial endogenous signal.

Determining the amount of host, symbiont, and ambiguous reads in sRNA datasets

For each organism's system we built a mixed reference that included both host and symbiont genomes. Then we mapped reads between 18 and 50 nt to our merged reference and worked with only those reads that matched perfectly to one or more sites. We decided to keep only perfect mapping read to help guide the evaluation of our assembly processes: any resulting assembled sequence that doesn't map back to the genome would be considered an assembly chimera. An implication for this decision is that we will loose potentially interesting reads, such as those that had additional nucleotides due to RNA post-transcriptional editing, those derived from heterozygous sites in the genome, or simply those with a small number of sequencing errors. Preliminar explorations with adult libraries suggest that RNA post-transcriptional editing happens in *Heligmosomoides*. I considered the mismatch position for all sequences with a single mismatch. The most abundant position for a mismatch to occur is the final 3' position, which suggests that an extra nucleotide was added to these sequences at the 3' end. I observed this effect in both phosphatase-untreated (mono-P) and phosphate-treated (poly-P) adult libraries and through a broad read size range (18 nt – 30 nt) **Supplementary Figure 2**. Focusing on reads 21-24 nt long, I observed a bias torwards uracil/thymine at the 3' extreme for mono-P libraries, whereas, for poly-P libraries I found a bias torwards uracil/thymine and cytosine. It would be valuable to include pure *H. bakeri* EV libraries and *Caenorhabditis* libraries for further explorations to look for evidence of these observations in secreted products and to look for evolutionary conservation. Nevertheless, the priority for this part of the thesis was to propose assembly and disambiguation strategies, so the phenomenon of RNA editing will need to be considered in future work.

A read was considered to be parasitic (or mutualistic for *Bradyrhizobium*) if it mapped only to the symbiont genome, a read was considered to be a host read if it mapped only to the host genome, while reads that mapped equally well to host and symbiont were considered ambiguous.

There are different amounts of ambiguous reads (**Figure 6**). For *Botrytis* and *A. thaliana*, ambiguous reads are more abundant at 48 and 72 hours (6.8% and 6.3% respectively) than at 24 hours (1.1%). This highlights the importance of sRNAs that map perfectly to both organisms, these sRNAs may be related to infection progression. Ambiguous reads may be a strategy for *Botrytis* to target genes in *A. thaliana* or these may be produced by the plant to respond to the fungal infection. It may be interesting to compare possible producing loci for these ambiguous sequences in the fungal and plant genome.

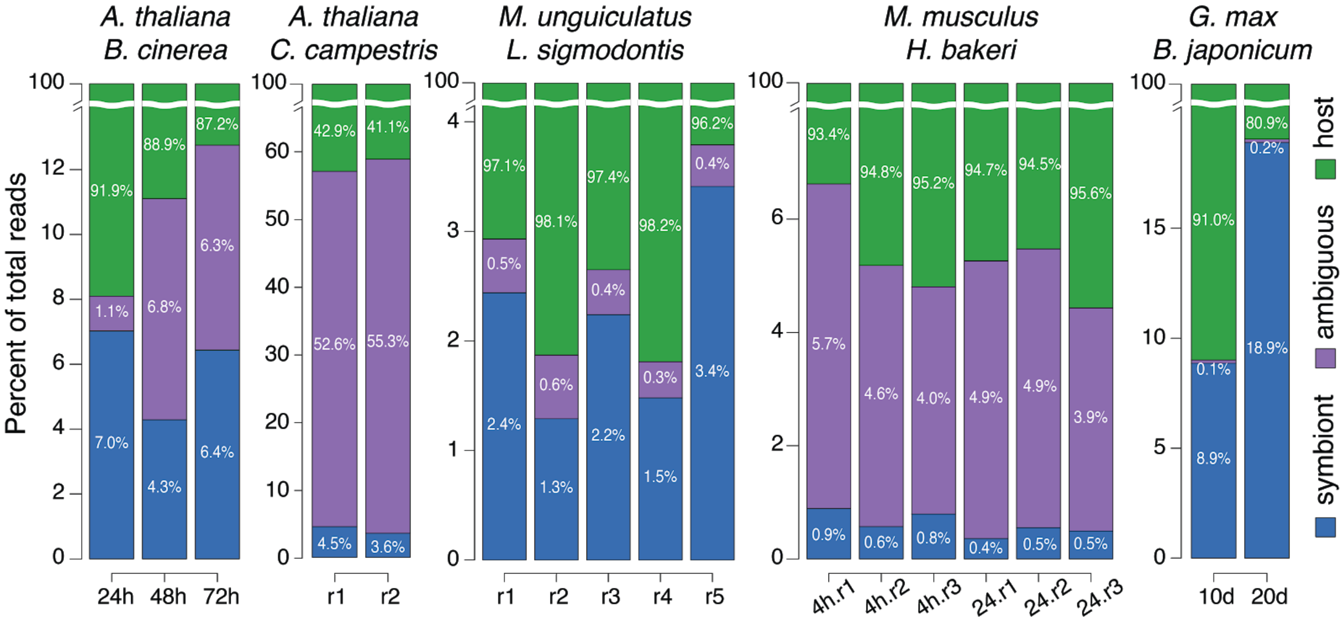


Figure 6. Fraction of ambiguous and symbiont reads for interacting libraries. The name of the two interacting species is shown for each experiment. Each bar represents all 18-50 nt reads from one sRNA-Seq sample, and bars are grouped by experiment. The Y-axes are independently zoomed and cut to to highlight the percent of symbiont (blue) and ambiguous (purple) reads. Host reads (green) always add up to 100%. Biological replicates are defined by "r", while other labels are hours post infection (*B. cinerea*), hours of incubation with EVs (*H. bakeri*) and days of nodule (*B. japonicum*).

The *C. campestris* - *A. thaliana* symbiosis system is the one with the most ambiguous reads: up to 55.3% of the reads map equally well to both plants. *C. campestris* signal ranges from 3.6% to 4.5%. This parasite signal is considerable given that these libraries are *A. thaliana* stem libraries 4 cm away from the *C. campestris* haustorium-*A. thaliana* stem interaction site. This suggests that *C. campestris* sRNAs may travel through *A. thaliana* vascular system and may reach other parts of the host plant. Ambiguous reads are 11- and 15-times more abundant than the unambiguous parasite reads. In their analysis of this data, Shahid and collaborators focused on enriched sRNAs present in interaction (*C. campestris*-*A. thaliana*) relative to parasite (*C. campestris*) libraries (Shahid et al. 2018). They did not explore possible parasite signal in host (*A.thaliana*) libraries.

For *Meriones* and *Litomosoides* the fraction of ambiguous reads ranges between 0.3% to 0.6% of mapped reads. Parasite signal in serum ranges from 1.3% to 3.4%. For this system parasite signal is higher than ambiguous reads. For this same data, Quintana and collaborators (Quintana et al. 2019) found 1.6% to 4.2% parasite signal and 0.3% to 1% ambiguous reads. There are several differences between our data processing and Quintana *et al*. that may explain these differences. Some processing differences include using a different reference genome, Quintana *et al*. used *M. musculus* and we used *M. unguiculatus*, we used reaper as an adapter trimming tool and Quintana *et al*. used cutadapt, just to mention a few examples. As additional context for filarial nematode sRNA data in serum, *Onchocerca* signal in serum from infected patients was between 5 and 127 reads/per million total host miRNA reads, which corresponds to up to 0.0127% of parasite signal (Quintana et al. 2015).

For mouse cells treated with *H. bakeri* EVs the amount of ambiguous sequences clearly outnumbers those reads that map perfectly to the nematode. Ambiguous reads range between 3.9% to 5.7% while parasite signal reaches up to 0.9% of the total reads. Thus, ambiguous reads are around 5 times more abundant than nematode reads. If we were to discard all ambiguous sequences, this would mean loosing a substantial amount of the reads that are potentially parasitic.

The system with the smallest fraction of ambiguous reads is *G. max* and *Bradyrhizobium* with 0.1% and 0.2%. This low amount of ambiguous reads could be due to the enormous phylogenetic distance found between bacteria and plants. On the other hand, this pair of organisms displays the greatest symbiont signal with 8.9% and 18.9% of the reads mapping to the bacterium genome. This symbiont-rich signal could the the result of high numbers of rhizobial bacteria in soybean nodules. Notheworthy, this is the only mutualistic relationship studied, it would be interesting to determine if bacterial sRNA contents relate to the number of bacterial cells in nodules.

A common practice is to discard or ignore reads that map equally well to both organisms due to the difficulty of determining their true origin. However, if we were to discard ambiguous sequences, we would be throwing away a substantial amount of data, especially for the *A. thaliana-C. campestris*, *A. thaliana-Botrytis*, and *Heligmosomoides*-*Mus musculus* systems. It is a possibility that sRNAs relevant to the symbioses would be discarded.

Ambiguity in host-symbiont sRNA-Seq reads is influenced by read length, genome size and phylogenetic distance

We wanted to get more insights into the origin of ambiguous sequences in our chosen host-symbiont systems. Previous studies have addressed similar problems showing that read length, genome size and phylogenetic relationships are relevant (Fofanov et al. 2004). In this section we address these factors using “k-mers” (nucleotide sequences of length k) as a proxy for sRNA-Seq reads. Results of this section were obtained by Dr. Obed Ramírez-Sánchez, a postdoctoral researcher (2018-2020) in the Abreu-Goodger lab.

Read length

Read length is a key factor for k-mers to be shared between genomes. To show this we used an estimation of shared k-mers for two random genomes with the same sizes as the *A. thaliana* and *Botrytis* genomes, and calculated the fraction of shared k-mers across different values of k. The number of shared k-mers decreases as k increases (**Figure 7A**). Nearly 80% of k-mers of length 12 are shared between these two genomes, in stark contrast to only around 0.25% of the k-mers of length 18 being shared. Just by adding a single nucleotide, the chance of a k-mer being shared decreases substantially.

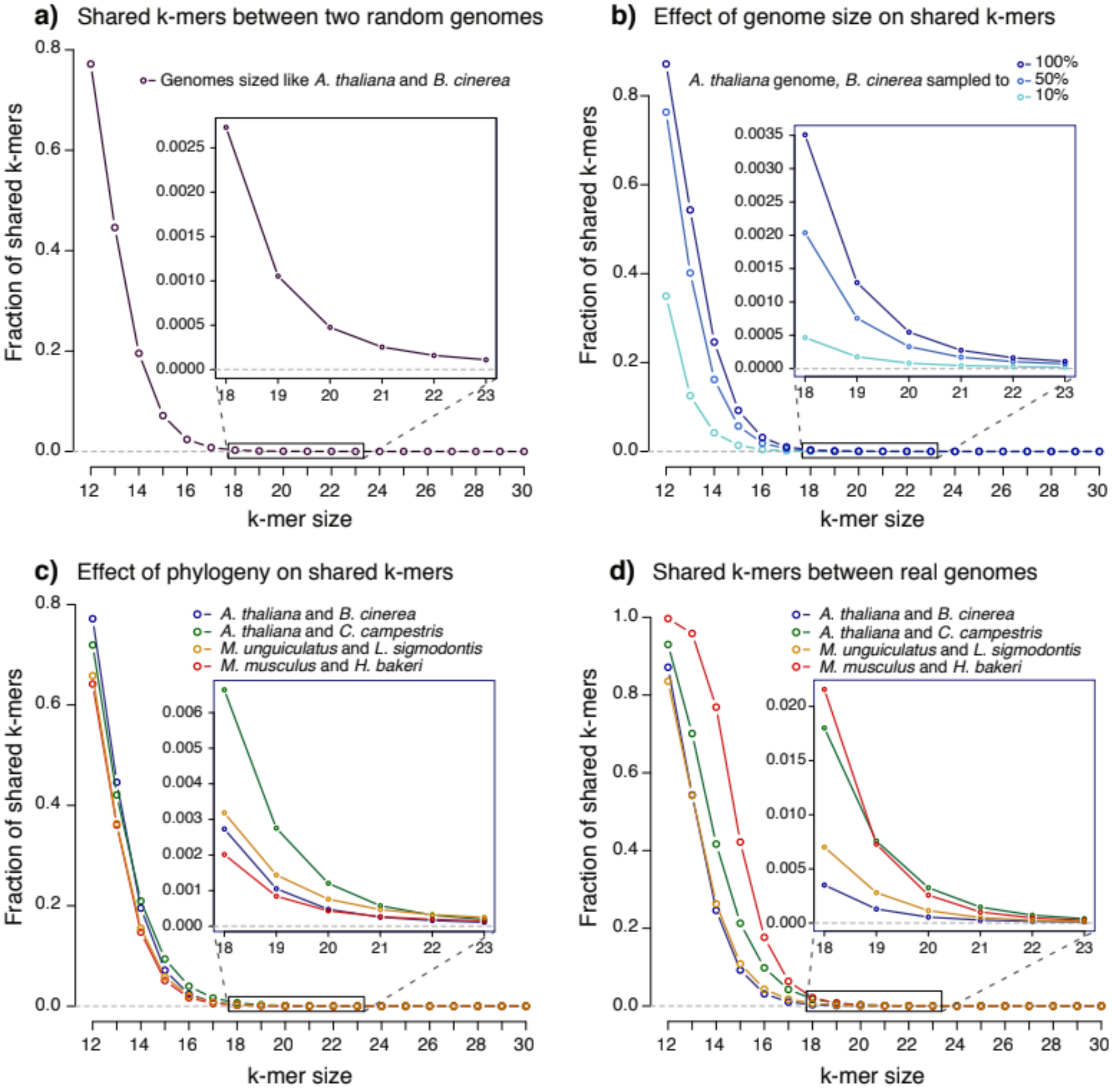
Genome size

Another factor to consider is genome size. The larger two genomes are, the more likely that they will share k-mers. This was tested by down-sampling *B. cinerea* genome to 50% or 10% of its length, and measuring the shared k-mers with *A. thaliana*. The smaller the sample the less shared k-mers (**Figure 7B**). Genome size clearly affects the amount of shared k-mers.

Phylogenetic distance

Genomes are related by common ancestry. The closer two organisms are, the more k-mers they will share. In order to compare the effect of phylogenetic distance, but controlling for genome size which also affects shared k-mers, we first down sampled each genome to match the size of *Bradyrhizobium japonicum* (9.1 Mb), the smallest genome included (**Figure 7C**). *C. campestris* and *A. thaliana* are two plant genomes belonging to the eudicots clade, and are the more closely related genomes. This is reflected by being the two genomes that share the most k-mers. *Meriones* and *Litomosoides* are two animal genomes and come second in shared k-mers. *A. thaliana* and *Botrytis* and *M. musculus* and *H. bakeri* have similar low fractions of shared k-mers. This is expected for *A. thaliana* and *Botrytis* having a long diverging time, 1,576 +/- 88 Ma (Wang, Kumar, and Hedges 1999), however *Heligmosomoides* and mouse are two animal genomes and this low proportion may appear to be unexpected. We interpret that this low proportion is likely because the genome of *Heligmosomoides* is rich in repetitive sequences exclusive to this organism, 58.3% of its genome is composed of repetitive elements (Chow et al. 2019). A random sample of this nematode genome will be enriched in these unique sequences that are less likely to provide shared k-mers. The low amount of shared k-mers between *Heligmosomoides* and mouse reveals other relevant factors for shared-kmers such as genome composition and genome complexity, which will not be further explored in this thesis. Lastly, *Glycine* and *Bradyrhizobium* have the longest diverging time and this is reflected in the smallest proportion of shared k-mers.

We calculated the real number of shared k-mers for each pair of full-length genomes with jellyfish (see Methods). In stark contrast with the previously mentioned down-sampling approach, *H. bakeri* and *M. musculus* share the highest number of k-mers (**Figure 7D**). These are the two biggest genomes being 700 Mb and 3.2 Gb, suggesting that genome size is the major determining factor for shared k-mers. *Bradyrhizobium* is by far the smallest genome (9.1 Mb size) compared, and it shares the fewest k-mers with its host *G. max,* even though the soybean genome is large (0.97 Gb). *C. campestris* and *A. thaliana* come second regarding shared k-mers. *Meriones* and *Litosomoides* are in third place, and last comes *A. thaliana* and *Botrytis*, with relatively small genomes and large phylogenetic separation.



a) Shared k-mers between two random genomes

b) Effect of genome size on shared k-mers

c) Effect of phylogeny on shared k-mers

d) Shared k-mers between real genomes

Figure 7. Factors influencing the number of shared k-mers between pairs of genomes. In all figures X-axes represent k-mer size and Y-axes represent the fraction of shared or ambiguous k-mers. A) Random genomes of sizes equivalent to *B. cinerea* and *A. thaliana*. B) Shared k-mers between A. thaliana genome and downsampling of *B. cinerea* genome to 50%, 10% or complete genome. C) Downsampling of all genomes to the *B. japonicum* genome size. D) Actual fractions of ambiguous k-mers in each pair of complete genomes. Zoomed regions cover k-mer sizes of 18-23.

For all four factors analyzed in the current section (**Figure 7)**, the longer the k-mers are, less likely that these will be shared between two organisms. This observation suggests that increasing small RNA sequence lengths would help diminish the fraction of ambiguous reads. One way of leaveraging this observation would be to assemble small RNA sequence reads before attempting to assign them to each genome, and this idea will be further explored in this thesis. But, we first wondered if the ambiguous small RNA reads come from particular places in the genome.

miRNAs, transfer and ribosomal RNAs are major contributors to ambiguous reads

We wanted to better characterize the origin of ambiguous sequences. To do so we focused on the more recently diverged genomes of *C. campestris* and *A. thaliana,* using the sRNA-Seq data generated by Shahid and collaborators in Axtell's lab (Shahid et al. 2018). We then extracted all the ambiguous sequences and determined their loci of origin in the *A. thaliana* genome, which is much better annotated than *C. campestris* (**Figure 8**).

Plant sRNA profiles typically have peaks at 21 and 24 nucleotides, which correspond to miRNAs and siRNAs respectively. These characteristic peaks are not that evident in *A. thaliana* stem libraries. Instead, these libraries show an enrichment for shorter sequences and depletion of longer sequences which suggests some degradation of the RNA sample (**Figure 8A**). The 24 nt peak is more evident in all libraries containing *C. campestris* tissue. In the *A. thaliana/C. campestris* libraries the 24 peak represents in average 21.9% of the library (**Figure 8C**) and in *C. campestris* libraries it represents 19.6% (**Figure 8E**). In contrast, in the *A. thaliana* stem library the 24 nt peak represents only 6.9% of all the library (**Figure 8A**). In the *A. thaliana/C. campestris* or parasite libraries the 24 nt peak is clearer than the 21 nt peak, this suggests that siRNAs are the most abundant sRNA class in *C. campestris*. More than half (53%) *A. thaliana* stem library are ambiguous reads, while these represent only 25% in *A. thaliana/C. campestris* libraries and 20.7% in parasite libraries.

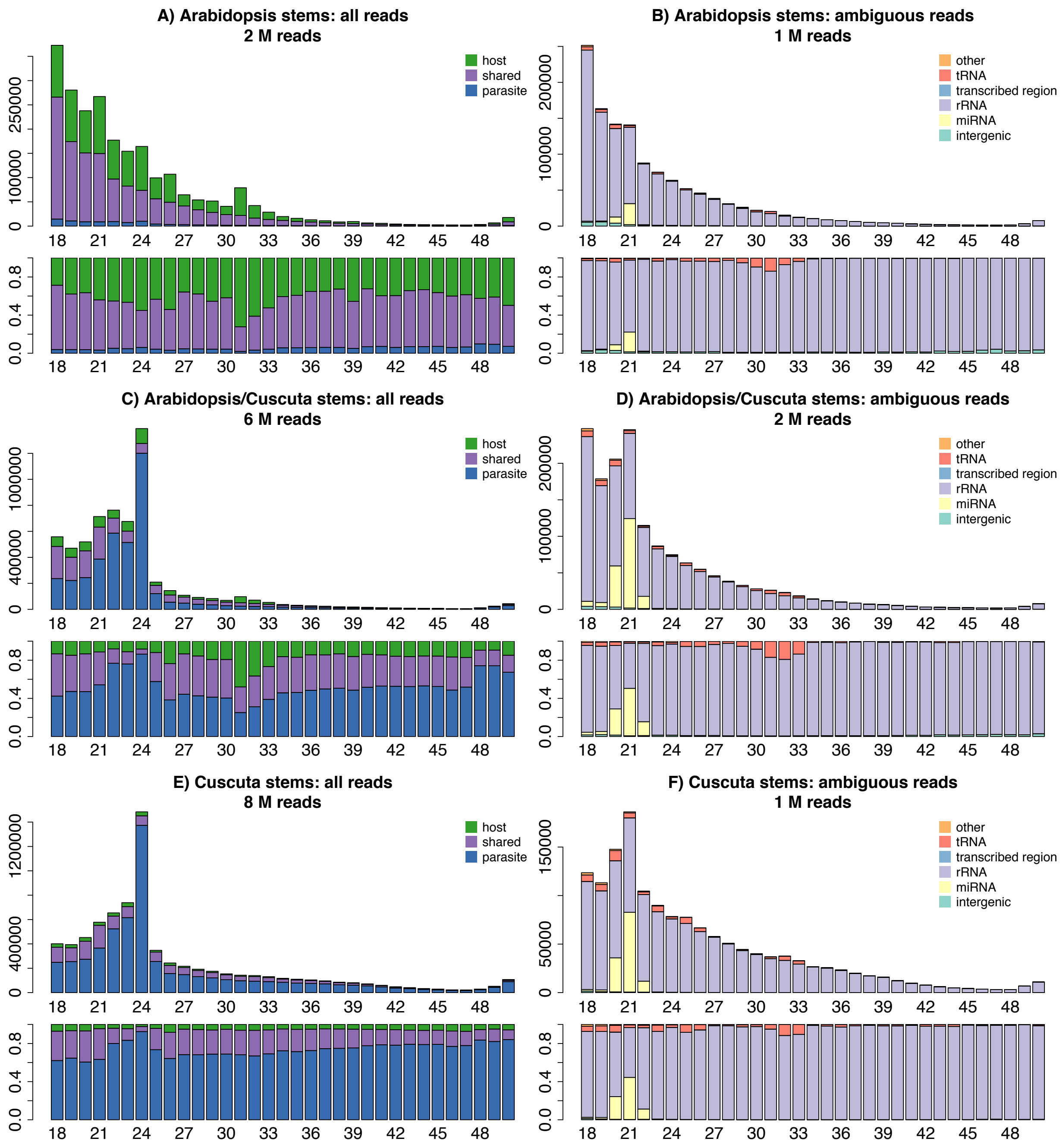


Figure 8. Genomic origin of ambiguous reads for *A. thaliana* and *C. campestris* samples. Each bar represents the sequenced reads of 18-50 nt size. Bar height represents the actual number of reads (top) or the fraction of reads (bottom). A), C) and E) Read mapping categories split according to read length: host (green), symbiont (blue) or ambiguous (purple). B), D) and F) Genomic annotation of ambiguous reads only: intergenic (light green), miRNA (yellow), rRNA (light purple), tRNA (red), uncharacterized transcribed regions (light blue) or other annotation (orange). Libraries were made from A) and B) *A. thaliana* stems above the site of primary haustoria, C) and D) *A. thaliana* stems with a *C. campestris* haustorium attached, and E) and F) from *C. campestris* stems above the site of primary haustoria.

We then explored the regions that produce ambiguous small RNA sequencing reads by extracting the annotations associated to loci producing ambiguous reads in the *A. thaliana* genome. We used *A. thaliana* annotation because, as it is the most intensely studied plant, we reasoned that it should have better annotations than *C. campestris*.

Ambiguous reads from the *A. thaliana* stem libraries originate mostly from rRNA (91.6%), miRNA (3.2%) and to a lesser extent tRNA loci (2.6%). Ribosomal signal is distributed across all read lengths (**Figure 8B**). This suggests that these reads may be the result of fragmentation of longer ribosomal molecules. As expected, miRNAs are found between 20 to 22 nucleotides, albeit at a rather slight level of detection. This low level of detection for miRNAs may explain why the 21 nucleotide is not that evident in **Figure 8A**. Regarding other libraries, a 21 nt peak becomes evident when considering only ambiguous reads for the *A. thaliana/C. campestris* (**Figure 8D**) and parasite libraries (**Figure 8F**). The 24 nt peak is not evident when tracing the origin for ambiguous reads in *C. campestris* libraries, this suggests that siRNAs are not well conserved between *A. thaliana* and *C. campestris*, and thus do not form an important part of the ambiguous reads. There's more miRNA signal for *A. thaliana/C. campestris* (13.1%) and parasite (8.8%) libraries than in the host library (3.2%).

Shahid and collaborators focused on sRNAs enriched in the *A. thaliana/C. campestris* relative to *C. campestris* libraries, they found 43 miRNAs that met this criterium. One of these belongs to the conserved MIR164 family, the other 42 have low sequence similarity with known miRNA loci. Neither the mature nor the complement sequence map perfectly to the host genome. In our results, these 42 sequences should be found in the parasite portion of the plots (blue bars). The majority of the interaction-enriched miRNAs are 22-nt, and these are an uncommon size since plant miRNAs are typically 21-nt in length. Plant 22-nt miRNAs are associated with secondary siRNAs production that amplify the silencing signal (H. Chen et al. 2010).

There is 22 nt miRNA signal in *C. campestris-A. thaliana* (14.4% of 22 nt signal) and *C. campestris* (10.7% of 22 nt signal) libraries **Figure 8D** and **Figure 8F**, and this signal is not evident in *A. thaliana* libraries (0.8% of 22 nt signal). These results suggest that these ambiguous 22 nt miRNAs are produced by *C. campestris* but can map to *A. thaliana* genome. These 22 nt miRNAs should not be any of those 43 miRNAs reported by Shahid *et al*. as those do not map to *A. thaliana* and therefore are not ambiguous. We find conserved plant miRNAs among ambiguous reads across all *A. thaliana-C. campestris* libraries such as MIR159, MIR319a, and MIR396a. MIR159 is often among the most abuntant RNAs, it is expressed throughout the plant, but displays higher expression in shoot and root meristematic regions (Millar, Lohe, and Wong 2019). MIR319 displays lower expression than MIR159 and it is restricted to specific tissues and developmental stages (Y. Li et al. 2011). MIR396 has been associated with plant immunity regulation and fungal resistance (Soto-suárez et al. 2017).

It is very notorious how a few RNA categories that represent a minimal fraction of the genomes produce the majority of the ambiguous sRNA reads. Ribosomal RNA loci represents 0.081% of the *A. thaliana* genome and comprises between 67.5% to 90.9% of ambiguous reads depending on the library, this translates to a 833 to 1,222 fold enrichment. miRNAs primary transcript loci cover 0.045% of the *A. thaliana* genome and contribute between 2% to 27% of ambiguous reads, this translates to a 44 to 600 fold enrichment. Finally, tRNAs comprise 0.047% of the *A. thaliana* genome and contribute between 2.4% to 4.3% of ambiguous reads, this represents a 51 to 91 fold enrichment.

For many organisms, such as vertebrates, a few RNA classes contribute greatly to the sRNA transcriptome such as miRNAs, ribosomal RNA, tRNAs, specially in somatic tissue (Armisen et al. 2009). At least a fraction of these loci may be conserved through several clades. However, there are other organisms in which siRNAs are the dominant sRNA class instead of miRNAs, such as *H. bakeri*, where repetitive elements are a great contributor (~40% of an adult nematode library) to the sRNA transcriptome (Chow et al. 2019). Repetitive elements typically display faster evolution rates than housekeeping loci such as rRNA, tRNAs or conserved miRNAs, therefore it is less likely that repetitive elements contribute to the pool of ambiguous sequences.

Ribosomal RNAs and transfer RNAs are most famous due to their role in protein synthesis. However, recent studies suggest additional functional roles for both rRNAs and tRNAs. Some examples of rRNA-derived fragments (rRFs) have been found in human, mouse, zebrafish, plants and fungi (Lambert, Benmoussa, and Provost 2019). Most of the rRFs described so far are generated by RNAi machineries such as those involved in phasiRNAs, piRNAs or miRNAs production (Lambert, Benmoussa, and Provost 2019). For example, mouse miR-712 is produced from the ribosomal internal transcribed spacer 2 (ITS2), a subset of this sequence forms a stem-loop structure that is recognized by the miRNA producer machinery. The inhibition of miR-712 with antisense oligos had a direct effect on plaque size in a atherosclerosis (arteries flow obstruction) mouse model, showing functional roles for rRFs (Son et al. 2013). Regarding tRNAs, Chiou and collaborators discovered that activated T cells secrete EVs enriched with specific tRNA-derived fragments (tRFs). The enrichment for 3’ or 5’ tRFs was not evident in resting T cells, which suggest that activation is necessary to load these fragments into EVs. These tRFs may inhibit T cell activation as antisense oligos targeting these EV-enriched tRFs resulted in enhanced activation, although the authors do not predict targets for their tRFs (Chiou et al. 2018). Another study reported tRFs produced by *Bradyrhizobium* that get loaded in the soybean Argonaute and downregulate host genes resulting in nodulation modulation (Ren et al. 2019). These examples suggest that by throwing away rRNA and tRNA sequences we could lose some interesting sRNAs that may be important for the symbioses.

To discard ambiguous miRNA sequences would be even worse. There are reports of highly conserved mature miRNAs that are transferred between organisms that can be identical (Buck et al. 2014). These could really be relevant sequences for interaction and to discard them would be neglecting part of a biological phenomenon. Examples of highly conserved miRNAs that can have identical sequences in some plants include MIR155, MIR159, MIR166, etc (Chavez-Montes et al. 2014). Examples of highly conserved miRNAs in animals include miR-100 and let-7, these mature sequences are identical in *Heligmosomoides* and mice (Buck et al. 2014).

On the other hand, it is important to mention that even ultra-conserved sequences may have point differences either in their functional form or in a longer precursor. For example, miRNAs have a hairpin precursor structure from which the mature miRNA is excised. The mature sequence tends to display higher conservation than the other portions of the hairpin.

Current sequencing technologies generate sufficient depth to detect variants of sRNAs. Small RNA variations may arise from imperfect enzymatic activities, such as imperfect cleavage by a Dicer protein. For miRNAs, the mature miRNA is typically the most highly detected sRNA. However, at high sequencing depths other components such as the miRNA complement or fragments of the hairpin can be detected. Some tools, such as miRDeep2, rely on the detection of these precursors to predict novel miRNAs from high throughput sequencing data.

As shown in the previous section, longer sequences are less likely to be shared between two genomes just by chance. If we could extend ambiguous sequences even by one or two nucleotides, we may be able to disambiguate them. This leads us back to the idea of assembling sRNAs in order to help distinguish their correct genome of origin.

Small RNA sequencing assembly

Most efforts regarding the assembly of RNA molecules have focused on messenger RNAs. Tools such as Cufflinks (Trapnell et al. 2010) and Stringtie (Pertea et al. 2015) perform mRNA assembly by first aligning reads to a genome sequence. On the other hand, there are tools for *de novo* mRNA assembly such as Trinity, SOAPdenovo and Trans-ABysSS. Previous efforts for sRNA assembly include the detection of viruses in sweetpotato in 2012 by Kashif and collaborators (Kashif et al. 2012) and the genome assembly of a bell pepper endornavirus by Sela and collaborators (N. Sela, Luria, and Dombrovsky 2012). Kashif and collaborators used velvet to assemble reads of lengths 21-24 nt, the assembled contigs were used to retrieve NCBI sequences via BLAST searches. The viral sequences retrieved were then used as references to align sRNA reads with MAQ and to build assemblies with these mappings. This combined strategy of *de novo* and genome-guided assemblies led to the identification of six different viruses from a RNA pool of 11 sweetpotato plants (Kashif et al. 2012). Sela and collaborators used BFAST to align reads to a reference viral genome, achieving 100% base coverage in this genome-guided approach (N. Sela, Luria, and Dombrovsky 2012).

Genome-guided assembly tools for sRNAs include segmentSeq (Hardcastle, Kelly, and Baulcombe 2012), the UEA sRNA workbench (Stocks et al. 2018) and ShortStack (Shahid and Axtell 2014). These tools define the boundaries of loci that produce sRNAs in a genome. We will refer to the results of these tools as genome-guided assemblies or simply clusters through the rest of this work. We used ShortStack to define and quantify genome-guided assemblies. ShortStack offers several advantages over other tools for sRNA annotation: with a single command it aligns reads to the reference genome, and defines, quantifies and gathers useful information about each cluster. ShortStack internaly uses ViennaRNA tools to test for hairpin folding capacity of miRNA candidate clusters. Shortack outputs several useful files such as an output table that can be directly used for differential expression analysis (Counts.txt), another table (Results.txt) that contains extensive information about each cluster, such as location, length, strand, the most abundant read for each cluster, complexity and a score for repeated arrangement of aligned small RNAs “phasing”. SegmentSeq provides cluster quantifications and report strand of clusters, but lacks several useful descriptions such as sRNA sizes in cluster, major read in each cluster and completely lacks any secondary structure utility. The UEA sRNA workbench provides several tools to analyze sRNA-Seq data, however to achieve the same functionality as ShortStack it needs to combine several of its tools such as SiLoCo for sRNA cluster definition, quantification, strand and repetitiveness descriptions;

ta-siRNA to test phasing in clusters and miRCat for miRNA annotation. The UEA sRNA workbench uses a graphical user interphase which is not appropriate for high performance cluster computing. Some of their tools have some functionality restrictions, for example SiLoCo accepts only full-length perfect matches to define clusters, while ShortStack allows for 1 or 2 mismatches. Additionaly, SiLoCo cluster quantifications are normalized by default, while ShortStack reports raw counts which are preferable for differential expression analysis tools such as edgeR or DESeq2.

There might be scenarios where a reference genome is not available for a symbiosis system or the genome quality is deficient. In these cases, a *de novo* sRNA assembly would be an attractive alternative to genome-guided assembly, or even the only possibility.

In our group, Obed Ramírez-Sánchez tested 6 tools for *de novo* assembly: Oases, rnaSpades, SOAPdenovo, Tadpole, Trans-ABySS and Trinity-inchworm. Throughout this work, we will refer to the results that he obtained as *de novo* assemblies or simply contigs. Trinity-inchworm performed well in extensive testing such as percent of contigs mapping to genome, reads re-mapping to assembled contigs, etc. and was chosen as the *de novo* tool used in this work (**Supplementary Figure 3**).

sRNA assembly decreased the number of ambiguous reads for all symbiosis pair datasets

We consider ambiguous reads as those that map equally well to the host or symbiont genome. Reads mapping to more than one assembled sequence is a similar problem to that of reads mapping to different isoforms or gene paralogs. This is a problem that has been addressed previously by tools such as ERANGE (Mortazavi et al. 2008), RSEM (B. Li et al. 2010) and ShortStack (Shahid and Axtell 2014). These programs assume that the proportion of mapping reads to a single location can be used to estimate the number of ambiguous reads to distribute to each multimapping loci. We implemented the same principle to assign multimapping reads to a probable producing locus in the host or symbiont genome (see Methods). We applied this idea to both genome-guided (clusters) and *de novo* assemblies (contigs). Many ambiguous reads were successfully assigned with either assembly strategy (**Figure 9**). For the *A. thaliana* - *B. cinerea* pair, 4.9% of the reads are ambiguous in the unassembled approach, and were reduced to 0.04% for contigs and to 0.2% for clusters. Ambiguous reads for the *A. thaliana* - *C. campestris* system are 53.5% for reads, 24.5% for contigs and 51.4% for clusters. For the *M. unguiculatus* – *L. sigmodontis* pair ambiguous reads are 0.4% with the unassembled approach, 0.03% for contigs, and 0.2% for clusters. For *M. musculus* and *H. bakeri* ambiguous reads are 4.7% for unassembled reads, 0.04% for contigs and 0.03% for clusters. For *G. max* and *B. japonicum* these are 0.2% for reads, 0.3% for contigs, and 0% for clusters. In some cases, the percentage of ambiguous reads for the *de novo* assembly is lower than the genome-guided assembly. This could happen if multiple similar clusters could produce a sRNA while a single contig would condense this information for the *de novo* assembly. In a previous effort with reads in the 18-32 nt range clusters outperformed contigs in all cases, this also stresses that contigs assembly benefited from including longer sequences (here we used sequences up to 50 nt long). In almost all cases assemblies performed better than unassembled reads, with a single exception for *G. max* – *B. japonicum* where unassembled reads had 0.2% relative to 0.3% ambiguous reads for contigs.

Regarding parasite fractions, for *A. thaliana* and *B. cinerea* the percentage of parasite signal are very similar for our three approaches, 5.9% for reads 5.9% for contigs and 6.1% for clusters. For the *A. thaliana* and *C. campestris* pair we have 4.2% for reads and an increase with both assembly strategies 28% for contigs and 14.1% for clusters. For *M. unguiculatus* and *L. sigmodontis* the numbers are similar with 2.2% for reads, 2.3% for contigs and 2.3% clusters. For *M. musculus* and *H. bakeri* the smallest fraction of parasite signal is for unsassembled reads with 0.6%, this increased a bit to 0.9% with contigs and increased up to 2% with clusters. Finally, for *G. max* and *B. japonicum* we have even more parasite signal for reads 13.7% than for contigs 12.8%, and similar signal to that of clusters (13.9%). Further work would be necessary to try to relate these differences in parasite or ambiguous reads signal to particular loci in each symbiotic pair.

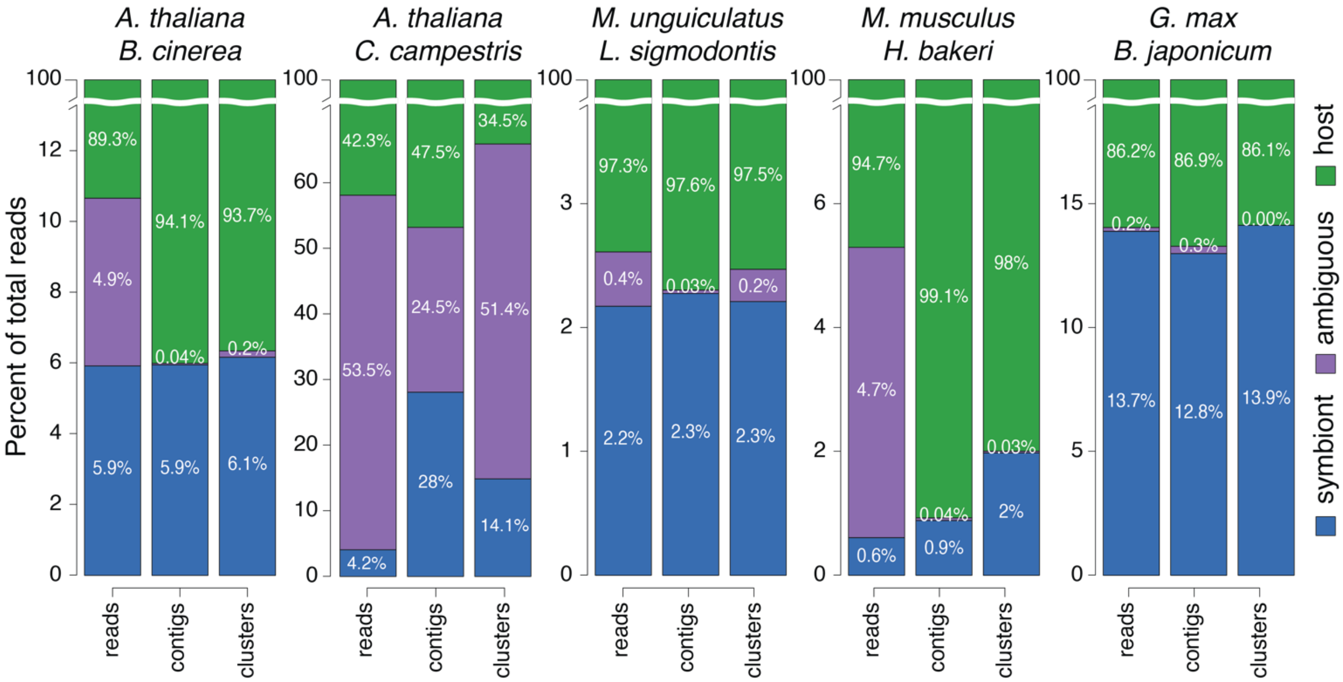


Figure 9. Fraction of ambiguous reads with and without assembly. The name of the two interacting species is shown for each experiment above the sets of three bars. All 18–50 nt reads were classified and the percent of each category were averaged across each experiment's samples. The first bar of each group represents unassembled reads, the second de novo contigs, the third genome-guided clusters. The Y-axes are independently zoomed and cut to highlight the percent of symbiont (blue) and ambiguous (purple) reads. Host reads (green) always represent the remainder of 100%.

It is worth mentioning that ShortStack can reduce ambiguous reads to 0%, however, its multi-mapping reads distribution becomes arbitrary for the last assigned reads. As previously mentioned, ShortStack uses uniquely mapping reads as “guides” to distribute multi-mapping reads. However, there may be some multi-mapping reads that do not have uniquely mapping reads in their surroundings. ShortStack distributes these reads randomly among the multiple mapping sites and by doing so reduces ambiguous reads to 0%. Our approach is more conservative, we also make use of uniquely mapping reads as guides but, when there’s no possible guidance for organism ambiguous reads, these are reported as ambiguous instead of randomly distributing them. Our ambiguous reads distribution approach also allows for fairer comparisons between genome-guided and *de novo* assemblies. Otherwise ambiguous reads for genome-guided assembly would always be 0% and higher for the de novo assembly strategy.

In this chapter we used assembly strategies to determine the origin of reads for sRNA-Seq libraries of mixed organisms. In chapter 2 we apply a genome-guided assembly strategy and a differential expression analysis to identify *H. bakeri* sRNAs in mouse cells.

Conclusions

For some symbiotic pair datasets ambiguous reads are quite abundant and can even outnumber the parasite signal, stressing the need to make use of these reads and to avoid discarding them.

The number of shared k-mers between two genomes is influenced by read length, genome size and phylogenetic distance. Shorter reads may map simply by chance, as read length increases this probability decreases. Bigger genomes may result in more shared k-mers and hence more random mappings than smaller genomes. Phylogenetic distance is also a relevant factor, two genomes belonging to recently diverged lineages share more k-mers than two genomes of lineages that diverged a long time ago. Additional factors influencing shared k-mers include genome complexity and repetitive elements, further work is needed to explore the contribution of these two factors.

High levels of ambiguity in host-parasite sRNA-Seq reads is caused by conserved sequences like ribosomal, transfer and miRNAs in the *A. thaliana – C. campestris* symbiotic pair.

Both genome-guided and *de novo* assembly approaches reduce ambiguity of host-symbiont sRNA-Seq reads. The *de novo* approach is an attractive option for organisms that do not have a genome assembly.

We designed a strategy to disentangle mixed sRNA-Seq data of two or more organisms and tested it on five different pairs of interacting organisms.

Perspectives

1. A description of 3’-added nucleotides across all available *H. bakeri* libraries, this would provide evidence for possible sRNA editing. If we find editing evidence, then we could then ask if editing happens under specific conditions such as HES, EVs or infection, or if it is a generalized phenomenon in this parasitic nematode.

2. Explore the source of ambiguous reads for the other host-symbiont species pairs: *Botrytis cinerea - A. thaliana*, *Litomosoides sigmodontis - Meriones unguiculatus, Bradyrhizobium japonicum - Glycine max.* This would allow us to verify that the same type of highly conserved non-coding RNA loci (rRNA, miRNA, tRNA) contribute substantially to ambiguous reads for pairs of species at diverse phylogenetic distances. Alternatively, we might discover new reasons for higher than expected ambiguity.

3. To explore loci contribution of rRNA, miRNA and tRNA to ambiguous sequences. Based on our results it is unlikely that these three RNA classes contribute equally to the pool of ambiguous sequences, and within each category some loci may contribute more than others.

4. Explore if any rRNA or tRNA has read pileups that would suggest they may be loaded into an Argonaute protein, this would reveal unexplored Ago-loaded sRNA classes for most of these host-symbiont pairs.

Chapter 2: Detection of *Heligmosomoides bakeri* sRNAs in mouse cells

Methods

All experimental procedures were performed by Dr. Franklin Chow, a postdoctoral researcher (2015-2018) in the lab of our collaborator Dr. Amy Buck at the University of Edinburgh, UK.

H. bakeri life cycle and EV isolation

Mice of the CBA x C57BL/6 F1 (CBF1) line were infected with 400 L3 *H. bakeri* larvae by gavage (force-feeding, through a tube leading down the throat to the stomach). Adult nematodes were collected from the small intestine 14 days post infection (dpi). These nematodes were washed and maintained in serum-free media as previously reported (Johnston et al. 2015). EVs were collected from the adult worm's culture media from 24-92 hours post-harvest from the mouse (the first 24 hours collected media was excluded to reduce host contamination). Eggs were removed by centrifugation at 400 g and the supernatant was then filtered through 0.22 µm syringe filter (Millipore) followed by ultracentrifugation at 100,000 g for 2 hrs in polyallomer tubes at 4 °C in an SW40 rotor (Beckman Coulter). Pelleted material was washed two times in filtered PBS at 100,000 g for 2 hrs and re-suspended in PBS. The pelleted *H. bakeri* EVs, were quantified with Qubit Protein Assay Kit (Thermo Fisher), on a Qubit 3.0.

Mouse cells uptake assays

Intestinal epithelial cells (MODE-K, from now on referred as IECs) and bone marrow-derived macrophages (BMDM) were maintained as reported previously (Vidal et al. 1993). Uptake experiments were done with 2.5 µg EVs or 25 µg HES (EVs represent <10% of HES) per 50,000 cells for 4 and 24 hrs at 37 °C in a 5% CO_2 incubator. *H. bakeri* EV-untreated cells served as controls for the two incubation times. Cells were washed with PBS buffer before RNA extraction with a miRNAeasy mini kit (Qiagen), according to manufacturer's instructions. Three biological replicates were generated per condition. The RNA integrity number was assessed with the Agilent RNA 6000 Pico Kit on an Agilent 2100 Bioanalyzer. As a control for endogenous uptake 2.5 µg of IECs exosomes were applied to BMDM.

Table 6. Mouse cell libraries used to sequence small RNAs. Libraries marked with an asterisk (24 hrs libraries) were additionally subjected to RNA-Seq (see Chapter 4).

#	Cell type	Treatment	Incubation time hrs	#	Cell type	Treatment	Incubation time hrs
1	BMDM	No treatment	4	13*	BMDM	No treatment	24
2	BMDM	No treatment	4	14*	BMDM	No treatment	24
3	BMDM	No treatment	4	15*	BMDM	No treatment	24

4	BMDM	2.5µg *H. bakeri* EVs	4	16	BMDM	2.5µg *H. bakeri* EVs	24
5	BMDM	2.5µg *H. bakeri* EVs	4	17	BMDM	2.5µg *H. bakeri* EVs	24
6	BMDM	2.5µg *H. bakeri* EVs	4	18	BMDM	2.5µg *H. bakeri* EVs	24
7	BMDM	25µg total HES	4	19*	BMDM	25µg total HES	24
8	BMDM	25µg total HES	4	20*	BMDM	25µg total HES	24
9	BMDM	25µg total HES	4	21*	BMDM	25µg total HES	24
10	BMDM	2.5µg IECs exosomes	4	22	BMDM	2.5µg IECs exosomes	24
11	BMDM	2.5µg IECs exosomes	4	23	BMDM	2.5µg IECs exosomes	24
12	BMDM	22.5µg IECs exosomes	4	24	BMDM	2.5µg IECs exosomes	24

#	Cell type	Treatment	Incubation time hrs	#	Cell type	Treatment	Incubation time hrs
25	IECs	No treatment	4	34	IECs	No treatment	24
26	IECs	No treatment	4	35*	IECs	No treatment	24
27	IECs	No treatment	4	36*	IECs	No treatment	24
28	IECs	2.5µg *H. bakeri* EVs	4	37*	IECs	2.5µg *H. bakeri* EVs	24
29	IECs	2.5µg *H. bakeri* EVs	4	38*	IECs	2.5µg *H. bakeri* EVs	24
30	IECs	2.5µg *H. bakeri* EVs	4	39*	IECs	2.5µg *H. bakeri* EVs	24
31	IECs	25µg total HES	4	40*	IECs	25µg total HES	24
32	IECs	25µg total HES	4	41*	IECs	25µg total HES	24
33	IECs	25µg total HES	4	42*	IECs	25µg total HES	24

Small RNA library preparation and sequencing

Total RNA samples were treated with RNA 5' polyphosphatase (Epicenter) according to manufacturer's instructions, before library preparation. Libraries for sRNA sequencing were constructed using CleanTag sRNA library preparation kit following manufacturer's instructions. For all samples 1:2 dilutions of both adapters were used with 18 amplification cycles (TriLink Biotechnologies). Libraries of 140-170 bp length were size-selected and sequenced on an Illumina HiSeq 2500 in high-output mode with v4 chemistry and 50 bp single end reads, by Edinburgh Genomics at the University of Edinburgh (Edinburgh, UK). This insert size was chosen to focus on the small interfering guides of exWAGO, the only secreted Argonaute protein detected within *Heligmosomoides* EVs to date (Chow et al. 2019).

Genome-guided assembly

We used ShortStack 3.8.5 (Axtell 2013) to define and quantify the sRNA producing regions (clusters) of a mixed reference genome including *M. musculus* and *H. bakeri*. I included all libraries described in **Table 6**, with the exception of libraries M_HES_4_3 and M_neg_24_1 as these didn't showed the characteristic 22U miRNA peak. We also included adult and pure EV libraries in their polyphosphatase-treated and untreated versions in our genome-assembly to be able to compare cluster expression in mouse cells to these libraries. I used the following ShortStack parameters bowtie_m: all, dicermax: 24 dicermin: 20, foldsize: 300, mincov: 1, mismatches: 1, mmap: u, pad: 1 and ranmax: 50000. These parameters favor the definition of short clusters relative to those defined in (Chow et al. 2019).

Differential expression analysis

To perform differential expression analyzes, we used the ShortStack output file Counts.txt. In this file rows represent clusters and columns correspond to libraries, and each cell harbors the number of times a sequence occurs in a given library. We discarded unmapped reads found in Counts.txt before testing for differential expression.

Differential expression analyzes were done using the edgeR package (McCarthy, Chen, and Smyth 2012). Lowly expressed features (individual sequences, *de novo* assembled contigs or genome-guided assembled clusters) were filtered; only those that had at least one count per million in at least two libraries were kept. EV-treated or HES-treated IECs or BMDM libraries were compared with untreated control libraries, regardless of the incubation time (4 or 24 hrs). We performed two separate analyses for IECs and BMDM. To find differentially expressed features, a generalized linear model (GLM) likelihood ratio test was used, always fixing the common dispersion to 1.626, which was estimated for unassembled individual sequences. This allowed a fairer comparison between the three levels of assembly. False discovery rate (FDR) was calculated and features that mapped to the nematode, had an FDR < 0.1 and a positive log fold-change were considered up-regulated (Up) *Heligmosomoides* sRNAs.

Defining sRNA classes by length and first nucleotide

The first nucleotide and length of each sequence mapping to the genome-guided clusters was calculated using custom R scripts and the Rsamtools 2.2.3 package (http://bioconductor.org/packages/Rsamtools). Reads between 21-24 nucleotides and beginning with a Guanine were classified as “22G”. Reads between 21-24 nucleotides and beginning with a Thymine were classified as “22U”. These criteria were set by observing the properties of pure EV and IECs libraries (**Figure 15**).

Extracting individual sequences from differentially expressed *H. bakeri* clusters

To get the read sequences from clusters we generated a list of all upregulated *H. bakeri* clusters in any contrast. With this list we extracted all reads that were assigned to these clusters according to ShortStack merged_alignments.bam BAM file with Rsamtools, considering only libraries that should have *H. bakeri* signal (EV or HES treated).

Genomic origin for *H. bakeri* sRNAs

We used a hierarchical annotation of the *H. bakeri* genome, where every nucleotide is assigned to a single genomic region category such as rRNA, tRNA, coding exon, intron, repetitive sequence, etc. As a simplification, only the annotation associated to the middle nucleotide of each read was considered. To find overlaps between mapping reads and annotations, these two sources of information were loaded in R using the genomicRanges package (Lawrence et al. 2013). Overlaps were found with the findOverlaps function with select argument “all”, along with a condition to stop if any of the reads overlapped more than one feature.

Expression comparison of *H. bakeri* detected clusters in mouse

To generate the expression comparison plots, we calculated average counts per million (cpm) for all clusters (mouse and nematode) before differential expression (and low expression filter) with edgeR cpm function (M. D. Robinson, McCarthy, and Smyth 2009). Then, we compared the expression of all *H. bakeri* clusters between treatments with a Shiny app (C. Winston et al. 2020) that we developed, by choosing pairs of columns to display from the *H. bakeri* cpm matrix.

Genome features visualization

Cluster and annotation coordinates were loaded to Integrative Genomics Viewer (IGV) ver. 2.5.2. (J. T. Robinson et al. 2011)

Results and discussion

We know that *H. bakeri* vesicles can be internalized by mouse cells (Buck et al. 2014)(Coakley et al. 2017), but we didn’t know if the sRNA is released from EVs inside the host cells. On the other hand, we also don’t know if HES sRNAs may also produce a detectable signal inside mouse cells. The main objective of the current chapter and the following experiment was to detect and quantify internalized *H. bakeri* sRNAs (Hb-sRNAs) in mouse cells. In Chapter 3 we predict targets in the host transcriptome for those nematode sequences detected in mouse cells in the current chapter. In Chapter 4 we test for a regulatory effect of these Hb-sRNAs on host transcripts.

To analyze mouse cells treated with nematode secretions using sRNA sequencing data we chose a genome-guided sRNA assembly approach that is similar (but not exactly the same) to that described in Chapter 1, along with a differential expression analysis of identified sRNA producing regions (termed clusters throughout this chapter). It is worth mentioning that the results presented in this and following chapters are independent to those generated for Chapter 1, which were published in (Bermúdez-Barrientos et al. 2020).

Our experimental design

The experiments were performed by the group of our collaborator Dr. Amy Buck. Parasite treatments were *H. bakeri* vesicles (EV), total secretion (HES) or no treatment (see Methods for further details). Two incubation times were used: 4 and 24 hours, as we don't know the dynamics of EV internalization. Two different cell types were used: mouse intestinal epithelial cells (IECs) using MODE-K cell line (Vidal et al. 1993) and mouse bone marrow-derived macrophages (BMDM). IECs were shown previously to be able to internalize *Heligmosomoides* EVs (Buck et al. 2014). BMDM were included as our collaborators showed that these cells take up EVs more efficiently than IECs and *Heligmosomoides* EVs suppress macrophage alternative activation (AAM) in BMDM (Coakley et al. 2017). As a control to account for endogenous vesicle treatment, BMDM were incubated with IECs exosomes. After incubation, cells were washed with a buffer solution to remove superficially associated but not internalized vesicles (see Methods). A total of 42 libraries were generated and subjected to sRNA sequencing (**Table 6**).

Our clusters

When constructing our clusters, we included *H. bakeri* adult and EV libraries in their mono-P and poly-P versions with the aim of making our mouse cells clusters comparable with these nematode libraries. A drawback of our cluster search is that our clusters won't be directly comparable with those described previously by our group (Chow et al. 2019). However, our previous clusters were too long, which can result in combining sRNA producing regions that may overlap distinct genomic features. We chose parameters that favor shorter clusters and by doing so, we aimed to minimize the chances of mixing clusters that overlap with different genomic features.

Our cluster definition process resulted in 4.4 M clusters, 1.6 M (37.4%) of these belong to mouse and 2.7 M (62.6%) belong to *H. bakeri*.

The longest *H. bakeri* cluster is Cluster_2777500 which is 7 kb long, and it overlaps a DNA transposon hAT Tip100 (forward strand) that appears to have a LINE BovB retrotransposon (reverse strand) interrupting it. Its most abundant read is a 22G, which suggests that these mobile elements may be silenced by the WAGO RNAi pathway (**Figure 10**).

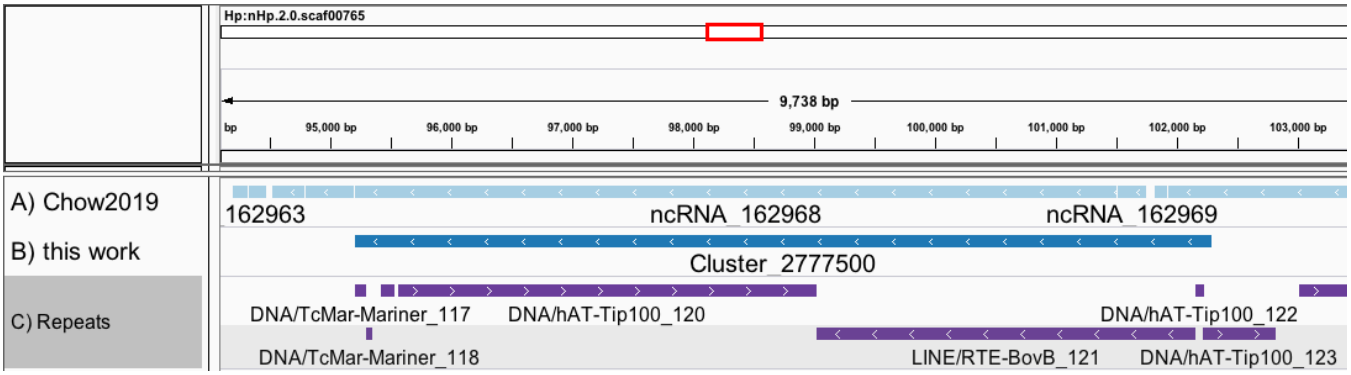


Figure 10. Cluster_2777500 is the longest *Heligmosomoides* cluster. A) Chow 2019 clusters, B) clusters constructed in this work, C) repeat masker annotation. Right pointing arrows denote forward chain features, left pointing arrows denote reverse chain features.

The longest cluster in mouse is Cluster_1602271 which is found in the mitochondrial genome. This cluster is 3.1 kb long and overlaps with a portion of a mitochondrial rRNA, NADH dehydrogenases 1 and 2 and three tRNAs. The reason that this cluster overlaps six genomic elements could be that the mitochondrial genome is highly compact.

Regarding expression levels, the highest expressed nematode cluster is Cluster_4387540, this cluster is 293 nt long and overlaps with the gene HPOL_0002297601 and a ncRNA annotated by Rfam as CeN72 (**Figure 11**). These ncRNAs have been found in *C. elegans* and closely related genomes. Noteworthy, 7 out of 12 (58%) *H. bakeri* CeN72 elements are located in this scaffold. These ncRNAs display a characteristic modification 2,2,7-trimethylguanosine (TMG), this modification was used to experimentally validate CeN72 RNAs by combining TMG targeting antibodies and RNA arrays (Jia et al. 2007). Cluster_4387540 displays higher expression in mono-P vs poly-P libraries and its most abundant read is 30 nt long.

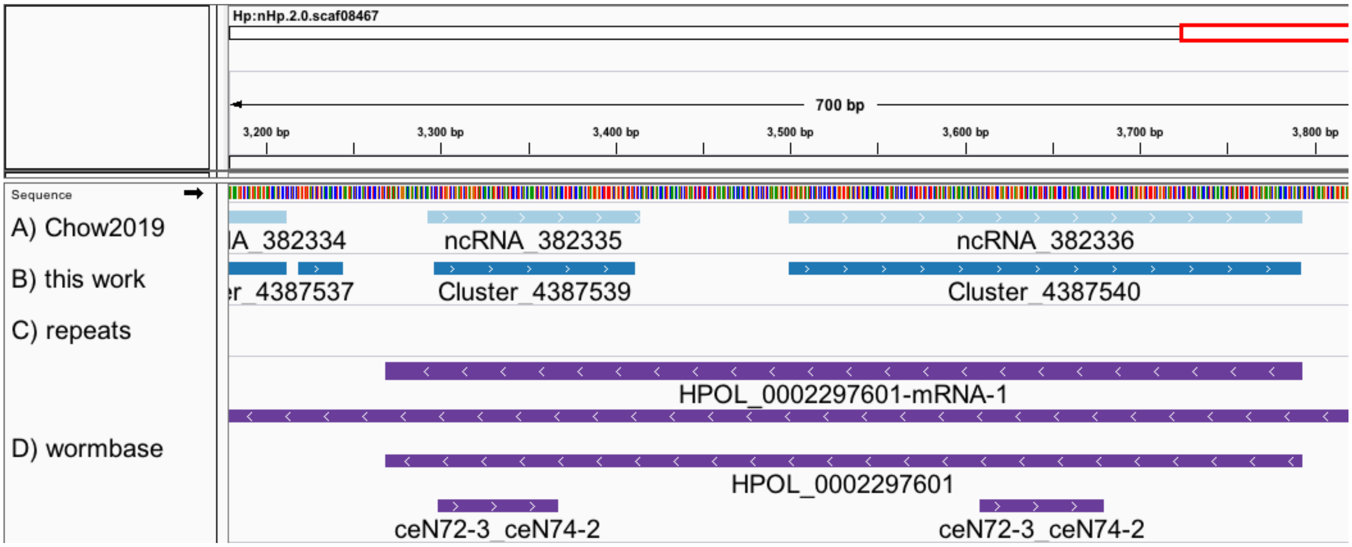


Figure 11. Cluster_4387540 region and annotations. A) Chow 2019 clusters, B) clusters constructed in this work, C) repeat masker annotation, D) Wormbase parasite annotations. Right pointing arrows denote forward chain features, left pointing arrows denote reverse chain features.

The most highly expressed cluster in mouse is Cluster_761871. This 47 nt cluster overlaps a 5.8S rRNA gene. This is an expected result as rRNA comprises a major portion of a cell's total RNA.

With our clusters defined, we proceeded to a differential expression analysis that would help us detect *H. bakeri* sRNA clusters under a statistical framework.

Differential expression analysis aids detection of parasite sRNAs in mouse cells

With our experimental design, multiple comparisons or contrasts are possible. These contrasts result in different numbers of detected upregulated *H. bakeri* clusters, that are summarized in (**Table 7**). We expect the nematode signal in cells to increase upon treatment with the secretion product, for this reason we will only focus in *H. bakeri* up-regulated (Up) clusters and ignore non-differentially expressed clusters and downregulated clusters, which we will refer collectively as non-Up clusters. We detected a total of 306 distinct *Heligmosomoides* clusters in mouse cells with all different comparisons.

Table 7. Summary of differential expression analysis results

Contrast	up Hb clusters	Up Mm+Hb clusters	% Hb up clusters	up DE Hb counts	non-DE Hb clusters
IECs_EV_vs_no_treatment	301	310	97.1	70,222	4,158
IECs_EV_vs_no_treatment_4hrs	201	204	98.5	63,431	4,254
IECs_EV_vs_no_treatment_24hrs	101	104	97.1	51,686	4,357
IECs_HES_vs_no_treatment	143	149	96.0	57,903	4,316
IECs_HES_vs_no_treatment_4hrs	84	86	97.7	56,535	4,376
IECs_HES_vs_no_treatment_24hrs	58	62	93.5	47,225	4,399
BMDM_EV_vs_no_treatment	29	50	58.0	29,597	3,001
BMDM_EV_vs_no_treatment_4hrs	6	10	60.0	27,884	3,030
BMDM_EV_vs_no_treatment_24hrs	12	25	48.0	28,686	3,019
BMDM_HES_vs_no_treatment	22	73	30.1	3,657	3,005
BMDM_HES_vs_no_treatment_4hrs	6	11	54.5	2,085	3,030
BMDM_HES_vs_no_treatment_24hrs	5	81	6.2	27,683	3,015
BMDM_exo_vs_no_treatment	1	6	16.7	26,492	3,035

Using IECs we are able to detect a greater number of *Heligmosomoides* clusters than with BMDM. The number of detected clusters ranges between 58 and 301 for IECs while the number of clusters for BMDM ranges only between 5 and 29. The endogenous uptake control for BMDM incubated with IECs exosomes detects only 1 *H. bakeri* cluster. This result contrasts to EV uptake reported by Coakley and collaborators (Coakley et al. 2017), where our collaborators reported higher EV uptake by BMDM and RAW macrophages than by IECs intestinal epithelial cells. Some scenarios that may explain this are: (1) macrophages internalize more *H. bakeri* EVs and degrade internalized Hb-sRNAs more effectively than

IECs. (2) Hb-sRNAs action also results in their degradation and then we are not able to see them in BMDM, in this scenario we would expect a stronger repression effect in BMDM relative to IECs.

The EV treatment (301, 201, 101 clusters) gives a greater number of clusters in IECs than the HES treatment (143, 84, 58). The number of detected clusters even doubles with EV relative to incubation with HES. This also means that fewer micrograms of EV (2.5 μg) are required to detect Hb-sRNAs in IECs relative to HES (25 μg). The BMDM EV treatment allowed the detection of slightly more *H. bakeri* clusters (29, 6, 12) than the HES treatment (22, 6, 5).

We considered the percentage of *H. bakeri* clusters found among up-regulated clusters (Hb + Mm) as a way of measuring the strength of nematode signal. In this comparison a higher *H. bakeri* percentage of the total of up-regulated clusters relates to more parasite signal for a particular treatment or contrast. We were able to detect a higher percentage of *H. bakeri* clusters in up-regulated clusters for IECs (ranging between 93.5% to 98.5% of total up-reg clusters host + parasite) than for BMDM ranging from 6.2% (HES treatment at 24 hrs) up to 60% (EV treatment at 4 hrs) (**Table 7**). This suggests that our nematode treatments have a greater effect on mouse sRNA populations of BMDM relative to IECs, which results in more host upregulated sRNAs in BMDM.

We also noticed that there's a higher percentage of nematode clusters in IECs EV treatment (97.1, 98.5%, 97.1%) than in HES treatment (96%, 97.7%, 93.5%). This result also holds for BMDM, this line also shows a higher percentage of parasitic clusters in EV (58%, 60%, 48%) than in HES treatment (30%, 54%, 6.2%). This suggests that HES treatment results in more host sRNAs being upregulated than EV treatment.

We detected more sRNA clusters in IECs at 4 hr relative to 24 hr, being 201 vs 101 clusters for early and late incubation respectively for EV and 84 vs 58 clusters for HES treatment. This shows that Hb-sRNAs signal decreases, but does not completely disappears after 24 hours in the intestinal epithelial cells.

We also detected higher percentages of *H. bakeri* clusters contribution to up-regulated clusters in 4 hours relative to 24 hours. In IECs with EV treatment at 4 hours the percentage of parasite up-regulated clusters is 98.5% while being 97.1% at 24 hrs, although this difference is quite small. In HES-treated IECs the percentage of up-regulated clusters is 97% at 4 hrs and 93.5% at 24 hrs. Regarding BMDM, with EV treatment the percentage of nematode up-regulated clusters is 60% at 4 hrs and 48% at 24 hrs. This effect is more dramatic for HES-treated BMDM, in this scenario the percentage of parasite clusters is 54.5% at 4 hrs and this drops to only 6.2% at 24 hrs. This suggests that at 24 hours there are more host sRNA loci responding to nematode secreted components than at 4 hours, this effect was more subtle for IECs than for BMDM and seems to be more dramatic for HES than for EVs.

IECs with EVs provides the highest number of counts for *Heligmosomoides* with 70,222. The lowest number of counts is given by BMDM incubated with HES at 4 hrs with only 2,085 read counts. In general, we detected higher parasite counts in IECs than in BMDM.

To explore if different clusters are detected in different cell types or treatments, we built a Venn diagram with those two contrasts that retrieved the most clusters for each cell type (BMDM with HES, BMDM with

EVs, IECs with HES and IECs with EVs), these results are shown in **Figure 12**. A total of 18 clusters (5.8% of all DE-detected *H. bakeri* clusters) were found in both cell types and in both treatments (EVs or total HES). A total of 158 clusters (51.6%) were exclusively found in IECs with EVs and 115 (37.5%) clusters were found also in IECs with HES. Macrophages contributed with few exclusive clusters. Only one cluster, Cluster_4048248, was exclusively found in BMDM treated with EVs and only three clusters, Cluster_4387540, Cluster_4200193 and Cluster_3414970, were exclusively identified in BMDM with HES treatment. This also reveals that most of the Hb-clusters detected in BMDM were also detected in IECs.

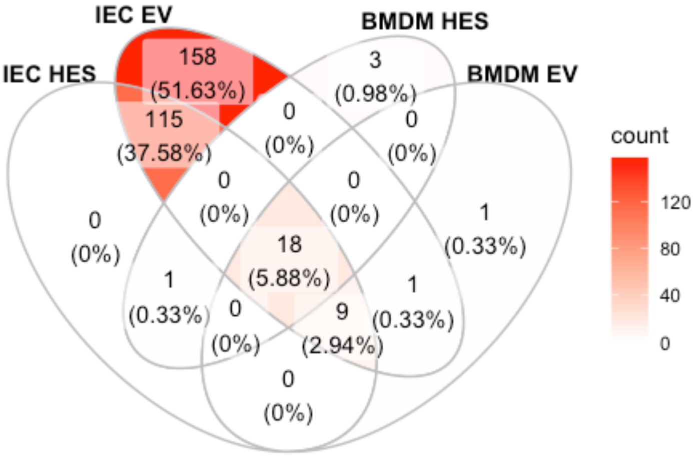


Figure 12. Venn diagram of the *H. bakeri* clusters detected in each differential expression comparison. Diagrams sections are colored on a red gradient according to the number of clusters within.

We will focus on some of the most interesting comparisons, the differential expression analysis (DEA) corresponding to IECs and BMDM cells each treated with EV or HES irrespective of time. All 13 DEA figures are available in the supplementary material (**Supplementary Figure 4**).

What we see in the following plots is a representation of all sRNA clusters for *M. musculus* and *H. bakeri* with each dot representing a sRNA cluster (**Figure 13**). We have four cluster categories, first we have clusters belonging to mouse or nematode, that are further divided into up-regulated clusters due secretion treatment (Up) and those clusters that are not up-regulated (non-Up). This last category includes both clusters that do not show any evidence of differential expression, and clusters that are down-regulated. We decided to ignore down-regulated clusters as our main focus is to detect *Heligmosomoides* clusters and these should be among up-regulated features.

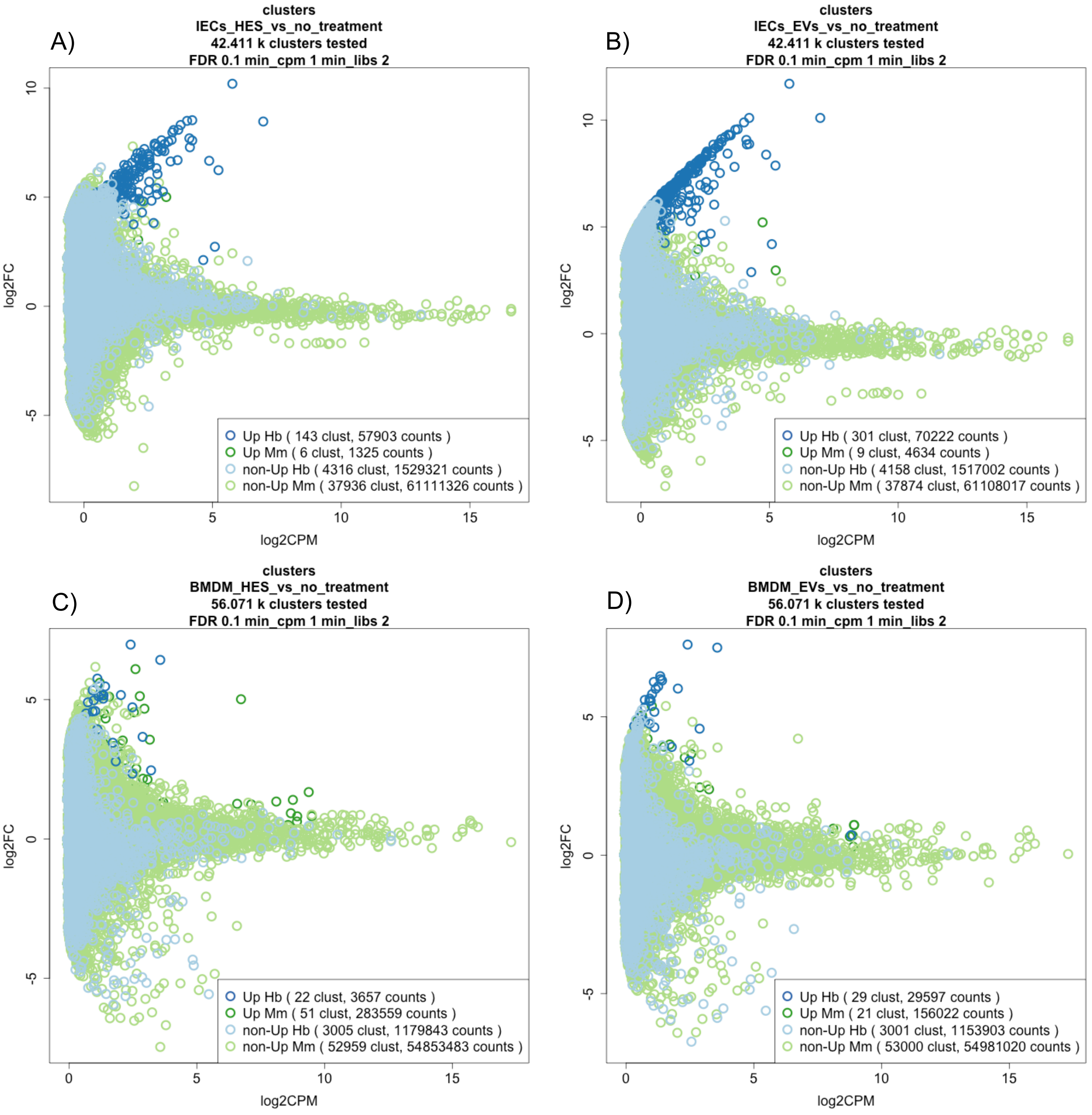


Figure 13. Mean abundance plots with organism mapping information for mouse cells incubated with *Heligmosomoides* EVs or HES. The X axis represents clusters expression in log2 CPM scale. The Y axis represents fold change ratio between compared treatments in log2 scale. Each dot represents a cluster for either *H. bakeri* (Hb, blue dots) or *M. musculus* (Mm, green dots). Upregulated clusters are colored in a dark tone and non-upregulated clusters are colored in light tone. In the legend, clust are the number of distinct clusters, counts are the number of reads mapping to these sequences. A) IECs with HES, B) IECs with EVs, C) BMDM with HES and D) BMDM with EVs. In all cases nematode-treated libraries were compared to control untreated libraries. Only clusters with one CPM in at least two libraries were included in the analysis. We used a FDR of 0.1.

The parasite signal is found in the top portion of these figures. In IECs the nematode signal has a shape similar to that of the tip of a pickaxe, formed by dark blue circles, that extends beyond the bulk of light-colored clusters that do not have enough evidence for differential expression (DE). This peak is wider for HES (146 clusters) and narrower for EV (306 clusters) in intestinal epithelial cells, this difference could be due to higher clusters expression variance among HES-treated replicates than that of EV-treated ones. In BMDM, this nematode signal tip is not as clear for HES as it is for EV treatment, in both cases I found less signal than that of IECs, that involves a difference of an order of magnitude of fewer detected clusters.

Some of the DE clusters display higher levels of expression and lower fold change increase, these are the result of the FDR making a curved cut along the tested clusters. It is possible that some of these clusters belong to the host as these are also expressed without the nematode secretion treatment, as evidenced by less dramatic fold changes.

An evident result from these figures is the great number clusters that were assigned to *H. bakeri* and lack evidence for differential expression in these analyses (80% to 97.1%, **Figure 13**, light blue clusters). The proportion of clusters that do not show differential expression is 1,440 out of 1,586 clusters (90.7%) in IECs with HES treatment, 1,277 out of 1,583 (80.6%) clusters in IECs with EVs, 658 out of 677 (97.1%) in BMDM incubated with HES and 654 out of 677 (96.6%) for BMDM with EV treatment. The properties of these upregulated and non-upregulated clusters will be discussed later.

A subtle but noticeable effect is found when comparing IECs and BMDM figures, there seem to be differences in the relative positions of the distribution of nematode clusters with respect to those of the host. The nematode clusters distribution seems to be skewed towards positive fold changes in IECs. Another plausible scenario, and not mutually exclusive to the previous one, may be that more sRNA loci would become activated upon external stimuli (herein EV or HES) in BMDM. An interesting comparison for this discussion is that of IECs exosomes being applied to BMDM as a negative control for *H. bakeri* signal. In this MA plot we may see that the upper part of the plot corresponding to mouse clusters is wider than that of the lower portion, which could be explained by mouse sRNA responding to endogenous exosomes being produced by another cell type. We also know that in this exosome treatment setting no nematode secretion was applied, so all apparent parasitic genome mapping reads are false positives. It is worth mentioning that, in this case, the fold change distribution of nematode clusters signal is near to symmetrical (**Figure 14**), which suggests that the asymmetries of nematode signal distributions in BMDM with either EVs or HES holds some true parasitic clusters.

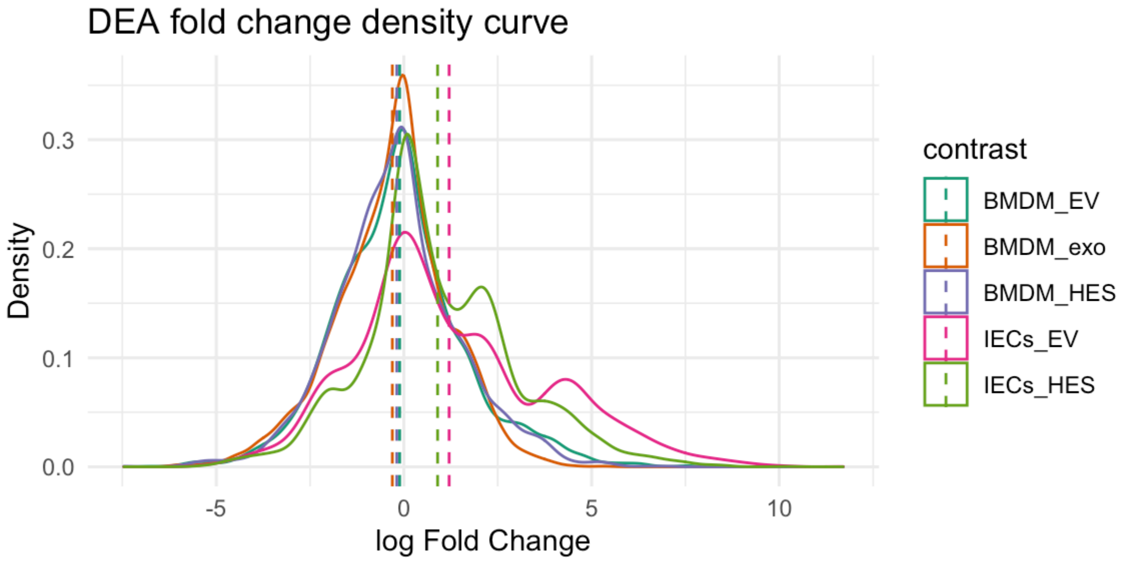


Figure 14. Fold change distributions for differential expression analyses with nematode secretions (EV or HES) or intestinal epithelial cells exosomes as a negative control. Dashed lines represent means for each contrast.

In the following sections of this chapter, we will compare the properties of the two different sets of clusters for *Heligmosomoides*, those that are upregulated (Up) and those that are not upregulated upon parasitic stimuli (non-Up).

Validation of differentially expressed parasitic sRNAs

The sRNA profiles obtained from pure *Heligmosomoides* and *M. musculus* samples differ greatly (**Figure 15**). RdRP products (22G reads) are the dominant class of *Heligmosomoides* libraries (92% of EV libraries are 22G considering a 21-24 nt range), while miRNAs (22U reads) are the dominant class in the *Mus musculus* sRNA profile (79.1% are 22U, considering a 21-24 nt range). A 22G-rich profile is more likely to be found in *H. bakeri* libraries than in those produced by *M. musculus*. We can take advantage of these differences and use them as a fingerprint (sRNA class fingerprint) to distinguish nematode from host sequences in samples containing material from both organisms.

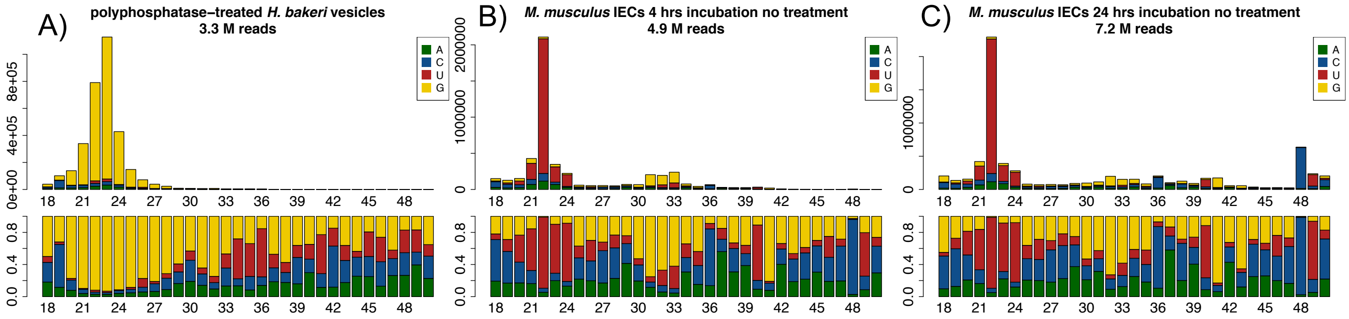


Figure 15. Small RNA class fingerprint obtained from pure *H. bakeri* vesicles or *M. musculus* IECs. First nucleotide preference for A) polyphosphatase-treated *H. bakeri* vesicles, and *M. musculus* IECs libraries at B) 4 hours and C) 24 hrs incubation without treatment. The y-axis represents the count average across replicate libraries.

We took our two sets of upregulated (Up) and non-upregulated (non-Up) *Heligmosomoides* clusters defined according to our differential expression analysis. We then compared the length distribution and the first nucleotide (sRNA class fingerprint) from all the reads that mapped to these two sets of clusters (**Figure 16**).

This information is represented in **Figure 16**. In a multi-panel figure organized as a matrix with rows and columns (**Figure 16**). The sRNA class fingerprint for both IECs with HES (70.9% of 22Gs) and EV (75.2% of 22Gs) treatments looks very similar to the profile of pure *H. bakeri* EVs (75% of 22Gs) (**Figure 16 A and B**). On the other hand, non-upregulated sRNA profile in IECs is enriched with 22U sRNAs (40.6% for HES and 40.5% for EV) as well as shorter 18 and 19 nt reads. Upon considering mapping location, the 22U sRNAs in non-Up are mostly ultra-conserved miRNAs such as let-7a-5p, miR-100-5p and miR-9a-5p.

Reads mapping to up-regulated *H. bakeri* clusters in BMDM with HES treatment somehow resembles the sRNA profile of pure EV libraries (43.5% vs 75%), but it is a bit noisier, there are considerable amounts of 20 nt reads that begin with an adenosine (**Figure 16D**). Regarding BMDM with EVs, the up-regulated signal is dominated by short (18-20 nt) sequences that begin with a cytosine, however a smaller distribution of 22G is evident (12.9%), this suggests that we were able to detect true parasitic signal even in these conditions (**Figure 16C**). This 22G signal is not evident in macrophages treated with IECs exosomes (**Figure 16E**), where we only see reads 18-20 nt and these are enriched for cytosine, this also suggests that the short cytosine reads from BMDM with EVs is host signal. Non-upregulated nematode reads in both BMDM contrasts also contain 22U signal from the same three miRNAs previously mentioned (let-7a-5p, miR-100-5p and miR-9a-5p), and an additional U rich peak at 19 nt reads.

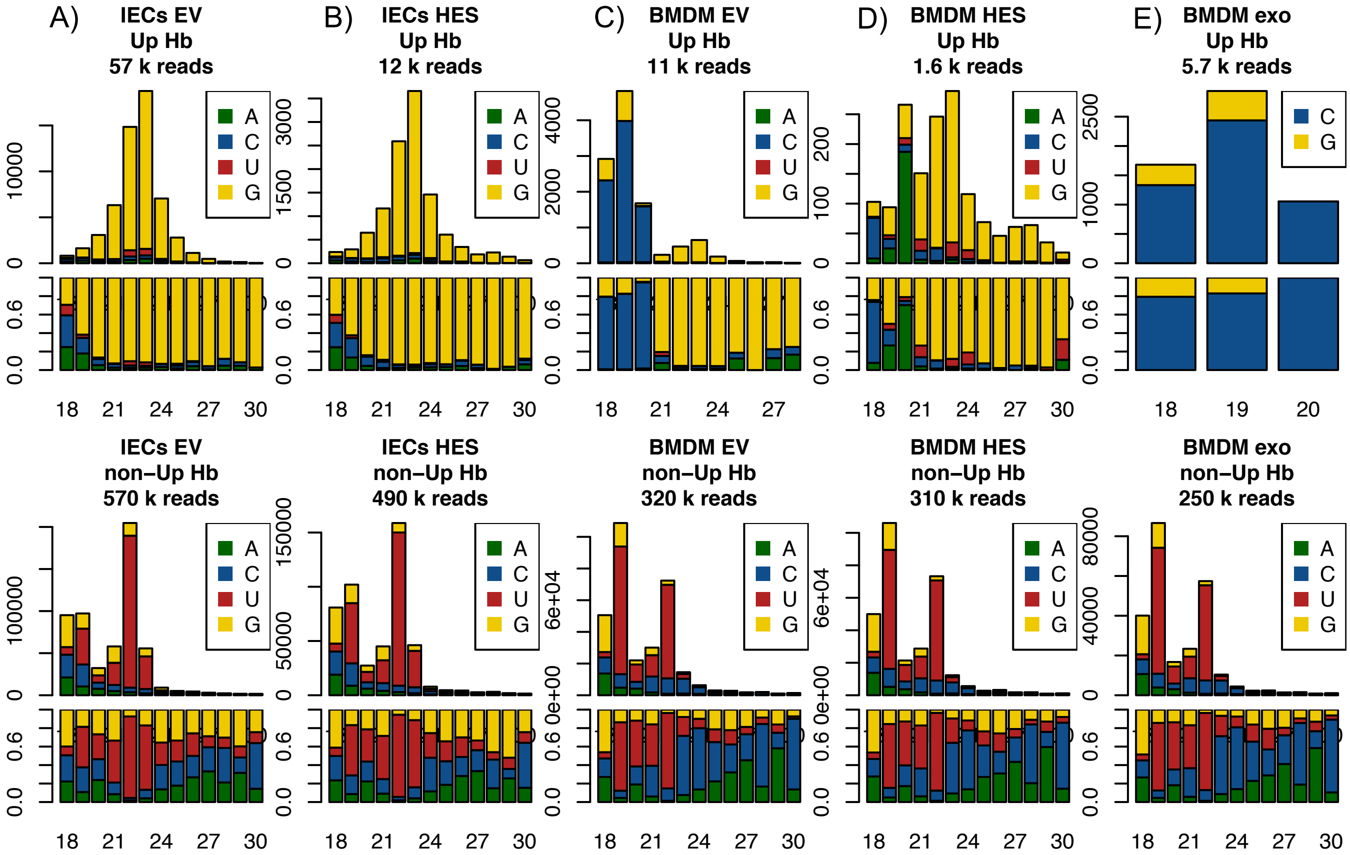


Figure 16. Small RNA class fingerprints obtained from reads mapping to upregulated (top half) or non-upregulated (bottom half) *H. bakeri* clusters detected in *Mus musculus* cells. Each column corresponds to a different comparison, A) IECs + EV, B) IECs + HES, C) BMDM + EV, D) BMDM + HES, E) BMDM + IEC exosomes. Small RNAs from upregulated *H. bakeri* clusters are found in the first row, sRNAs from non-upregulated *H. bakeri* clusters are found in the second row. Each column represents a different comparison, the first column represents IECs with HES-treatment, the second represents IECs with EVs, the third corresponds to BMDM incubated with HES, finally the fourth column represents EV-treated BMDM.

We find short *H. bakeri* reads signal in both upreg and non-upreg sets from BMDM. Could this suggest that *Heligmosomoides* sRNAs are being degraded in macrophages? These shorter reads are not evident in IECs. To assess this question it would be valuable to check if the shorter sequences detected are trimmed versions of longer reads found in the BMDM libraries and absent from parasite untreated libraries.

We conclude that the sRNA class fingerprint is consistent with up-regulated clusters reads being produced by *H. bakeri*, giving us confidence that our detected nematode reads are truly parasitic.

Genomic origin of *H. bakeri* sRNAs detected in mouse cells

We were curious about the nature of the *H. bakeri* loci that produced the nematode sRNAs signal detected in mouse cells. We extracted the annotation for each read assigned to the *Heligmosomoides* genome and did this for upregulated and non-upregulated nematode clusters as two different sets, in a very similar way that we did for the first nucleotide and read length analysis. We then compared these two profiles to the genomic origin profile of the pure EV libraries.

For simplicity, I will only show the genomic origin profiles for four DEA contrasts, the full genomic profiles for all comparisons are available in the supplementary material (**Supplementary Figure 5**).

We used the same genomic feature summary scheme as that used in (Chow et al. 2019) with rRNAs, miRNAs, tRNAs, yRNAs, mRNA sense, mRNA antisense, introns as well as some additional categories that I will explain a little more. Retroelements represent retrotransposons such as LINES, transposons represent DNA mobile elements, simple repeats include micro-satellites and low complexity sequences and novel repeats are repetitive sequences that are unique to the *H. bakeri* genome. Finally, other ncRNA includes piRNAs, snRNAs, snoRNAs, as well as other non-coding RNAs contained in the Rfam database.

The genomic profiles of the *H. bakeri* clusters discovered in treated IECs somehow resemble the profile of pure EV libraries, and this similarity is present in both EV and HES treatments (**Figure 17A&B**). Both profiles consist of 31% of reads originating from intergenic regions, 18% from novel repeats, 12% from transposons, 10%-11% from introns, between 10-12% from retroelements, 9% antisense to mRNA and only 3% of miRNA contribution. HES treatment is different in that we detected 6.9% yRNA signal which is not evident in EVs and miRNAs decrease from 3% to 0.8% in HES. Buck *et al.* identified a strong yRNA signal in the supernatant and low yRNA signal in EVs (Buck et al. 2014), our results are consistent with this previous report since supernatant is basically a EV-depleted HES.

The average phosphatase treated *H. bakeri* EV libraries are composed of 22.2% retroelements, 15% novel repeats, 10.8% transposons, 10.6% introns, only 2.7% of miRNA, and 26.84% reads mapping to intergenic regions. Retroelements are sub-represented in up-regulated sequences in treated IECs, here they comprise only 10-12% while in pure EV they can reach up to 22.2%, being the most abundant genomic category. The novel repeats category increases from 15% in EVs to 18% inside treated IECs, transposons also increase from 10.8% to 12%. Finally, the percentage of intergenic reads increases from 26.8% to 31% inside IECs.

miRNAs are the major contributor for non-up-regulated nematode-mapping reads in IECs comprising 47-51%. Introns are the second most abundant category for this set with 12-16%, and rRNA comes third with 12%.

In the genomic origin profile of BMDM incubated with HES (**Figure 17C**), we find yRNAs as the most abundant category with 20%, in second place we find rRNA with 16%, in third come novel repeats with 11%, then introns with 10% and so on. The presence of miRNAs is not even evident in the figure. The major differences between the profiles of up-regulated reads with HES and EV profiles is the over-representation of yRNAs and rRNAs in HES, and the under-representation of the three most abundant EV categories: retroelements (22% in EVs, <10% in BMDM with HES), novel repeats (15% in EVs, 11.2% in BMDM) and transposons (10.8% in EVs, 4.9% HES in BMDM). As above, this result agrees with yRNAs detection by Buck *et al.* (Buck et al. 2014), therefore we detect yRNAs signal in HES treatments regardless of the cell type (IECs or BMDM).

The genomic profile corresponding to BMDM with EV treatment shows an aberrant behavior in comparison with the rest of the profiles with parasite treatment shown here (**Figure 17D**), it consists of 84-92% antisense to mRNA, ~3% novel repeats, 1-2% retroelements. The BMDM with EV profile is similar to that of our control treatment for endogenous exome uptake, BMDM with IECs exosomes (**Figure 17E**).This profile is composed by 100% reads that are antisense to mRNAs although there was only one up-regulated cluster missasigned to *H. bakeri* in this comparison. The resemblance between BMDM with EVs and BMDM with exo suggests that there is low parasite signal in the EV treatment for macrophages and that the sole up-regulated cluster missasigned in the exosome contrast is also present in BMDM with EVs. We would expect that the signal present in BMDM with EVs originating from novel repeats, retroelements and transposons corresponds to a minority of expected 22G sRNA class fingerprint (**Figure 16**). This could be demonstrated if we build figures similar to those of **Figure 16** but containing genomic origin for reads instead of the first nucleotide.

The genomic profile of non-upregulated *H. bakeri* clusters in BMDM cells consist of 30-33% introns, 29-31% miRNAs and 20% rRNAs as the three major categories. The profiles are almost identical for HES, EVs or IECs exosome-treated macrophages, we even detect signal from *H. bakeri* novel repeats in the exosome contrast despite being a negative control. This again raises concerns of merely using mapping information to assign reads to an organism. The finding of miRNAs and rRNAs could be expected given their high degree of evolutionary conservation, however no direct explanation is apparent for the intronic regions being this abundant between nematode and mouse macrophages. Further explorations would be needed in order to explain the abundance of intron mapping reads.

It is worth mentioning that the genomic profiles of non-Up *H. bakeri* sRNAs may still contain true parasitic signal, since the methods we use are designed to limit the rate of false discoveries, at the expense of increasing the false negative results. This is a compromise that needs to be taken in any statistical analysis.

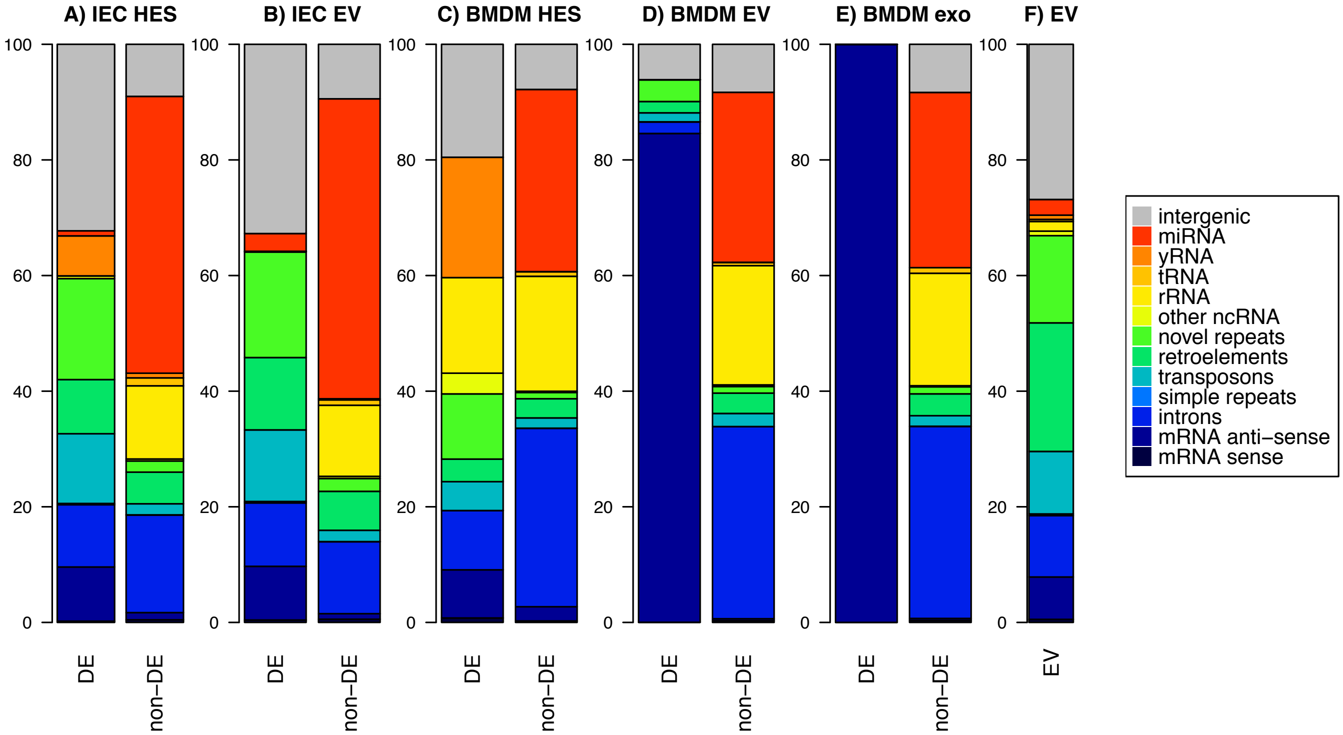


Figure 17. Genomic origin profile for *H. bakeri* clusters detected in *Mus musculus* cells. The results for four differential expression comparisons are shown A), B), C), D) and E), each one represented by a pair of columns for upregulated (left) or non-upregulated (right) reads mapping to *H. bakeri* clusters. F) The genomic profile for polyphosphatase-treated EV libraries for adult and EVs are shown as a reference.

With these results, we have shown that the reads associated to up-regulated nematode clusters have properties that are similar to those sRNAs produced by *H. bakeri*. To consider 22G sRNAs seems to be a good indicator of *H. bakeri* signal (**Figure 16**). With this in mind, we revisited our differential expression analysis shown in **Figure 13**. This time we show only *H. bakeri* clusters and highlight 22Gs over those belonging to 22U and other clusters (**Figure 18**). One of the first things that caught our attention was that some of the Up-detected clusters (yellow circles) appear to be longer than non-Up ones (yellow triangles), longer clusters may be translated into increased statistical power, and thus detectability. Secondly, in all cases, the top half of the plots (positive logFC) has more 22G clusters than the lower half of the plots (negative logFC). This effect also involves those clusters that do not pass the differential expression threshold (FDR <= 0.1, represented by yellow triangles). From these observations, we conclude that our 306 *H. bakeri* Up clusters and their associated 15,133 sequences are high-quality but conservative results (there's evidence for *H. bakeri* sRNA clusters that didn't pass the stastistical threshold) of the total parasite signal inside mouse cells.

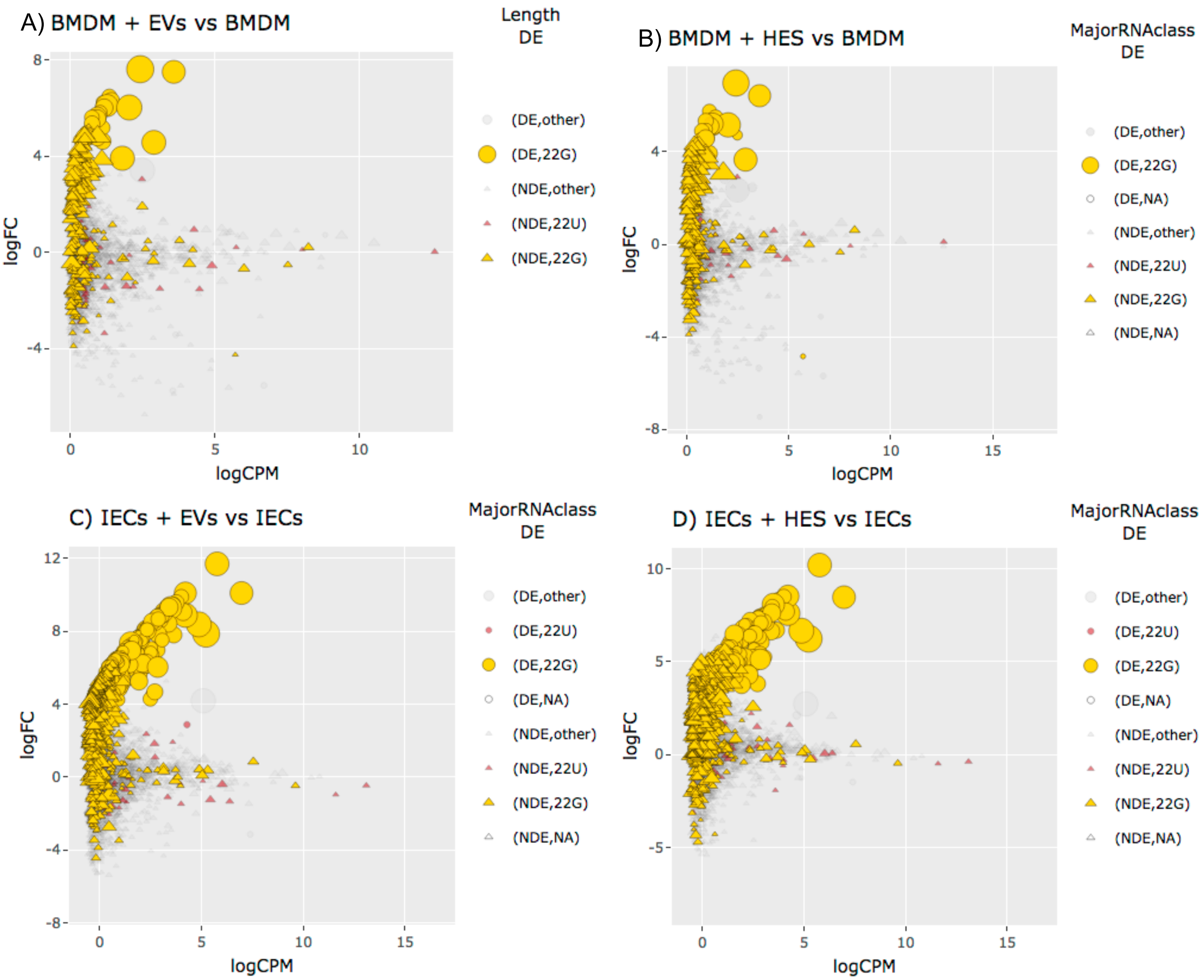


Figure 18. Mean abundance plots for *H. bakeri* clusters. 22G clusters are shown in yellow, 22U clusters are shown in red, other clusters are shown in grey. Those clusters with an FDR ≤ 0.1 are shown as circles, cluster that do not pass this threshold are shown as triangles. In all cases the size of the point reflects the size of the cluster.

Comparison of *H. bakeri* detected clusters in mouse cells

Our experimental design allows us to ask multiple questions that are of interest such as: do we detect the same clusters in intestinal epithelial cells (IECs) and in macrophages (BMDM)? Do we detect different sRNAs in EV vs HES treatment? Do we see changes in sRNAs that are influenced by early (4 hrs) and late (24 hrs) incubation times?

The next figures compare *Heligmosomoides* cluster expression between different libraries. We classified clusters as 22G (yellow), 22U (red) or other (grey) depending on its most abundant read. We also use the size of dots to represent the length of sRNA producing clusters, therefore bigger dots represent longer clusters. In the following plots when there is good correlation between the expression of clusters between two libraries, clusters are found near the plot diagonal. Taking this to the extreme, if we compare the expression levels of a library with itself, all clusters will be aligned into a single diagonal line starting from the bottom left to the top right part of the plot. Thus, deviations from the diagonal tell us about clusters that are expressed to higher or lower levels in a library relative to another.

And what happens when a cluster is expressed in a library but expressed to very low or null levels in another one? In these cases, differences in cluster expression are represented by deviations from the diagonal, in the most extreme case of null expression, this cluster will appear along the axis where it is expressed and with low or null CPMs for the other library. This is precisely what happens when comparing BMDM treated with IECs exosomes vs BMDM treated with *H. bakeri* EVs. 22G clusters (yellow) are found near the X axis, which corresponds to the EV-treated libraries. As we have previously shown, mouse cells produce few 22G (**Figure 19**), which results in low or null expression for these clusters in exosome treated libraries, this is in accordance with IECs exosomes being a negative control treatment for nematode signal.

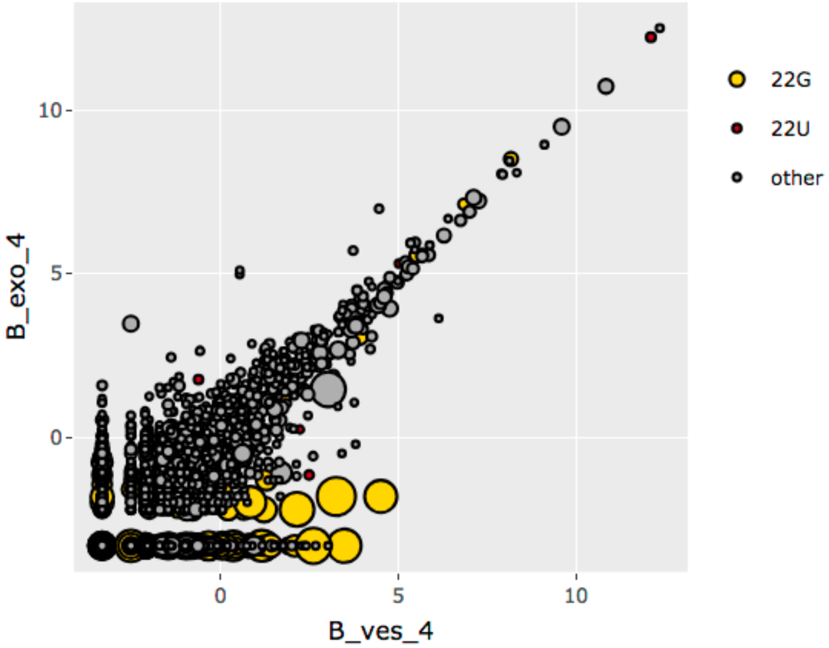


Figure 19. *Heligmosomoides* cluster expression comparison for EV-treated BMDM libraries (X axis) and IECs exosome-treated BMDM libraries (Y axis). Each dot represents a sRNA producing cluster that are color coded according to the major read classification 22G are yellow, 22U are red and other are grey. The size of the dot reflects the size of the cluster. X and Y axes show log2 CPM clusters expression levels.

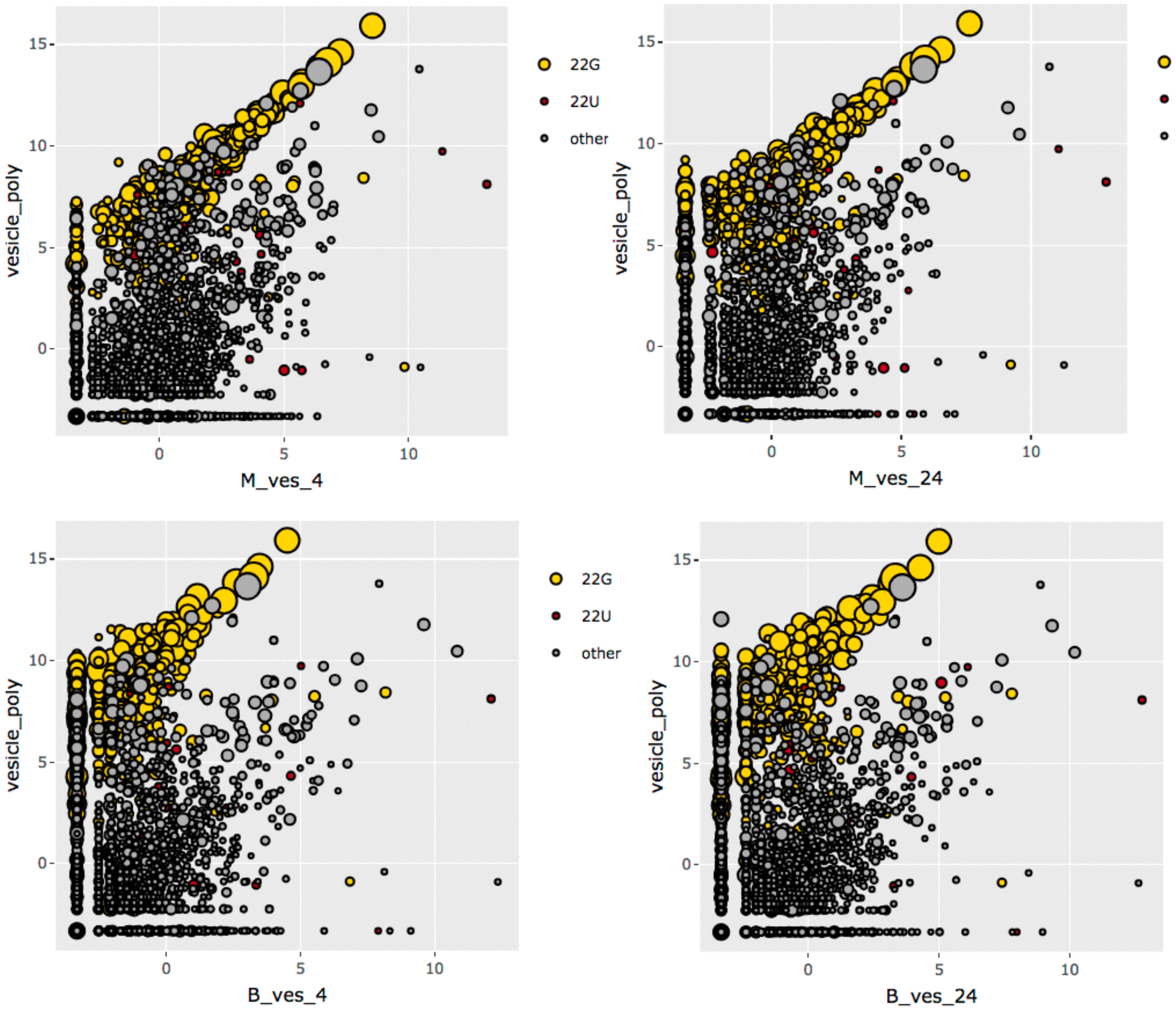


Figure 20. *Heligmosomoides* cluster expression comparison for libraries from EV-treated mouse cells (X axis) and EV pure libraries (Y axis). Each dot represents a sRNA producing cluster. Clusters are color coded according to the major read classification 22G are yellow, 22U are red and other are grey. The size of the dot reflects the size of the cluster. The first row of figures includes IEC libraries, the second row includes BMDM libraries. The first and second columns of figures include 4 and 24 hr respectively. X and Y axes show log2 CPM clusters expression levels.

Our decision to include pure EV libraries in the process of cluster construction allows us to compare those clusters detected in mouse cells with their expression in pure EV libraries.

When comparing detected clusters in EV-treated cells vs pure EVs, we observe two main clusters populations separated by a gap. Those clusters categorized as 22G are located on the top left part of the plot, smaller clusters categorized as other are located on the middle part of the plot (**Figure 20**). This segregation pattern may be explained by two possibilities, the first one is that the 22G population is the true *H. bakeri* signal and those clusters located in the middle of the plot are the host noise (wrongly mapping to the nematode genome). The second possibility is that the middle cluster population may be true nematode clusters signal that comprises mainly 5' monophosphate sRNAs that are present in a smaller proportion in poly-P EVs. In support of this second possibility, we observe a similar two populations distribution pattern when comparing EV poly-P and EV mono-P libraries (**Figure 21**).

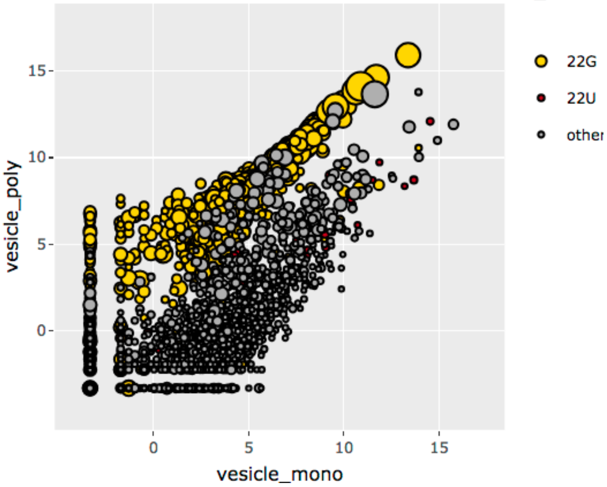


Figure 21. *Heligmosomoides* clusters expression comparison for phosphatase-untreated EV libraries (X axis) and phosphatase-treated EV libraries (Y axis). Each dot represents a sRNA-producing cluster. Clusters are color coded according to the major read classification 22G are yellow, 22U are red and other are grey. The size of the dot reflects the size of the cluster. X and Y axes show log2 CPM clusters expression levels.

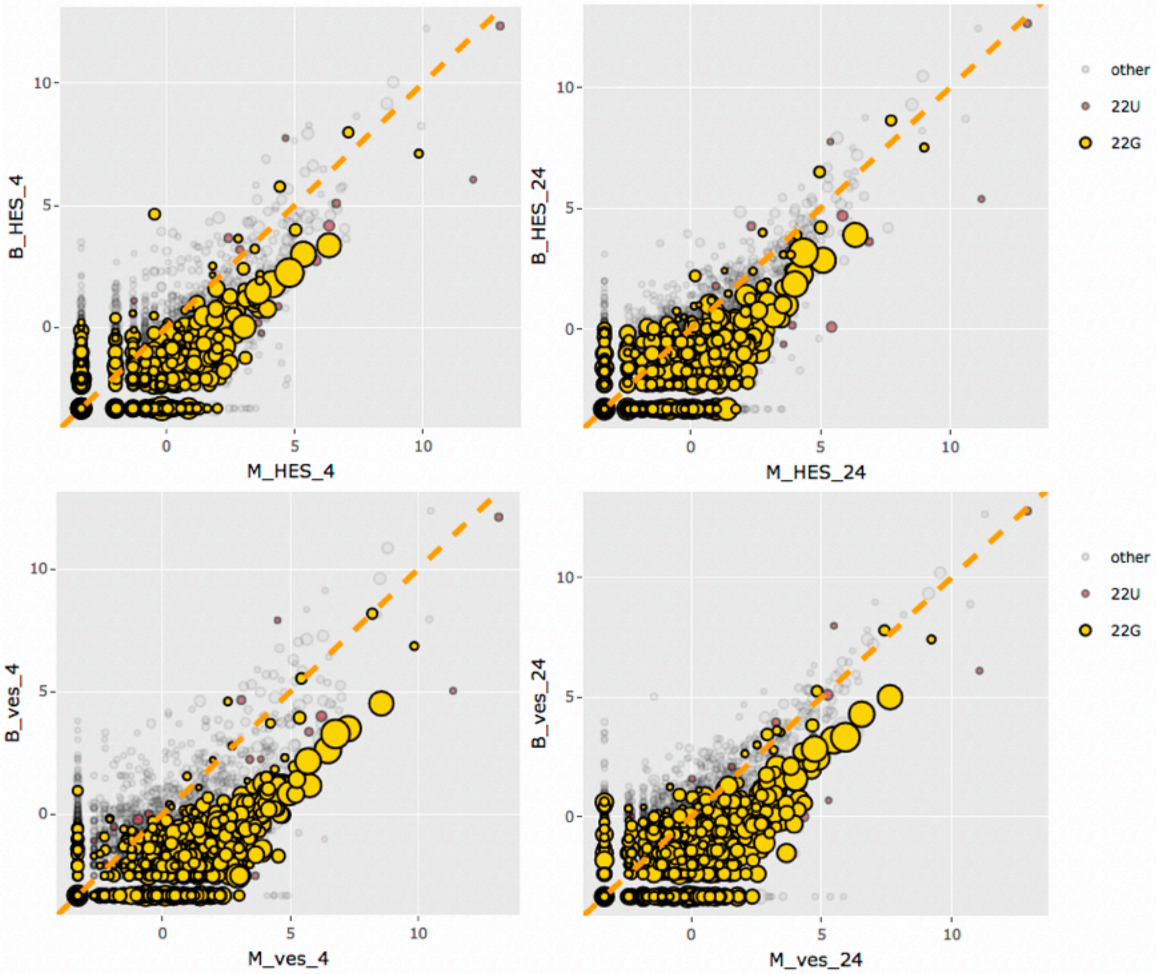


Figure 22. *Heligmosomoides* cluster expression comparison for IECs (X axis) and BMDM libraries (Y axis). Each dot represents a sRNA producing cluster that are color coded according to the major read classification 22G are yellow, 22U are red and other are grey. The size of the dot reflects the size of the cluster. The first row of figures includes HES libraries, the second row includes EV libraries. The first and second columns of figures include 4 and 24 hr respectively. X and Y axes show log2 CPM clusters expression levels.

Regarding IECs (X axis) vs BMDM (Y axis) comparisons (**Figure 22**), we noticed that 22G clusters (yellow) are biased to the lower side of the plot, below the diagonal line. This means that upon treatment there is higher expression of 22G clusters in IECs relative to BMDM. This is in accordance with a stronger signal detected in IECs vs BMDM (**Figure 12** and **Figure 13**). Nevertheless, the pattern suggests that there is simply a difference in the level of detection inside BMDM and not a major difference regarding which clusters can be internalized by both types of cells.

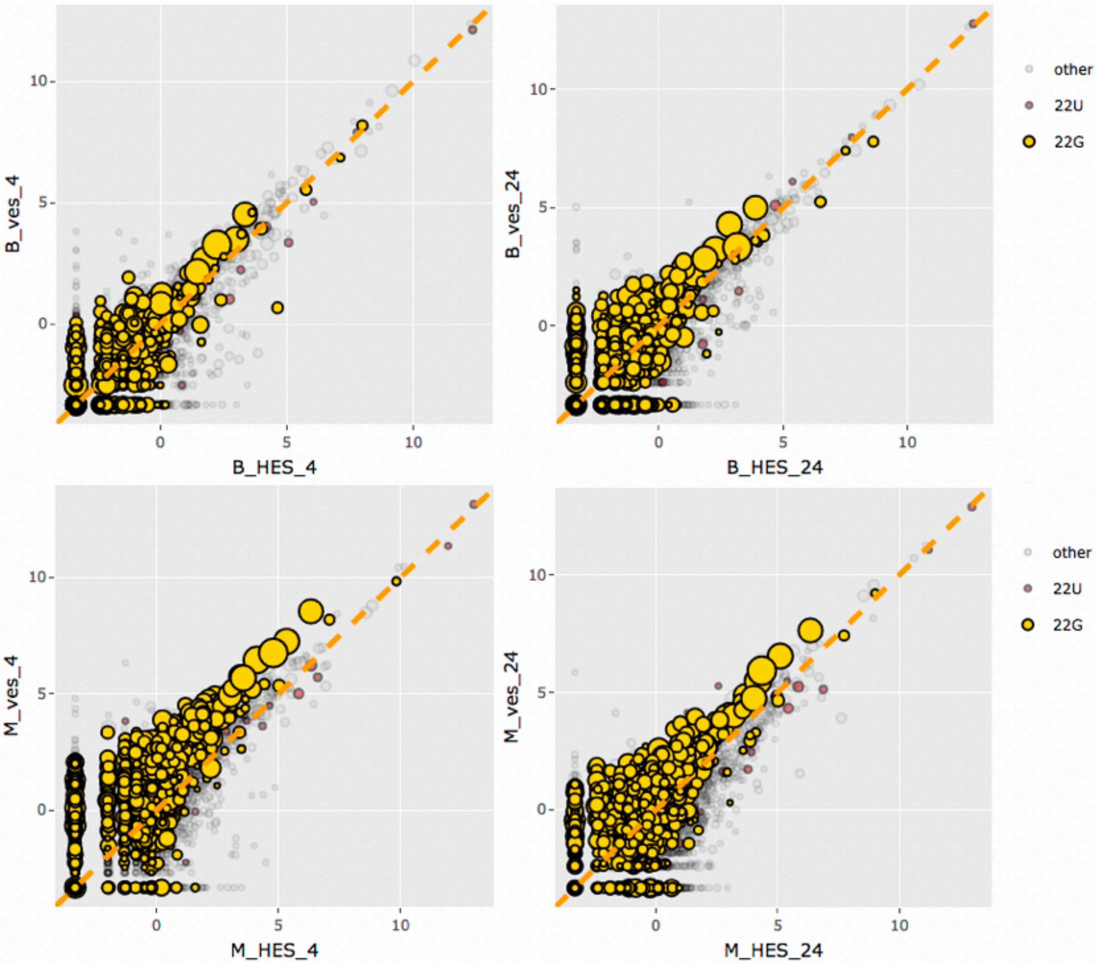


Figure 23. *Heligmosomoides* cluster expression comparison for HES (X axis) and EV libraries (Y axis). Each dot represents a sRNA producing cluster that are color coded according to the major read classification 22G are yellow, 22U are red and other are grey. The size of the dot reflects the size of the cluster. The first row of figures includes BMDM libraries, the second row includes IECs libraries. The first and second columns of figures include 4 and 24 hr respectively. X and Y axes show log2 CPM clusters expression levels.

When comparing EV vs HES treatments we noticed that there's a bias for 22G clusters to have higher counts in EV relative to HES, since 22G clusters tend to be located in the upper side of the plot (**Figure 23**). This could be due EV providing a higher concentration of 22G *H. bakeri* reads than HES in IECs (**Figure 16**). This bias is also present in macrophages but is more subtle.

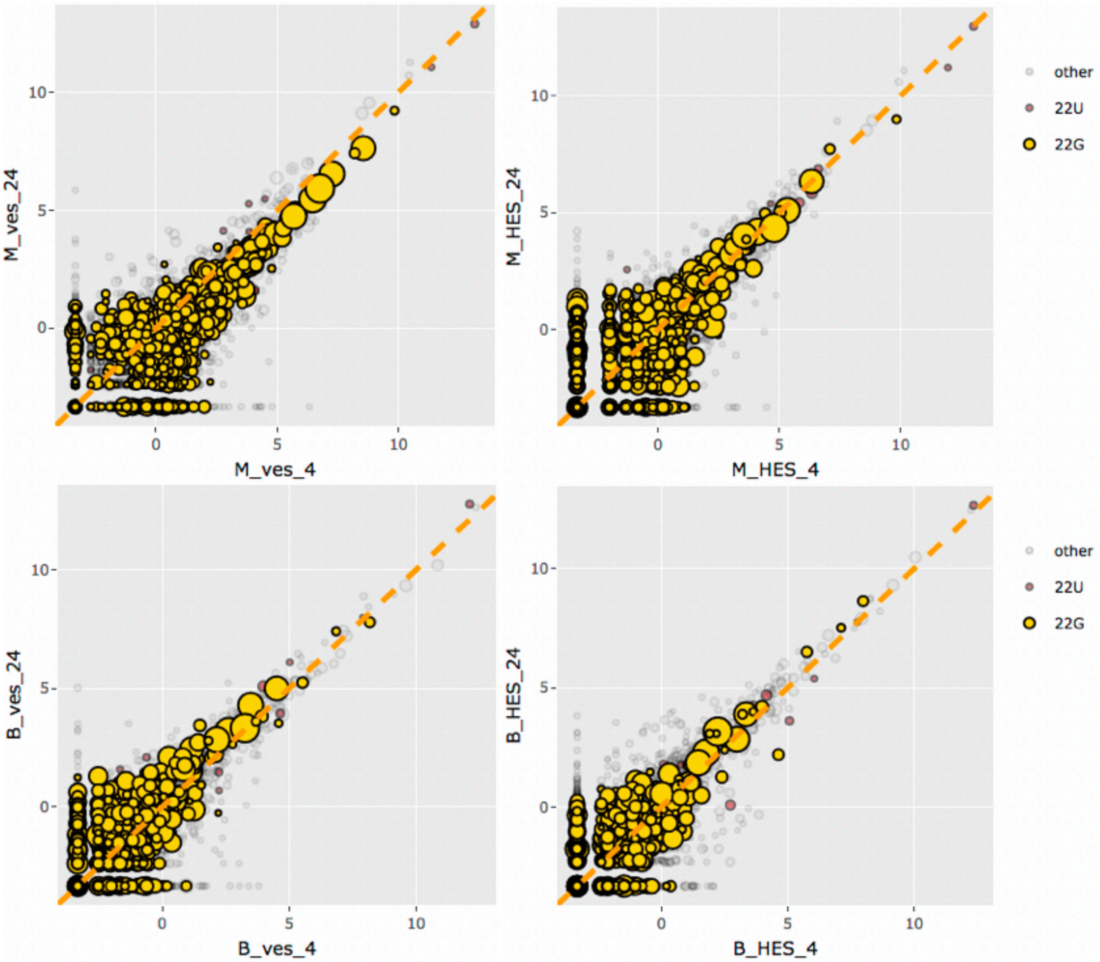


Figure 24. *Heligmosomoides* cluster expression comparison for 4 hr (X axis) and 24 hr libraries (Y axis). Each dot represents a sRNA producing cluster that are color coded according to the major read classification 22G are yellow, 22U are red and other are grey. The size of the dot reflects the size of the cluster. The first row of figures includes IECs libraries, the second row includes BMDM libraries. The first and second columns of figures include EV and HES respectively. X and Y axes show log2 CPM clusters expression levels.

When comparing the two incubation times (4 and 24 hours) we can see that 22Gs tend to be found below the diagonal for IECs with EV treatment (**Figure 24**), although the separation from the diagonal is smaller than for any of the previous comparisons. This indicates that there's stronger 22Gs signal at an early incubation time with EV. On the other hand, this effect is not evident for IECs that received the HES treatment. Regarding macrophages, BMDM with EV 22G signal seems to be slightly biased to 24 hrs rather than 4 hrs, this effect is not evident with HES treatment, although it may be worth doing a statistical test rather than a visual inspection. This apparent higher 22G signal at 24 hrs sounds counterintuitive, I would expect Hb-sRNAs to dimish, or remain at similar levels. This observation requires validation with experimental approaches such as qPCR quantification of selected Hb-sRNAs.

Conclusions

I was able to detect *Heligmosomoides* signal in mouse cells using a genome-guided assembly combined with a differential expression analysis. This strategy yielded 306 *H. bakeri* clusters that are associated to 15,133 different sRNA sequences. Our detected *H. bakeri* sRNAs have clear nematode properties such as *H. bakeri* genome-mapping, upregulation upon *H. bakeri* treatment (**Figure 13**), 22G enrichment (**Figure 16**) and similar genome origin to that of pure nematode EV (**Figure 17**). The detection of high-confidence *H. bakeri* sRNAs is a key step in determining if these foreign sRNAs exert a function in host cells.

We detected more nematode signal in intestinal epithelial cells (IECs) than in macrophages (BMDM). But in general, it is not apparent that specific sRNAs are only able to enter IECs. If sequencing depth of BMDM were to be increased, we would expect the same sRNAs to be detected in both cell types. EV-treatment contributed a greater number of *H. bakeri* clusters than HES, although HES treatment recovers yRNA mapping reads than are absent from EVs (**Figure 17**).

Our results demonstrate that just relying on mapping information for sRNA-Seq data may result in wrong organism assignment (**Figure 13**). Ultra-conserved miRNAs such as let-7, miR-9a and miR-100 are a main source for *Heligmosomoides* assigned reads that do not show up-regulation, and are thus likely to contain a large amount of reads that in reality originate from the mouse genome.

Perspectives

- Investigate if there’s evidence for expression of *H. bakeri* CeN72 elements in mouse cells, as the top expressed *H. bakeri* cluster ovelapped one of these elements. We know that this cluster is enriched in mono-P vs poly-P libraries, we could additionally check if CeN72 elements are differentially expressed in EVs relative to adult.
- Explore if there’s evidence for degradation of Hb-sRNAs in macrophages. A first step for this would be to investigate if smaller sequences found in macrophages are subsets of longer Hb-sRNAs detected.
- Explore the cellular fate of EVs in macrophages and IECs with pure EVs and with total HES. Are EVs (and their Hb-sRNAs) degraded by macrophages?
- As I detected more 22Gs signal in macrophages at 24 hrs than 4 hrs, it may be woth to investigate if there is any evidence for Hb-sRNAs amplification in mouse cells.
- Investigate which are the most highly expressed yRNA and miRNA clusters detected in mouse cells and predict targets for these separately from 22G targets. The rationale behind this is that we know that 22Gs tend to be loaded into WAGO proteins, but yRNAs and miRNAs tend to be enriched in polyphosphatase-untreated libraries (mono-P), suggesting that they may be loaded into other Argonaute proteins (miRNAs tend to be loaded into ALG-1 in *C. elegans*).
- Explore which host miRNAs are differentially expressed due to HES or EV treatments in IECs or macrophages. In this work I focused in detecting Hb-sRNAs, thus host miRNAs are an unexplored field for this dataset. Changes in endogenous miRNAs could also be a valuable control when assesing the effect of sRNAs in host transcripts.

Chapter 3: Predicted host targets for *Heligmosomoides bakeri* sRNAs

In Chapter 1 we developed strategies to determine producing organism for sRNA-Seq data, in Chapter 2 we applied these strategies to detect nematode sequences inside mouse cells. In this chapter we predict host targets for those *H. bakeri* sRNAs found in Chapter 2.

Methods

Non-redundant Hb-sRNAs

We produced a set of non-redundant Hb-sRNAs using the generalized Levenshtein edit distance (Levenshtein 1966) implemented in the adist R function. This function calculates the edit distance between two sequences, and we used an edit distance threshold of 2 to consider to sRNAs as redundant. We also set a higher priority for sequences that display higher expression to those that are lowly expressed, this is done by using higher expressed sequences as references and removing those sequences that are similar according to an edit distance of <= 2.

To extract all Hb-sRNAs that matched seq5_x385 we used grep -B 1 command line instruction, these Hb-sRNAs were aligned with muscle v3.8.31 and visualized with jalview 2.11.1.0 (Waterhouse et al. 2009) (see Results and discussion).

End-to-end Hb-sRNAs target predictions in host genes

End-to-end target predictions for the 8,501 non-redundant Hb-sRNAs were done using TargetFinder version 19.02.2015 (Fahlgren and Carrington, n.d.) and a penalty score threshold of 8. TargetFinder works with a penalization scoring scheme. Perfect complementarity between a sRNA and its target would result in a score of 0. Mismatches and single nucleotide gaps are given a penalty of +1, G:U interactions are given a penalty of +0.5, penalty scores are doubled between positions 2-13 in order to give a higher priority for the seed region. We used the whole mouse transcriptome for target predictions, regardless of the transcript biotype. Transcriptome data was downloaded with the biomaRt 2.42.1 R package (Durinck et al. 2005). TargetFinder predictions in table format were loaded into R to perform further explorations discussed in this chapter.

Reward summed score

We transformed TargetFinder penalty scores to a score that could be summed to better reflect a transcript being targeted by multiple Hb-sRNAs. We inverted TargetFinder's penalty score, considering a 0-penalty score as a "reward score" of 6 and a penalty score of 6 as a reward score of 0. For this summed score we didn't consider sites with a penalty score above 6.

Those 306 detected parasite clusters found in Chapter 2 correspond to 16,889 unique *H. bakeri* sRNA sequences in the 18 to 30 nucleotide range (**Figure 25A**). At first glance, this profile is 22G rich which is in accordance with it's *H. bakeri* origin, however, when we consider the expression levels for each sequence (**Figure 25B**), it becomes obvious that there are 18 to 20 nt sequences begining with a cytosine that are quite abundant. These 18-20 C sequences are not found in typical pure EV or adult nematode libraries (Chow et al. 2019), so we decided to narrow the size range before making target predictions. Therefore, we kept 15,133 unique sequences in a 20 to 26 nucleotide size range. This size range encompasses that of sequences known to be loaded onto Argonaute proteins.

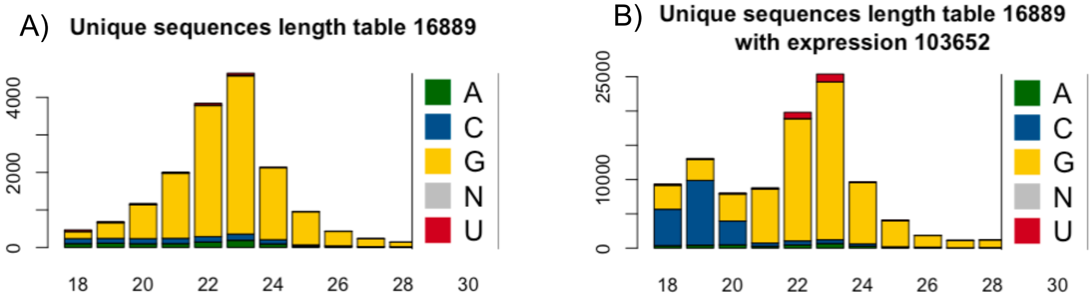


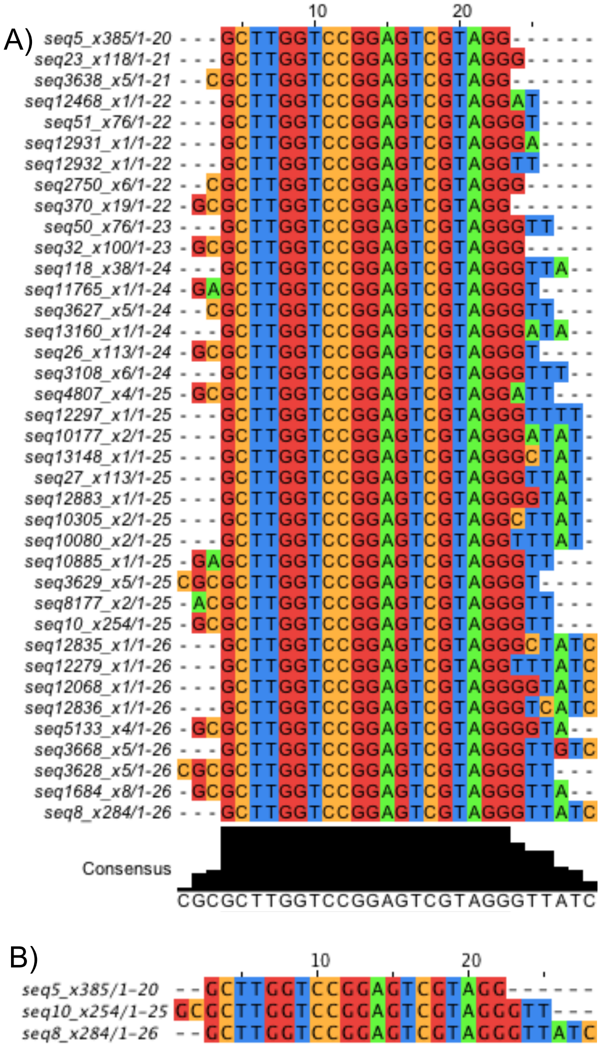
Figure 25. Small RNA first nucleotide profile for A) unique sequences and B) unique sequences with expression for sequences detected in mouse cells.

I decided to name the Hb-sRNAs according to their expression level. In this naming scheme seq1_x3092 is the top expressed individual sequence out of all 15,133 Hb-sRNAs with 3092 reads across all libraries with EVs or HES. Seq2_x1400 is the second most expressed sequence with 1400 reads and so on.

Reducing redundant targets

We found 15,133 sequences in mouse cells treated with *H. bakeri* secretion products (see Chapter 2), however, some of these sequences are essentially identical such as extension variants or sequences that differ by just a few nucleotides.

As an example, there are 38 sequences that share the same 20 nt with seq5_x385 (**Figure 26A)**. Sequence seq23_x118 contains an extra G on its 3' end, while sequence seq3638_x5 contains an extra C at that end. It is very likely that if we predict targets for all these 38 sequences most of them would share the same target transcripts, especially since they share the seed region that contributes most to the score. Many of these redundant sequences might be derived from the same nematode locus, with differences probably arising due to enzymatic activity variation or by editing at the 3' end.



```
A)
seq5_x385/1-20     ---GCTTGGTCCGGAGTCGTAGG------
seq23_x118/1-21    ---GCTTGGTCCGGAGTCGTAGGG-----
seq3638_x5/1-21    --CGCTTGGTCCGGAGTCGTAGG------
seq12468_x1/1-22   ---GCTTGGTCCGGAGTCGTAGGAT----
seq51_x76/1-22     ---GCTTGGTCCGGAGTCGTAGGGT----
seq12931_x1/1-22   ---GCTTGGTCCGGAGTCGTAGGGA----
seq12932_x1/1-22   ---GCTTGGTCCGGAGTCGTAGGTT----
seq2750_x6/1-22    --CGCTTGGTCCGGAGTCGTAGGG-----
seq370_x19/1-22    -GCGCTTGGTCCGGAGTCGTAGG------
seq50_x76/1-23     ---GCTTGGTCCGGAGTCGTAGGGTT---
seq32_x100/1-23    -GCGCTTGGTCCGGAGTCGTAGGG-----
seq118_x38/1-24    ---GCTTGGTCCGGAGTCGTAGGGTTA--
seq11765_x1/1-24   -GAGCTTGGTCCGGAGTCGTAGGGT----
seq3627_x5/1-24    --CGCTTGGTCCGGAGTCGTAGGGTT---
seq13160_x1/1-24   ---GCTTGGTCCGGAGTCGTAGGGATA--
seq26_x113/1-24    -GCGCTTGGTCCGGAGTCGTAGGGT----
seq3108_x6/1-24    ---GCTTGGTCCGGAGTCGTAGGGTTT--
seq4807_x4/1-25    -GCGCTTGGTCCGGAGTCGTAGGATT---
seq12297_x1/1-25   ---GCTTGGTCCGGAGTCGTAGGGTTTT-
seq10177_x2/1-25   ---GCTTGGTCCGGAGTCGTAGGGATAT-
seq13148_x1/1-25   ---GCTTGGTCCGGAGTCGTAGGGCTAT-
seq27_x113/1-25    ---GCTTGGTCCGGAGTCGTAGGGTTAT-
seq12883_x1/1-25   ---GCTTGGTCCGGAGTCGTAGGGGTAT-
seq10305_x2/1-25   ---GCTTGGTCCGGAGTCGTAGGCTTAT-
seq10080_x2/1-25   ---GCTTGGTCCGGAGTCGTAGGTTTAT-
seq10885_x1/1-25   -GAGCTTGGTCCGGAGTCGTAGGGTT---
seq3629_x5/1-25    CGCGCTTGGTCCGGAGTCGTAGGGT----
seq8177_x2/1-25    -ACGCTTGGTCCGGAGTCGTAGGGTT---
seq10_x254/1-25    -GCGCTTGGTCCGGAGTCGTAGGGTT---
seq12835_x1/1-26   ---GCTTGGTCCGGAGTCGTAGGGCTATC
seq12279_x1/1-26   ---GCTTGGTCCGGAGTCGTAGGTTTATC
seq12068_x1/1-26   ---GCTTGGTCCGGAGTCGTAGGGGTATC
seq12836_x1/1-26   ---GCTTGGTCCGGAGTCGTAGGGTCATC
seq5133_x4/1-26    -GCGCTTGGTCCGGAGTCGTAGGGGTA--
seq3668_x5/1-26    ---GCTTGGTCCGGAGTCGTAGGGTTGTC
seq3628_x5/1-26    CGCGCTTGGTCCGGAGTCGTAGGGTT---
seq1684_x8/1-26    -GCGCTTGGTCCGGAGTCGTAGGGTTA--
seq8_x284/1-26     ---GCTTGGTCCGGAGTCGTAGGGTTATC

Consensus          CGCGCTTGGTCCGGAGTCGTAGGGTTATC

B)
seq5_x385/1-20     --GCTTGGTCCGGAGTCGTAGG------
seq10_x254/1-25    GCGCTTGGTCCGGAGTCGTAGGGTT---
seq8_x284/1-26     --GCTTGGTCCGGAGTCGTAGGGTTATC
```

Figure 26. Example of redundant sequences. A) *H. bakeri* sRNAs sharing the same sequence as seq5_x385, B) Output of non-redundant sequences for seq5_x385.

Redundant sequences may bias our tests for target down-regulation (see Chapter 4). Nearly identical Hb-sRNAs binding sites would be counted as independent binding sites in target transcripts, and would result in inflated metrics that aim to consider multiple binding sites (see reward summed score below). We filtered redundant sequences with an R script using the adist function (see Methods). Our script gives priority to higher expressed sequences over lower expressed ones. As an example, of the 37 Hb-sRNAs that share the same sequence as seq5_x385, only two remained: seq10_x254 and seq8_x284 (**Figure 26B**). All results described from this point on were produced using this set of non-redundant sequences. We initially had 15,133 unique Hb-sRNAs and ended up with 8,501 non-redundant sequences (61% of the original sequences and 70.6% of the original expression).

There are alternatives to our edit distance approach, for example Cd-hit-est is a program that groups similar sequences and returns a representative one for each group. We didn't use cd-hit-est as this program keeps the longest sequence as representative of each cluster, and this is not convenient for target predictions as some sequence length may be longer than the size of Argonaute-loaded sRNAs. My approach gives preference to expression rather than sequence length.

A caveat to our strategy for reducing redundant sequences is that our editing distance method does not differentiate between seed regions and other portions of the sRNA. This deviates from known biology given that the seed region is of uttermost importance for the interaction of a sRNA-loaded Argonaute protein and its targets.

It’s worth mentioning that there are differences in expression levels for our Hb-sRNAs, the most highly expressed one has 3,092 read counts, on the other hand, there are 1739 (20.4%) out of the 8,501 non-redundant Hb-sRNAs that have only one read count. This also highlights the impact and success of the assembly strategy, there’s simply no way that such lowly expressed sequences would have been detected as differentially expressed with a non-assembled, individual-sequence approach.

End-to-end Hb-sRNAs target prediction in host

The objective for this chapter was to perform target predictions on the host transcriptome for those Hb-sRNAs detected in mouse cells in Chapter 2. A question that soon comes to mind is which transcripts may be the targets for our Hb-sRNAs? And, how many targets are there? It would be of little use to say that the whole mouse transcriptome may be targeted by Hb-sRNAs, on the other hand, we may not expect that only a couple host genes are being regulated by our set of thousands of Hb-sRNAs. The number of targets may lay in between these extremes with hundreds or perhaps thousands of host targets.

There are clusters that were detected in a particular condition and go undetected in another. There’s evidence that there’s parasite signal in BMDM despite low numbers of Hb Up-regulated clusters (**Figure 18**). Therefore, we predicted targets for all Hb-sRNAs associated to the universe of Hb Up-regulated clusters detected in any of the performed DEA contrasts, independently of whether any cluster was detected only in IECs or BMDM or only with one of the nematode treatments (EVs or HES).

We performed target predictions for only 8,501 non-redundant Hb-sRNAs using TargetFinder on the mouse transcriptome (see Methods). In order to explore the number of mouse targets we described how the penalty score influences the number of Hb-sRNAs that have at least one predicted binding site, the number of host target genes found and the median number of Hb-sRNAs binding sites for targeted transcripts. We described these parameters varying the penalty score from 0 (perfect complementarity) to 8 (a very lax penalty score). For this description we focused on the protein coding transcripts and used the longest transcript variant for each gene.

There’s only one Hb-sRNA that displays perfect sequence complementarity with its target (penalty score 0). So, at a penalty score 0 we have one Hb-sRNA and one targeted gene (

Table 8). Seq5874_x4 binds to the Ssh2 slingshot protein phosphatase 2, a gene implicated in actin cytoskeleton organization according to its GO terms.

With a penalty score of 1 there are 12 Hb-sRNAs that target 11 genes, this implies that one of these genes is targeted by more than one Hb-sRNA. Both seq558_x16 and seq5874_x4 target the previously mentioned Ssh2 phosphatase. Penalty scores of 0, 0.5, 1 and 1.5 yield almost exclusively one to one Hb-sRNA-transcript interactions. At a penalty score of 2, the same sRNAs start to target multiple genes as 154 Hb-sRNAs target 183 mouse genes.

With a penalty score of 4 almost half of the mouse genes (10,393, 47.8%) are targets for 4,086 Hb-sRNAs (48% of the Hb-sRNAs used for target prediction). It is worth mentioning that 4 is the default penalty score used by TargetFinder and this is a recommended threshold according to the authors to predict miRNA targets in plant genomes, where high-complementarity is required to cleave the target RNA molecule. This is also the threshold that has been used by other studies (Srivastava et al. 2014). Relaxing the penalty score to 5.5 results in 93% of the Hb-sRNAs having a target and 98.4% of the mouse genes being targeted.

I was curious to know if all Hb-sRNAs have at least one host target transcript. With our most lax penalty score threshold (penalty score 8) only six out of the 8,501 Hb-sRNAs don’t have any host target.

Another thing we explored was the median number of Hb-sRNA binding sites per targeted transcript in order to describe how commonly are mouse transcripts targeted by multiple Hb-sRNAs. From 0 till 4 penalty scores, the median of Hb-sRNA binding sites per targeted transcript holds at 1. From 4 onwards we observed increasing numbers that resemble a 2-based exponential growth, with 2 sites for a 4.5 penalty core, 4 sites for 5 penalty score, 9 sites for 5.5 and so on. This trend holds till penalty score 7.5 where we observe 101 median sites per targeted transcript, were we would expect 128. This deviation is even wider for the penalty score 8, as we would expect 256 binding sites and we observe just 141. We expect this reduction simply reflects a limit imposed by the length of the mouse transcripts. Further work is needed to explore if this observation has any biological implication such as Hb-sRNA binding site saturation.

We had a deeper look at the number of Hb-sRNA binding sites per targeted transcript for the recommended penalty score threshold 4. With a penalty score of 4, half of the targeted transcripts have 1 site, the third quartile of targeted transcripts corresponds to 2 sites and the maximum number of sites with this threshold is 19 sites for the gene D10Wsu102e “DNA segment expressed” by Wayne State University. Gene D10Wsu102e has a transcript length of 123,179 nt, it is the longest transcript considered in this exploration and unfortunately doesn’t have any particular functional annotation with the exception of a Pfam domain of unknown function PF15370.

Table 8. TargetFinder penalty score influence on the number of Hb-sRNAs having a target, the number of host gene target genes and the median number of Hb-sRNA binding sites per targeted transcripts.

Penalty score	Number of Hb-sRNAs	Number of gene targets	Fraction of Hb-sRNAs	Fraction of genes targets	Median sites per transcript	2^x
0	1	1	0.00012	4.60E-05	1	NA
0.5	1	1	0.00012	4.60E-05	1	NA
1	11	10	0.0013	0.00046	1	NA
1.5	31	33	0.0036	0.00152	1	NA
2	154	183	0.018	0.00841	1	NA
2.5	430	556	0.051	0.0256	1	NA
3	1,150	1,891	0.14	0.0869	1	NA
3.5	2,290	4,745	0.27	0.218	1	NA
4	4,086	10,393	0.48	0.478	1	1
4.5	5,765	16,139	0.68	0.742	2	2
5	7,110	20,042	0.84	0.921	4	4
5.5	7,917	21,400	0.93	0.984	9	8
6	8,292	21,686	0.98	0.997	20	16
6.5	8,419	21,731	0.99	0.999	37	32
7	8,466	21,746	1	1	65	64
7.5	8,488	21,750	1	1	101	128
8	8,495	21,754	1	1	141	256

We found a positive relationship between transcript length (log scale) and the number of predicted Hb-sRNAs binding sites (log scale) **Figure 27**. Therefore, longer transcripts tend to have more Hb-sRNAS binding sites than shorter transcripts. Here I only show results for penalty score 4, but figures for penalty scores 4, 5, 6, 7 and 8 are available in the supplementary material (**Supplementary Figure 6**). At penalty score 4 there are 10,393 mouse target transcripts for the Hb-sRNAs. The highest density of data occurs at 1 targeting sites with a transcript length between 3,500 and 4,500 nt. Some examples of outlier transcripts that display more targeting sites than the bulk of transcripts include Gm4559 with 4 targeting sites and a transcript length of 600nt and a GO term annotation of keratin filament, mitochondrially encoded NADH dehydrogenase 2 with 4 sites on a 1,038 nt transcript, involved in cell respiration. Txndc2, thioredoxin domain containing 2 with 5 sites and a length of 1,861 nt. Another example is the membrane bound vomeronasal 2, receptor 105 which has 11 sites and a transcript length of 9,899 nt. This is not an extensive list of outliers; these are some examples based on an interactive plot exploration.

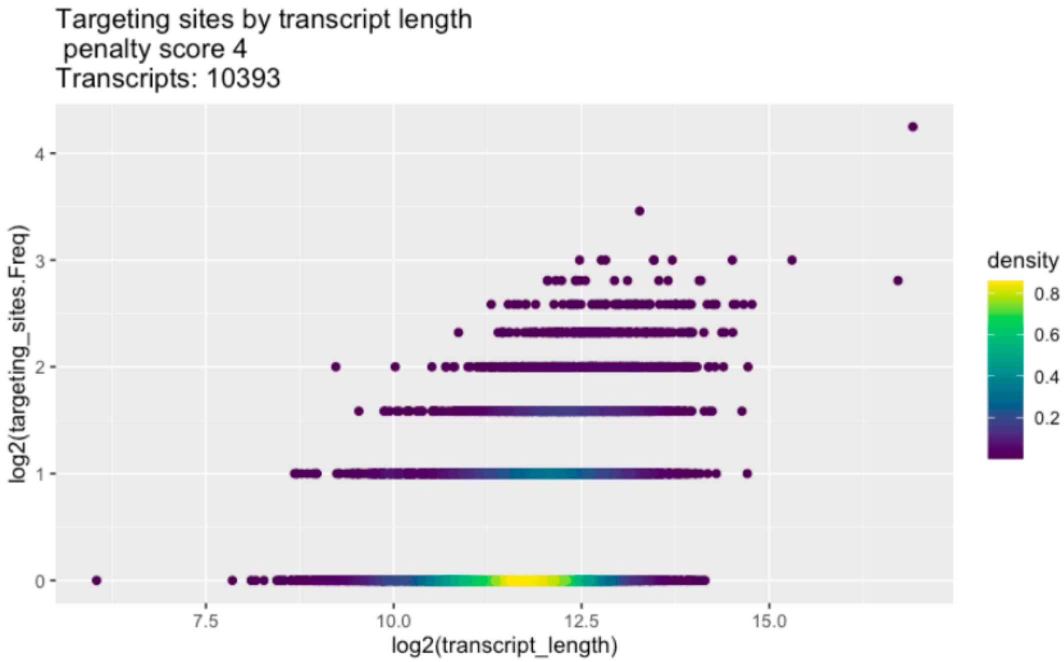


Figure 27. Relationship between host transcript length and number of predicted Hb-sRNAs binding sites.

By taking only the top 5% expressed Hb-sRNAs, we found genes of interest such as IL-33R. This receptor was previously reported by Buck et al. to be downregulated in IECs incubated with EVs (Buck et al. 2014). IL-33R is a predicted target for 9 independent Hb-sRNAs with a penalty score of 5 or less. If we extend the penalty score to 8, we find 19 different Hb-sRNAs targeting IL-33R just between positions 1950 – 2050 of the transcript (**Figure 28**). However, this finding may be due to 92.1% of the mouse genes being targeted at penalty score 5 with the whole set of Hb-sRNAs. Further explorations of mouse gene targeting coverage are needed with the top 5% expressed Hb-sRNAs to test the significance of this result.

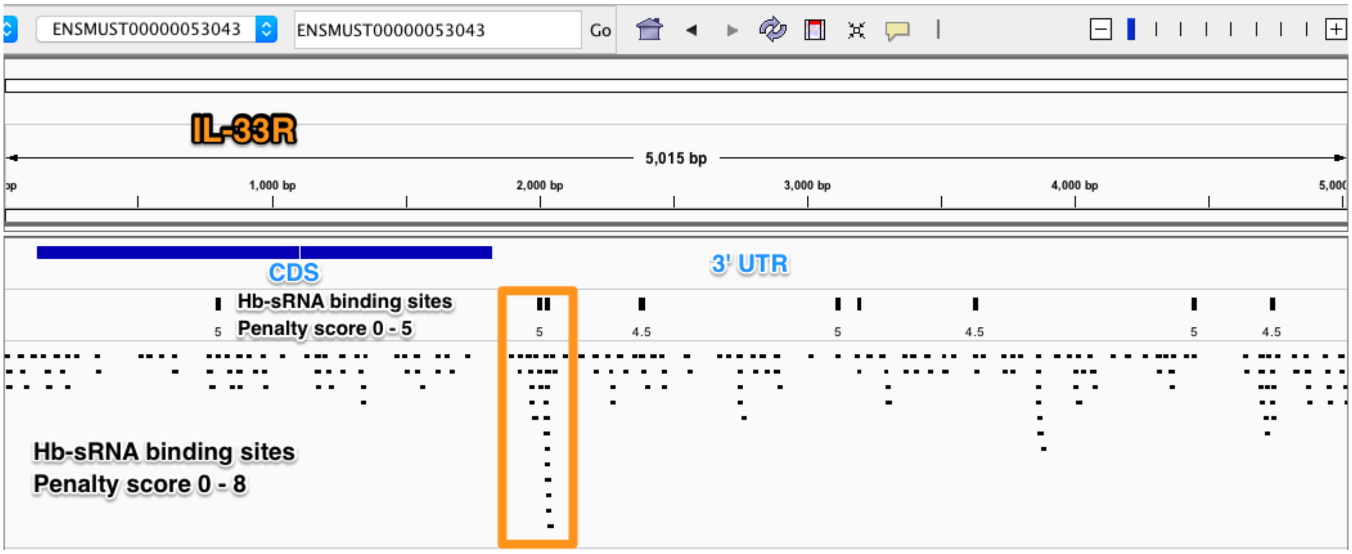


Figure 28. IL-33R (ENSMUSG00000026069), (ENSMUST00000053043). Low scoring Hb-sRNA binding sites. The coding region is shown in blue; the predicted Hb-sRNA binding sites are shown in black.

So far, we have only considered individual target sites for transcripts. We would also like to have a way to prioritize predicted targets that have multiple Hb-sRNAs binding sites over those that have fewer sites. A very simple approach to consider the number of target sites and the degree of sequence complementarity would be to simply sum the targeting sites scores. However, there's a problem with summing TargetFinder's scores, as these are penalty scores their sum doesn't reflect contribution. For example, the penalty score 0 reflects perfect complementarity between a sRNA and its target but adding 0 plus 0 give us a total of 0, which wouldn't reflect on a transcript having two perfect scoring sites relative to another one having just one. We decided to transform such scores so that these differences may be evident, we achieve this by transforming penalty scores to reward scores and to sum all the reward scores for each transcript, we call this number the "Reward summed score" (see Methods).

The reward summed score follows a very similar trend to that of the target binding sites in terms of how they are both influenced by the transcript length (**Figure 29**). This is expected, as the reward summed score depends heavily on the number of target binding sites. Again, we show only the figure with penalty score 4 as representative and the plots for penalty scores 4, 5, 6, 7 and 8 are available in the supplementary material (**Supplementary Figure 7**). Some outliers for the reward summed score include Gm4559 (also found as an outlier according to number of Hb-sRNA sites and transcript length) with a reward summed score of 3, CD209g antigen with a 1169 transcript length and a reward summed score of 4, phosphoenolpyruvate carboxykinase 2 with a transcript length of 4,625 and a reward summed score of 10.5, protein phosphatase methylesterase 1 which transcript is 2,772 nt long and its reward summed score is 10.

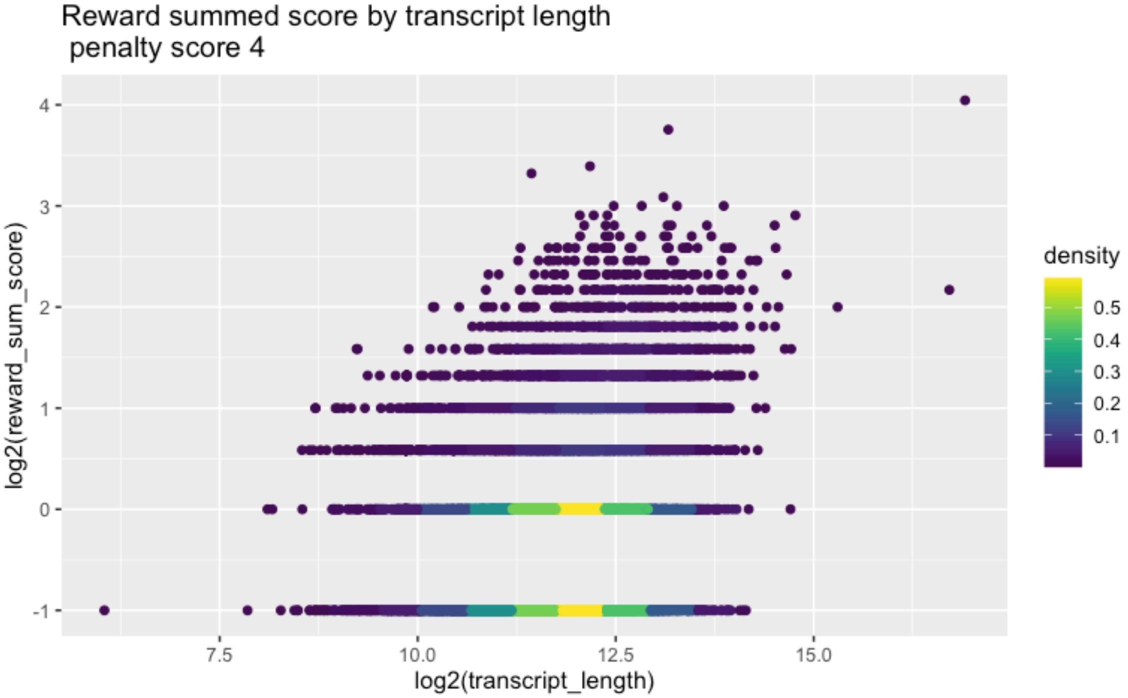


Figure 29. Relationship between host transcript length and target reward summed score.

As we are interested in immune system-related targets for Hb-sRNAs, I generate a separate plot highlighting the reward summed score for transcripts with the immune system process Gene Ontology annotation (GO:0002376) **Figure 30**. An interactive version for this plot is available in the Supplementary Material. Top immune system process targets include: RUN domain and cysteine-rich domain containing Beclin 1-interacting protein (Rubcn or 1700021K19Rik), eomesodermin, phosphoprotein associated with glycosphingolipid microdomains 1 (Pag1), cytotoxic and regulatory T cell molecule (Crtam) and CD274 antigen also known as programmed death-ligand 1 (Cd274 or PD-L1). I also highlighted Interleukins and interleukin receptors, Toll-like receptors and Arginases given their relevant role in immune system reactions.

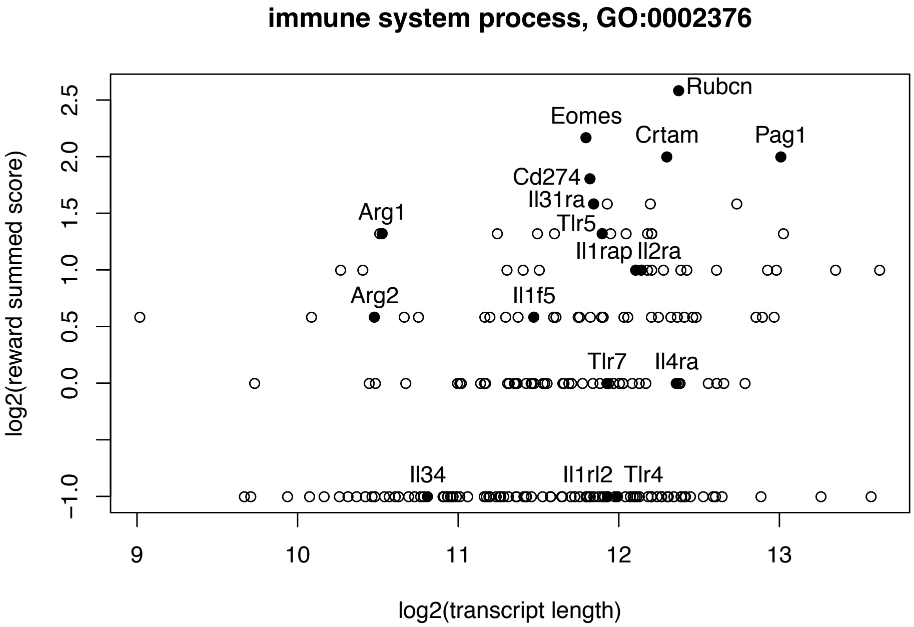


Figure 30. Relationship between host transcript length and target reward summed score for Immune system process genes at penalty score 4.

Rubcn is annotated as an autophagy-related protein according to Gene Ontology, it contains several Hb-sRNA binding sites with low penalty score **Figure 31**. Its lowest penalty score predicted binding sites are found within its coding region.

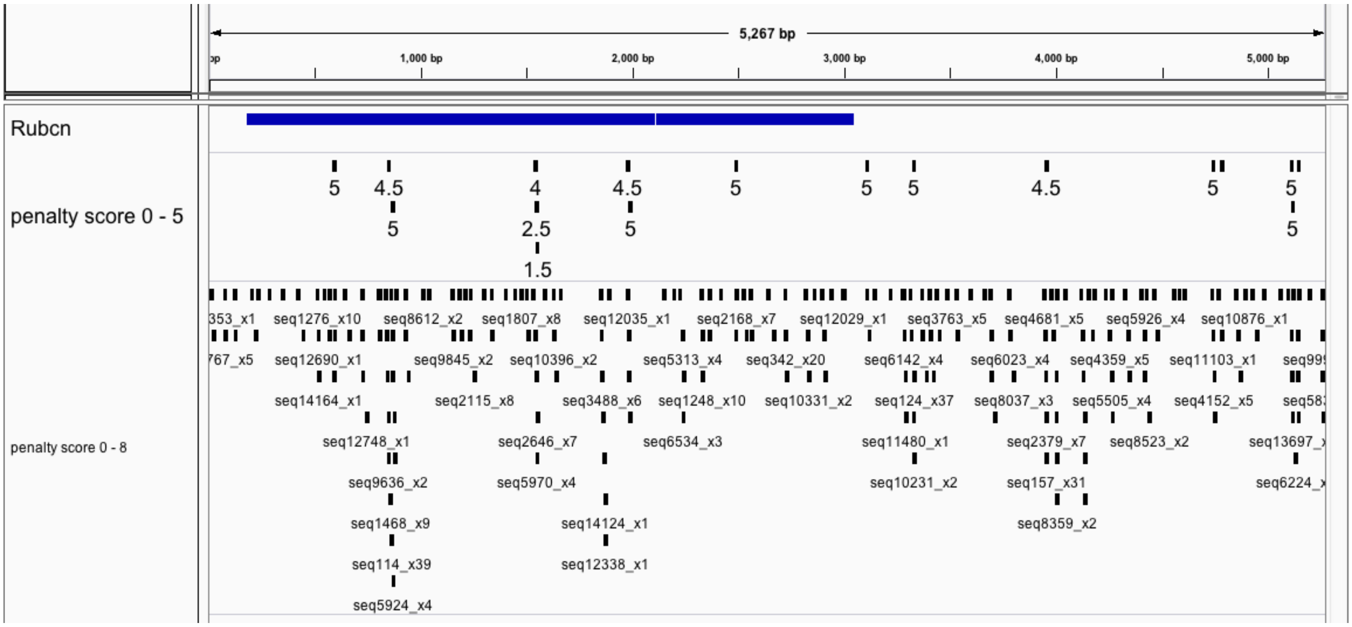


Figure 31. Rubcn (ENSMUSG00000035629), (ENSMUST00000089684). Low scoring Hb-sRNA binding sites. The coding region is shown in blue; the predicted Hb-sRNA binding sites are shown in black.

Conclusions

The end-to-end host target predictions for our set of non-redundant Hb-sRNAs reaches 47.8% of mouse protein coding genes according to TargetFinder authors' recommended parameters. However, host target coverage is highly sensitive to variations in the chosen penalty score, reaching more than 90% coverage with just a small increase in this threshold.

The excessive targeting issue is likely due to the high number of Hb-sRNAs used for target prediction. We predicted targets for thousands of Hb-sRNAs (8,501) despite applying a reduction based on sequence similarity. Importantly, we observed stark contrasts in Hb-sRNAs expression levels that may be relevant to prioritize Hb-sRNAs predictions for downstream analyses.

Perspectives

1. Describe the occurrence of Hb-sRNAs binding sites in different transcript regions of protein-coding genes: 5'UTR, CDS, 3'UTR. We know that endogenous miRNAs have a preference to target 3'UTRs and somewhat CDS, so any preference for the Hb-sRNAs could reflect on their unique biology.
2. Fit a model that relates transcript length with reward summed score and to search for outliers. We would like to derive a score that is less biased due to the length of the transcript.
3. Explore which high penalty scores lack biological significance. Is a penalty score of 8 still worth working with? Use a miRNA-transcript example and check the penalty score of a seed-only alignment. Nevertheless, for many of these refinements, it would be useful to have positive and negative controls of true Hb-sRNA targets with experimental validation.

4. Explore the possibility of predicting target sites on both transcript strands and to use the reverse-complement strand predictions as a negative control in further analyses.
5. Seed-based target predictions with TargetScan. We have some preliminary results, but these predictions quickly cover an even higher percent of the mouse transcriptome. Thus, it will be highly important to have controls as well as to validate how to consider a transcript targeted by multiple Hb-sRNAs.
6. To explore the length and targeting score relationship for genes involved in the immune response and search for high score outliers to prioritize Hb-sRNAs regulated genes.

Chapter 4: Assesing the effect of nematode secretion on mouse cells and testing for downregulation of predicted targets for *Heligmosomoides* sRNAs

In this last chapter we explore the effect of HES or EVs on mouse cells, using measurements obtained by RNA-Seq. We then test if there's any effect of the *H. bakeri* sRNAs present in the nematode material being applied, according to those we detected in Chapter 2 and our target predictions made in Chapter 3.

Methods

In order to design our RNA-Seq experiment, we had to select certain samples to focus on, since we had too many in our initial sRNA-Seq experiment to be cost effective. We focused on BMDM treated with HES and IECs incubated with HES or EVs. The reason for choosing IECs with EVs for RNA-Seq was that in this condition we detected the highest Hb-sRNAs signal (Chapter 2). We chose HES treatments for both cell types as some of the immunosuppression effects previously reported are stronger using HES that only EVs or supernatant (Buck AH. personal communication). The reason to choose 24 hours was that 4 hours was probably too early to detect an effect on mouse cells. The chosen libraries from which mRNA was sequenced are shown with asterisks in **Table 6**.

We used two different methodological procedures, a genome-based approach and a transcriptome-based approach (see Methods). Aligning reads to the reference mouse genome, we will refer to this as a genome-based approach from now on, and by doing pseudoalignment quantification of mouse transcripts expression, that we will refer to as transcriptome-based approach. Both approaches converge to a gene counts table and to a differential expression analysis in search for an effect of HES or EV treatment on host cells. True biological effects should be robust and detected by either methodological approach.

Read quality and adapter trimming

We performed quality analyses with fastqc version 0.11.2 (Andrews 2010) before and after adapter trimming. We removed the adapters and filtered low quality sequences with Trimmomatic version 0.32 (Bolger, Lohse, and Usadel 2014) using the following parameters: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15 and MINLEN:36.

Gene expression quantification

We had two different approaches for gene expression quantification: a genome-based approach and a transcriptome-based approach. The reason for using these two approaches is that we wanted to look for robust gene expression differences, and these should be present in both quantification methodologies. For the genome-based approach, we used the splice-aware aligner STAR version 2.7.0f (Dobin et al. 2013). To count the reads associated to genes we used featureCounts (Liao, Smyth, and Shi 2014) function which is part of the Rsubread package (version 1.34.2). We provided the GTF file with annotations and turned on the flag isPairedEnd. We obtained the GTF file with *Mus musculus* gene annotations from Ensembl release 96 (Yates et al. 2020). For the transcriptome-based approach we used salmon version 0.11.3 (Patro et al. 2017). We then used tximport version 1.8.0 (Soneson et al. 2015) to map the resulting transcript quantifications at the gene level.

Differential expression analysis

I divided the differential expression analyses (DEAs) according to cell type: Bone marrow-derived macrophages (BMDM) and intestinal epithelial mouse cells (IECs). Differential expression (DE) analysis was done using the edgeR package in R (M. D. Robinson, McCarthy, and Smyth 2009). A gene was considered for DE analysis if at least one library had 0.5 counts per million reads. This allowed us to consider genes with low-expression that might still be potential targets.

The comparisons of interest are the following to test that Hb-sRNAs do affect transcription of mouse genes and to asses the effects of *H. bakeri* products on these mouse cells:

- BMDM HES-treated vs BMDM untreated (DEA I)
- IECs HES-treated vs IECs untreated (DEA II)
- IECs vesicle-treated vs IECs untreated (DEA III)

Differential expression analyzes were done using the edgeR package. Lowly expressed host genes were filtered; only those that had at least 0.5 count per million in any of the libraries were kept. We used the method of trimmed mean of M-values (M. D. Robinson and Oshlack 2010) to calculate normalization factors to deal with different library sizes, using the calcNormFactors function (M. D. Robinson and Oshlack 2010). Then the common, trended and tagwise dispersions were estimated using the function estimateDisp (Phipson et al. 2016). Differential expression was determined using the “Generalized linear model likelihood ratio” test (McCarthy, Chen, and Smyth 2012) with the function glmLRT. We considered a false discovery rate (FDR) < 0.01 to define differentially expressed genes.

Test for downregulation of Hb-sRNAs target predictions in host

I took 5%, 10% or 25% of the highest expressed and lowest expressed Hb-sRNAs that had predicted targets from the total of 8,495. Then, I calculated a reward summed score for targets of these two sets of Hb-sRNAs with different penalty score thresholds 3, 4, 5 and 6. We then removed overlaps between highly and lowly expressed Hb-sRNAs. We then divided the predictions into best and worst predictions by taking 25% top targets and 25% bottom targets from a data frame that was ordered from highest to lowest reward summed score, this with the aim to evaluate if different reward summed score values were related to more or less repression of targets. We then compared the fold change distributions for mouse targets of the following different sets:

- highly expressed Hb-sRNAs best target predictions,
- highly expressed Hb-sRNAs worst predictions,
- lowly expressed Hb-sRNAs best predictions
- lowly expressed Hb-sRNAs worst predictions
- non-targets for a given penalty score threshold.

Fold change distributions were compared with the Mann-Whitney test implemented in the wilcox.test in the R stats package (Bauer 1972).

Results and discussion

Effects of HES and EVs on mouse cells

To assess the effects of HES and EVs on mouse cells, we sequenced 14 samples out of 42 libraries included in our sRNA-Seq experimental design (see Methods). I used two different approaches to analyze this RNA-Seq data, a genome-based and a transcriptome-based approaches (see Methods). In the genome-based approach, more than 80% of the aligned reads were assigned to a feature. Between 8% to 13% of reads did not overlap any feature, and 5% were ambiguous between two or more features (**Supplementary Figure 8**). On average, 90% of the assigned reads overlap with protein coding genes. Between 5-7% of the assigned reads overlap with processed pseudogenes. The rest of the reads overlap with other biotype categories (**Supplementary Figure 9**). In the transcriptome-based approach we get an average mapping rate of 87% to protein coding transcripts.

As part of the quality control for mouse RNA-Seq we produced a multidimensional scaling plot (MDS). MDS analysis is a technique used to reduce the multidimensional nature of thousands of genes to be able to represent this information in a two-dimensional plot. The distance between points in the resulting plot represents differences in the gene expression profile of libraries. Libraries that appear closer in an MDS plot are more similar than libraries that are plotted farther from each other. We expect replicate libraries to cluster together and libraries with different treatments to separate from each other.

According to the MDS plot including both BMDM and IECs, the first dimension separates these cell types, the second-dimension separates HES-treated from untreated BMDM libraries. All the IEC libraries cluster together and it is hard to tell if their distribution of secretion and control libraries is appropriate (**Figure 32A**). For this reason we made MDS plots for BMDM and IECs separately. For BMDM the first dimension separates BMDM HES-treated from BMDM untreated libraries. The second dimension separates B_HES_24_2 from B_HES_24_1 and B_HES_24_3 (**Figure 32B**). The libraries separate appropriately to perform a differential expression analysis and compare the effect of HES incubation relative to untreated libraries.

Regarding the IECs MDS plot, the first dimension locates IECs HES-treated at the leftmost part of the plot, IECs vesicle-treated appear in the middle and MODE-untreated libraries appear on the right extreme. This distribution of libraries suggests that the effect of EVs on IECs gene expression is more subtle than that of HES treatment. This effect is evident on both genome-based and transcriptome-based approaches (**Supplementary Figure 10**). The second-dimension places IECs vesicle-treated on top, and HES-treated cells and control libraries below (**Figure 32C**). MDS plots revealed appropriate separation of treatments and grouping of replicate libraries, therefore we proceeded with differential expression analyses.

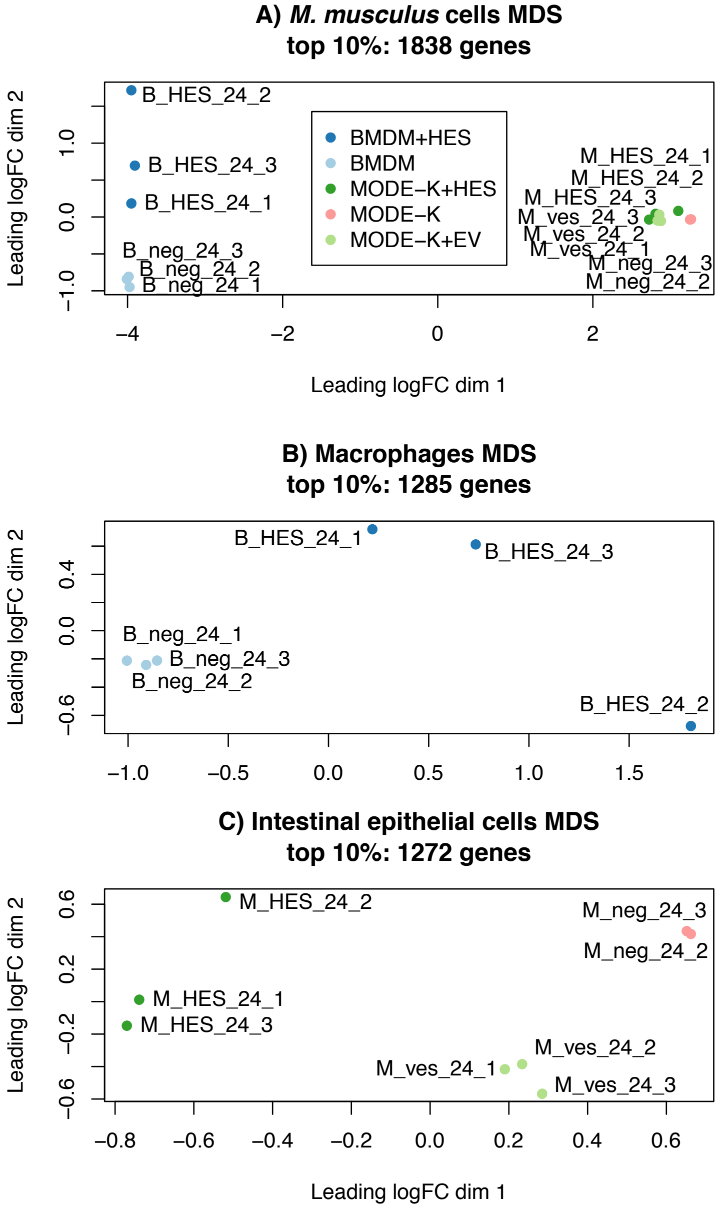


Figure 32. Multidimensional scaling plot for mouse cells treated with *Heligmosomoides* secretion products. A) BMDM + IECs B) BMDM C) IECs. In all cases we used the top 10% variable genes. Replicates are colored the same.

A summary for the number of differentially expressed genes across our three comparisons is found in **Table 9**. EVs exert the most subtle effect on IECs with only 16.8% changing significantly, in contrast HES incubation causes ~35.3% of genes to be either up or downregulated. In BMDM HES causes ~18.2% of genes to be differentially expressed. Interestingly, there's a slight bias to downregulation for IECs treated with EVs (up/down ratios: 7.6% to 9.2%, and 10.8% to 15.2%). This downregulation bias does not occur in IECs incubated with HES. This effect is evident regardless of the methodology used for the analyses.

Table 9. Summary of differentially expressed mouse genes with different tools.

	Genome-based approach	Transcriptome-based approach
	STAR software	Salmon software
BMDM HES		
Upregulated	1,574 (12.2%)	1,142 (8.1%)
Non-differentially expressed	10,498 (81.6%)	11,995 (85.3%)
Downregulated	778 (6.0%)	919 (6.5%)
IECs HES		
Upregulated	2,320 (18.2%)	1,902 (13.7%)
Non-differentially expressed	8,225 (64.6%)	10,268 (74.0%)
Downregulated	2,176 (17.1%)	1,702 (12.2%)
IECs EVs		
Upregulated	978 (7.6%)	536 (3.8%)
Non-differentially expressed	10,563 (83.0%)	12,471 (89.9%)
Downregulated	1,180 (9.2%)	865 (6.2%)

When I looked at the fold change and expression patterns, I noticed that for both cell types there are genes that weren't expressed and were turned on when exposed to *Heligmosomoides* secretions. This is represented in our MA plots as lines of genes with high positive logFC values that separate from the rest of the genes in the shape of distinct diagonals (**Figure 33**).

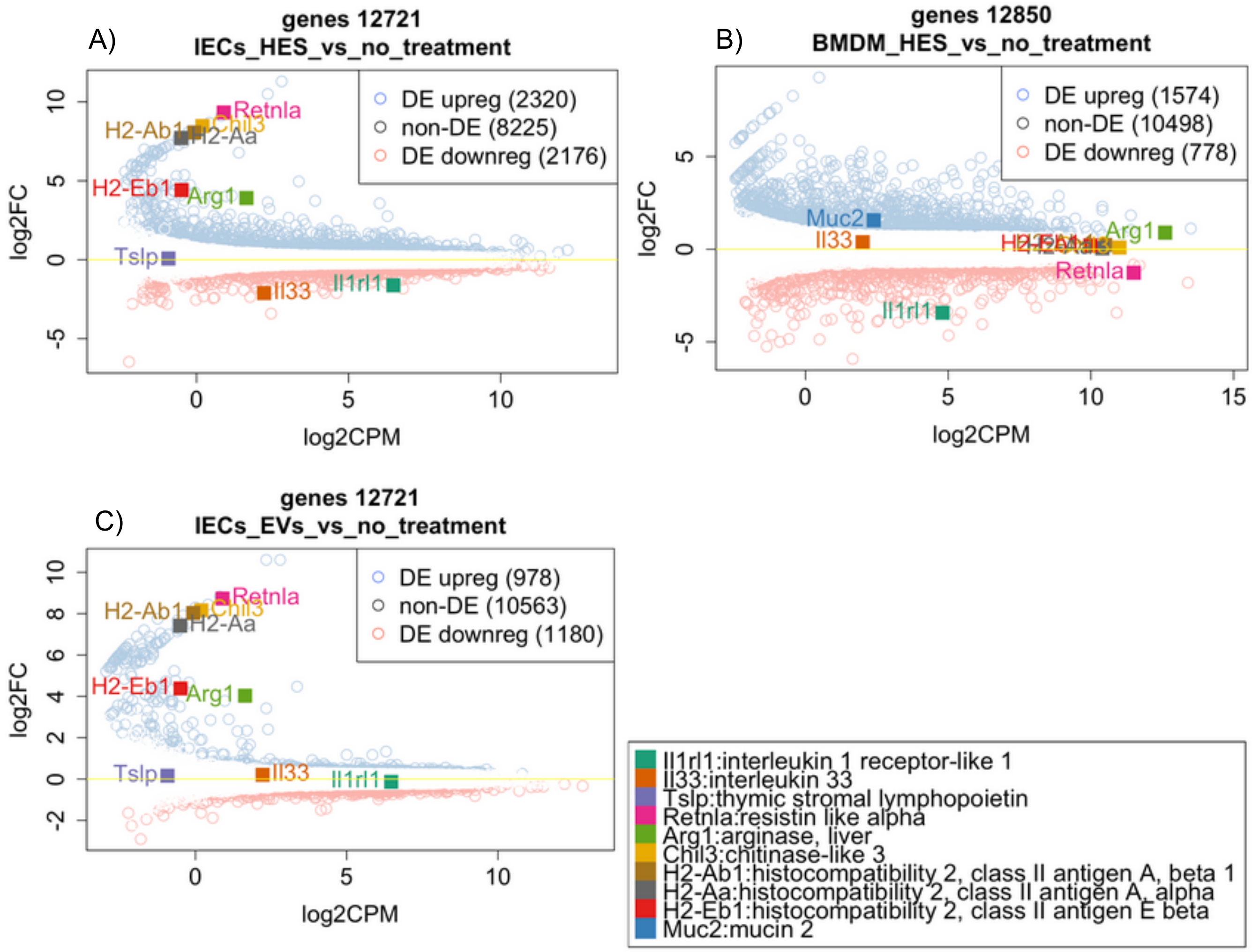


Figure 33. Mean abundance plots for mouse cells with *Heligmosomoides* secretion products. A) IECs HES vs control IECs. B) BMDM HES vs control BMDM. C) IECs vesicle vs control IECs. X axis represents average $\log_2$ counts per million. Y axis shows $\log_2$ fold change. Differentially expressed sRNAs are shown in red with an FDR <= 0.1.

Macrophages response to HES

For macrophages, the first thing I looked at was if there were any differences in the three typical markers for macrophage alternative activation: Resistin-like alpha (Retnla), Arginase-1 and chitinase-like 3 (Chil3/Ym-1). These three genes display high expression in BMDM with > 10 $\log_2$CPM. Retnla is repressed due to HES in macrophages, but not Arg-1 nor Chil3 (**Figure 33**). This result differs from previous observations by Coakley and collaborators, where they reported that *Heligmosomoides* HES, vesicles or even the supernatant suppressed the expression of these same genes (Coakley et al. 2017). However, a key difference between the current and Coakley experiments is that she added IL-4 and IL-13, which are Th2 and AAM trigger cytokines. We didn't add IL-4 nor IL-13, and therefore there's no trigger for AAM in our experimental setting. What we would see in our experiment is the isolated response from macrophages or IECs to *H. bakeri* secretions.

Macrophages also upregulate mucin 2 transcript due HES treatment, but its expression is not apparent in IECs **(Figure 33)**. Mucins are high molecular weight and glycosylated proteins produced by epithelial tissues in most animals. The gel-forming mucin 2 is particularly abundant in gut, where it is secreted into the intestinal lumen by goblet cells (Johansson and Hansson 2016). The absence of expression of mucin 2 for IECs may suggest that goblet cells are the dominant source for this protein, and that enterocyte contribution is modest. The upregulation of mucin 2 in macrophages still needs an answer, it is unclear to which degree macrophages secreted products reach the intestinal lumen.

As we are dealing with EVs and macrophages, I was curious to understand if the process of phagocytosis was affected. We had a look at the phagosome KEGG pathway and overlaid the HES-treated macrophages gene DEA fold change information (**Figure 34**). Major histocompatibility complex I and II were upregulated, as well as some phagocytosis promoting receptors such as Toll-like receptor 2 and CD14. Some downregulated genes that caught my attention were Rab5, Rab7, and RILP (Rab interacting lysosomal protein) as these proteins are relevant for intracellular vesicular trafficking. Rab5 is involved in early endosome generation and Rab7 is involved in late endosome maturation and targeting endosomes to the lysosome for degradation with the aid of RILP (Guerra and Bucci 2016). A closer inspection revealed that in this figure Rab7 condenses fold-change information for Rab7 and Rab7b. Rab7b is the protein that is truly down-regulated due to HES treatment in macrophages ($\log_2$FC -2.22, FDR 3.7×10^{-19}). If *H. bakeri* EVs were to have an effect on macrophages it may be important for them to avoid degradation by the lysosome. In fact, Coakley observed higher EVs and lysosomes co-localization in macrophages when anti-EV serum (containing antibodies) was applied to macrophages, and less co-localization when this serum was absent (Coakley et al. 2017). This suggests that at least a portion of EVs may be able to escape lysosome-mediated degradation. It would be interesting to determine if the downregulation of Rab7b, RILP or Rab5 in macrophages would result in reduced EV targeting to lysosomes by a co-localization experiment.

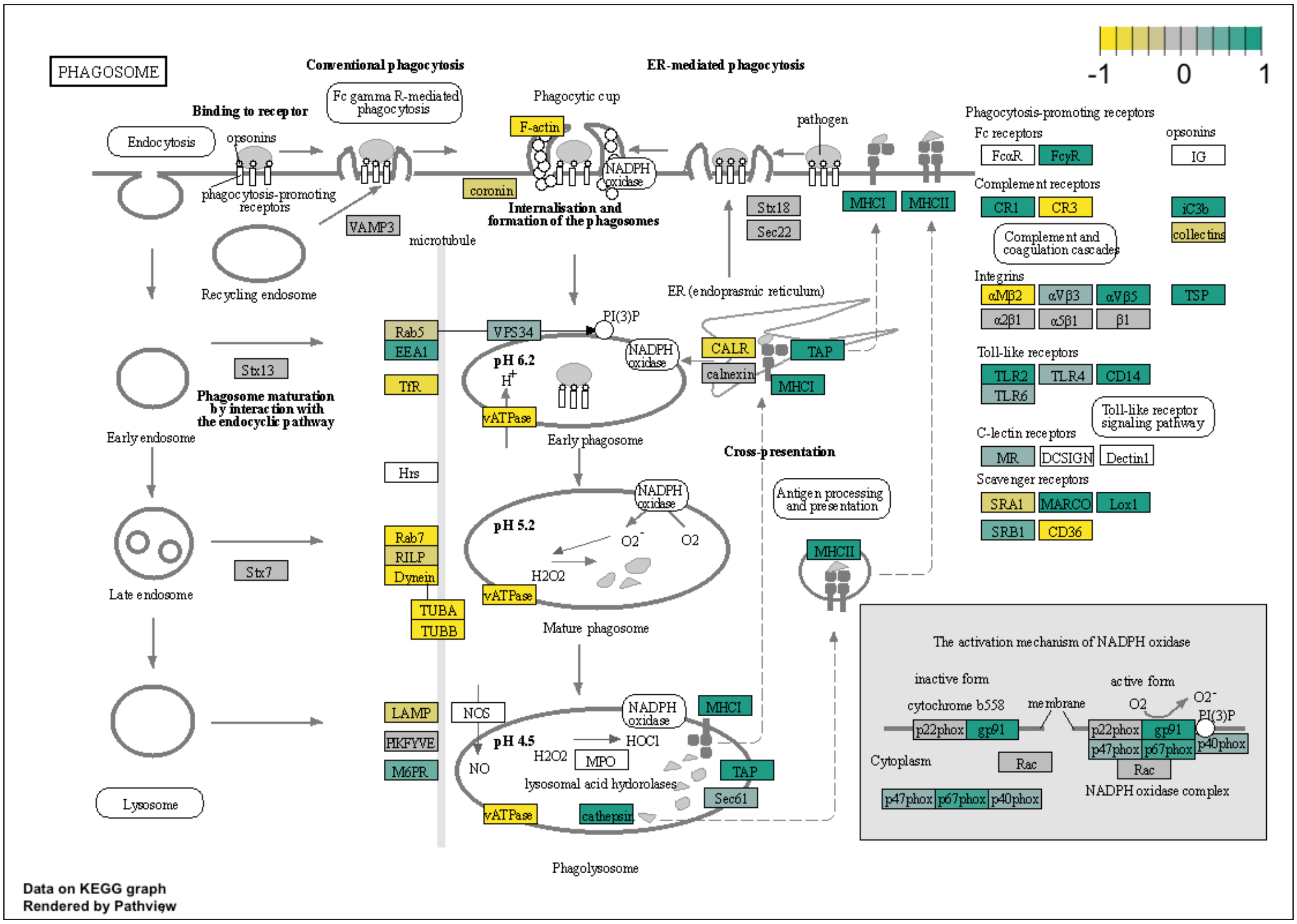


Figure 34. Phagosome pathway in BMDM treated with HES vs no treatment. Yellow-colored genes are downregulated, green-colored genes are upregulated, grey-colored genes do not show differential expression and white colors are absent from the differential expression analysis.

Intestinal epithelial cells response to HES or EVs

IECs turn on the AAM classical markers chitinase-like 3 and resistin-like alpha genes, as well as upregulate Arginase-1. These AAM markers, as well as three chains (H2-Ab1, H2-Aa and H2-Eb1) of the major histocompatibility complex II (MHC II), are found in the “turned on due to secretion” gene line for IECs (**Figure 33**). The expression of MHC II by IECs has been reported previously (Vidal et al. 1993), although some authors still question the capacity for IECs to present antigens.

The secretion of resistin-like alpha by IECs has been previously shown to attract eosinophils *in vitro* (Munitz et al. 2008). In the same study, intraperitoneal injection of Relm-alpha and posterior differential cell analysis resulted in an increase of neutrophils, eosinophils and lymphocytes (Munitz et al. 2008). IECs secretion of Relm-alpha may be a chemotaxis signal for myeloid cells upon *H. bakeri* infection.

Sutherland etl al. (Sutherland et al. 2018) reported the expression of Chil3 in lung epithelial cells, to the best of my knowledge there are no reports of Chil3 expression in IECs. Chil3 may also be a chemoattractant signal for eosinophils as previously shown by Owhashi and collaborators (Owhashi, Arita, and Hayai 2000).

Talavera et al. (Talavera et al. 2017) reported the upregulation of iNOS and Arginase II (but not Arginase I) in a rat IEC line upon exposure to LPS. They found that Arginase II inhibition resulted in increased apoptosis levels in the cell culture, presumably due to NO production. In the context of my results, Arginase I may have a role in producing ornithine, which has been shown to reduce *H. bakeri* larval mobility *in vitro* (Bieren et al. 2013). Further experiments may reveal if ornithine also affects adult nematodes.

IECs secrete cytokines such as IL-33, IL-25 and thymic stromal lymphopoietin (TSLP) to warn immune system cells about nematode infection (Maizels et al. 2012). We found that the IL-33 is repressed more than fourfold in IECs due HES (log_2FC 2.11, FDR $2.4x10^{-37}$), but shows no statistical difference in EV treatment (log_2FC 0.202, FDR 0.22) (**Figure 35**). TSLP doesn't respond to either secretion treatment (log_2FC 0.07, FDR 0.88 for HES, log_2FC 0.1, FDR 0.8 for EV). We didn't find any evidence of expression for IL-25 in IECs nor BMDM. It is interesting to speculate if tuft cells may be the primary source for IL-25 during gastrointestinal parasites infection (Gerbe et al. 2016), and if the gross population of intestinal epithelium cells may contribute in a limited fashion to the IL-25 pool. If this turns to be true, then I propose that tuft cells may be an attractive cell type to look for cell surface receptors that may detect gastrointestinal parasites.

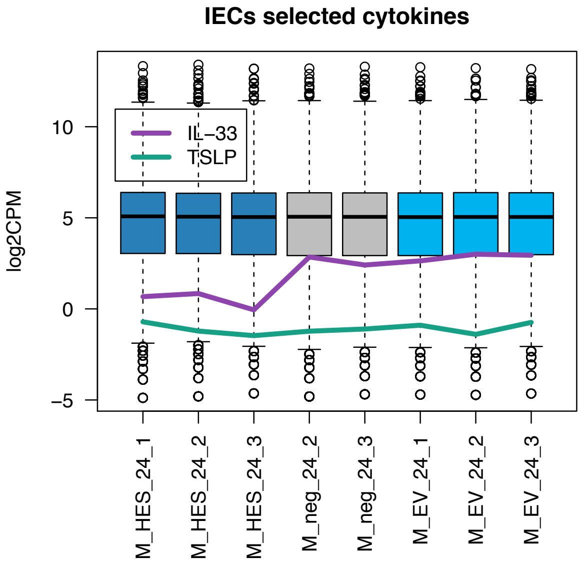


Figure 35. Influence of *Heligmosomoides* secretion products on the expression of intestinal epithelial cell selected cytokines. Nematode treated libraries are shown as blue boxes and control untreated libraries are shown in grey. M stands for IECs, HES stands for *Heligmosomoides* excretion-secretion product, neg stands for negative control, untreated library and EV stands for extracellular vesicles. The expression for IL-33 and TSLP are shown in purple and green respectively.

The IL-33 receptor ST2 subunit (Il1rl1) is repressed due HES in both macrophages (log_2FC -3.52, FDR $1.29x10^{-40}$) and intestinal epithelial cells (log_2FC -1.62, FDR $3.6x10^{-90}$) (**Figure 36**). This repression is not evident in EV-treated IECs (log$_2$FC -0.135, FDR 0.18) and, unfortunately, we lack EV-treated macrophages libraries to know if EVs influences the expression of this receptor in BMDM.

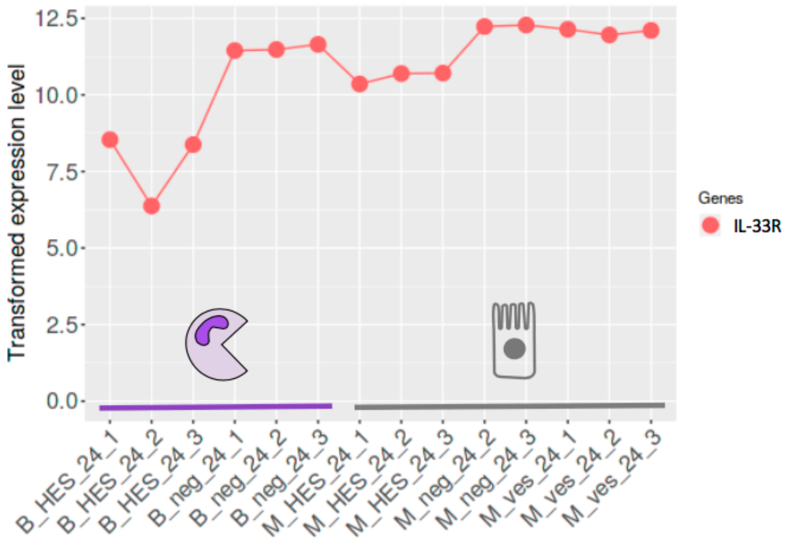


Figure 36. Interleukin 33 receptor expression is repressed in the HES treatment both in BMDM (B) and IECs (M).

I didn't find evidence of expression for other nematode-relevant genes such as: resistin-like beta (Retnlb), a protein reported to be produced by intestinal epithelial cells (specially goblet cells) and to directly affect nematode feeding (Herbert et al. 2009). Pla2g1b is an epithelial cell-derived lysophospholipase that is relevant to clear *Heligmosomoides* secondary infections (Entwistle et al. 2017). Pla2g1b was lost during the low expression filtering before the differential expression analysis.

Are *H. bakeri* small RNAs responsible for HES or EV effects?

Here we tested if the *H. bakeri* sRNAs present in HES or EVs have any influence on mouse cells gene expression due to binding to potential targets. To achieve this goal, we used the Hb-sRNAs detected in Chapter 2, the target predictions made in Chapter 3 and the differential expression results described in the previous section of this chapter.

Most sRNAs decrease the expression levels of their direct transcript targets. Therefore, we would expect Hb-sRNAs to repress mouse transcripts and this would be reflected in RNA-Seq reads count difference (fold change) between secretion-treated and untreated libraries (**Figure 37**). If this scenario occurs for many Hb-sRNAs, then the fold change distribution for targets should be biased relative to the fold change distribution of non-target transcripts towards more negative fold change values. We will refer to this fold change distribution bias as repression during the rest of this work. We will use cumulative distributions to present the results for this test, as these representations facilitate comparisons between two or more distributions. In cumulative distribution plots, the distribution of target genes should be shifted to negative fold change values (left part of the plot) relative to non-targets.

We compared fold change (logFC) distributions for five different sets: best targets or worst targets for highly expressed Hb-sRNAs (h-Hb-sRNAs), best or worst targets for lowly expressed Hb-sRNAs (l-Hb-sRNAs) and non-targets (see Methods). These sets were defined according to Hb-sRNA expression and the summed reward score (see Methods). We expect to find the highest repression signal for the best targets of the highly expressed Hb-sRNAs, a more subtle effect for the remaining three target sets. We would also expect any target set to display more repression than the non-targets set.

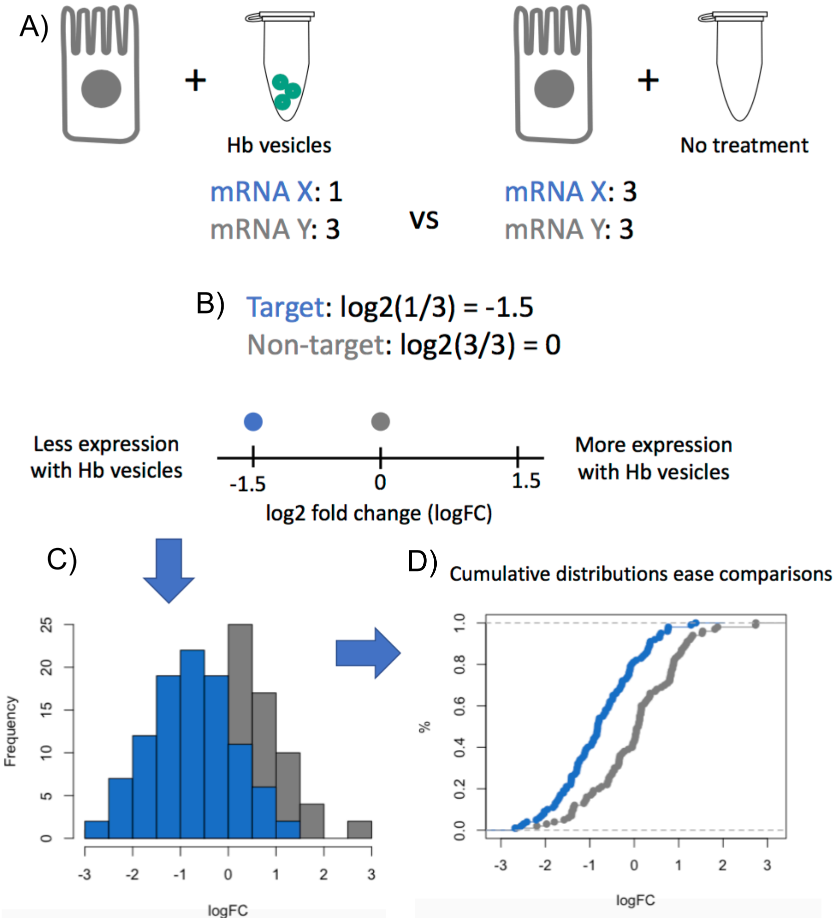


Figure 37. Strategy to test for an overall repression effect for *H. bakeri* sRNAs (Hb-sRNAs) in mouse cells. A) Hb-sRNAs treatment would reduce expression of targets in mouse. B) Repression is associated with negative fold change. C) An overall repression effect for Hb-sRNA targets would bias the fold change distribution of targets relative to non-targets. D) Cumulative distributions ease comparison of two or more sets.

In BMDM with HES, the best targets for highly expressed Hb-sRNAs are biased to more negative logFC values than worst targets for this same set of sRNAs (p-value = 5.2×10^{-07}). Top targets for highly expressed Hb-sRNAs also display more repression than non-targets (p-value = 1×10^{-13}) (**Figure 38A**). We also see a shift in the distribution of best targets for lowly expressed Hb-sRNAs, this shift is not present in the worst predictions. This bias is what we would expect if Hb-sRNAs would have a repression effect on the best host targets, the effect is subtle but enough to be detected by statistical tests.

We know that longer transcripts tend to have more predicted Hb-sRNAs binding sites and hence, better reward summed scores (**Figure 29**). When we look at the length distributions of these five different sets, we find that best predictions tend to have slighter longer transcripts than worst predictions, and any of these target sets tends to have longer transcripts than non targets (**Figure 38B**). This suggests that the repression effect is not due to longer transcript lengths for the best targets of highly expressed Hb-sRNAs, as these have a similar length distribution to that of the best predictions for lowly expressed Hb-sRNAs. These observations for transcript lengths distribution apply to all three contrasts.

For IECs with HES treatment top targets for h-Hb-sRNAs have a shifted logFC distribution to negative values than worst targets (p-value = $4x10^{-04}$) or non-targets (p-value = $2.3x10^{-08}$). The best predictions for l-Hb-sRNAs show a similar bias that suggests repression, this bias is nor present in the l-Hb-sRNAs worst targets (**Figure 38C**). We observe a similar effect on IECs treated with EVs (**Figure 38E**), best targets for h-Hb-sRNAs are biased towards negative fold change values relative to h-Hbs-RNAs worst targets (p-value = $1.8x10^{-08}$) or relative to non-targets (p-value = $7x10^{-13}$). Again, we observe a shift for the best predictions of l-Hb-sRNAs relative to their worst predictions or non-targets. We found very similar results with the transcriptome-based approach (**Supplementary Figure 11**).

I next wanted to know the identity of these h-Hb-sRNAs targets that behave as expected. We have 536 targets in macrophages and 538 targets in IECs with HES or EVs (**Figure 39**). The union of these biased best targets results in 722 unique genes, 372 (51.5%) of them are shared among the three comparisons, 121 (16.7%) are exclusive for the macrophage contrast, 38 (5.2%) are exclusive for IECs with HES and 45 (6.2%) are exclusive for EV-treated IECs. Among the best targets present in all comparisons we found slingshot protein phosphatase 2 (Ssh2), this was also the only mouse transcript that had a perfect complementary binding site to an Hb-sRNA (Chapter 3). We also find Interleukin-1 receptor-associated kinase 2 (Irak2), a kinase involved in Toll-like receptor signaling. According to Maizels lab, Irak2 transcript levels are reduced in Dendritic cells incubated with HES (Kemter 2016, PhD thesis dissertation).

During the search for an effect of Hb-sRNAs, we noticed a transcript length bias in IECs DE analyses (**Figure 40**). Longer transcripts tend to be upregulated relative to all transcripts, additionally, shortest transcripts have a slight tendency towards repression compared to all transcripts. These length biases are found in IECs with HES and IECs with EVs, but not in BMDM with HES. We still don't know the possible source of this length bias. This length bias can be a confounding factor in our tests for target effects.

These results suggest that there may be a down-regulation of the best target predictions for our Hb-sRNAs targets relative to worst predictions or non-targets in both cell types (BMDM or IECs) and with either nematode secretion product (EV or HES). This downregulation signal is supported by two different gene quantification strategies: genome-based with STAR and transcriptome-based with salmon (**Supplementary Figure 12**). However, to be completely convinced that this repression signal is true we would like to have additional controls, such as target predictions for permuted Hb-sRNA sequences as an additional negative control, as well as qPCR validation for selected targets.

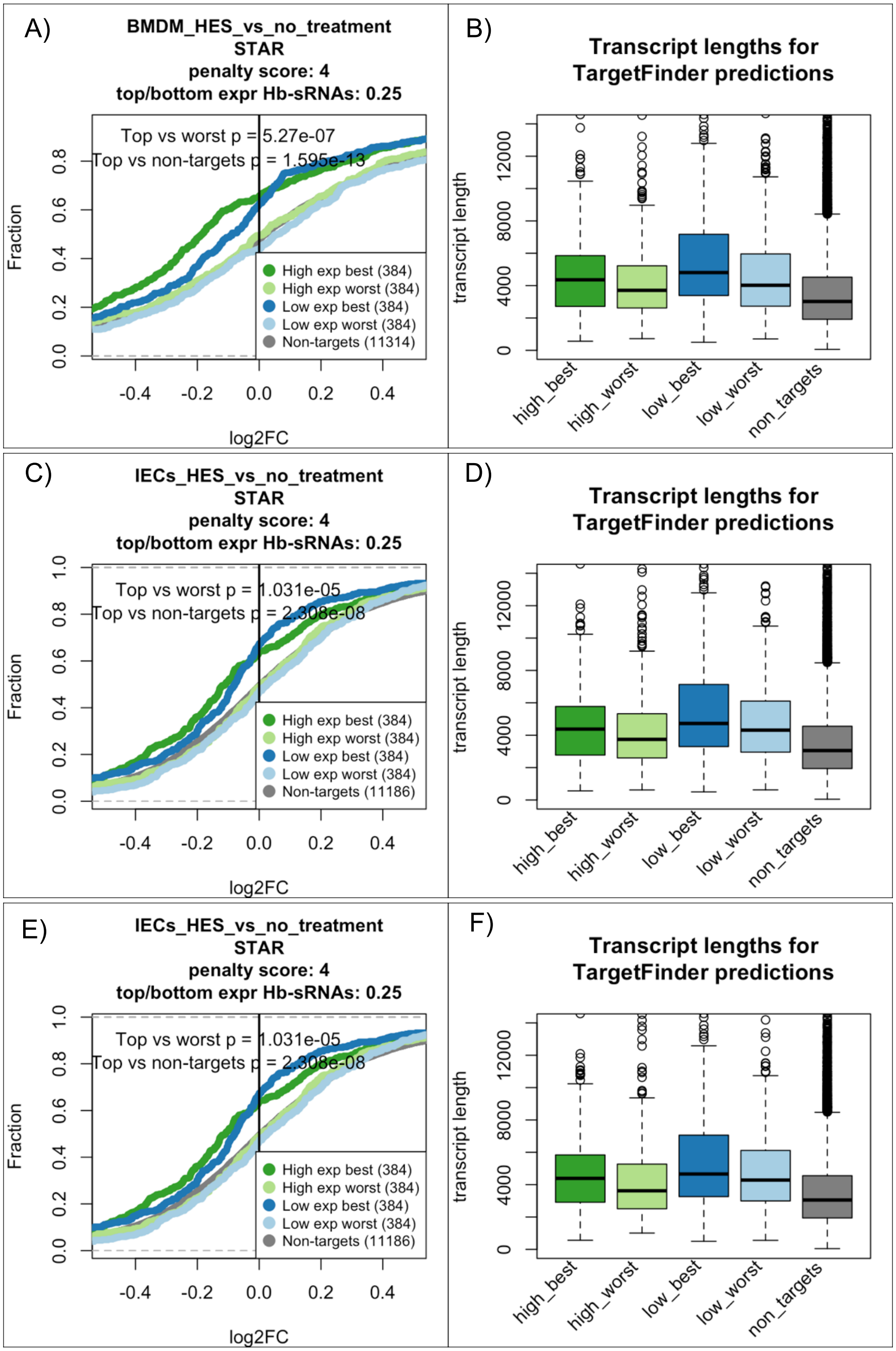


Figure 38. Cumulative distribution comparison between highly expressed *H. bakeri* sRNAs (Hb-sRNAs) best targets (dark green), worst targets (light green), lowly expressed Hb-sRNAs best targets (dark blue), worst targets (light blue) and non-targets (grey) for Hb-sRNAs in mouse. A) BMDM + HES C) IECs + HES D) IECs + vesicle. The shown p-values are the result of a Mann-Whitney test between best vs worst targets and top vs non-targets. B), D) and F) Transcript length distributions for the same sets shown in each left panel.

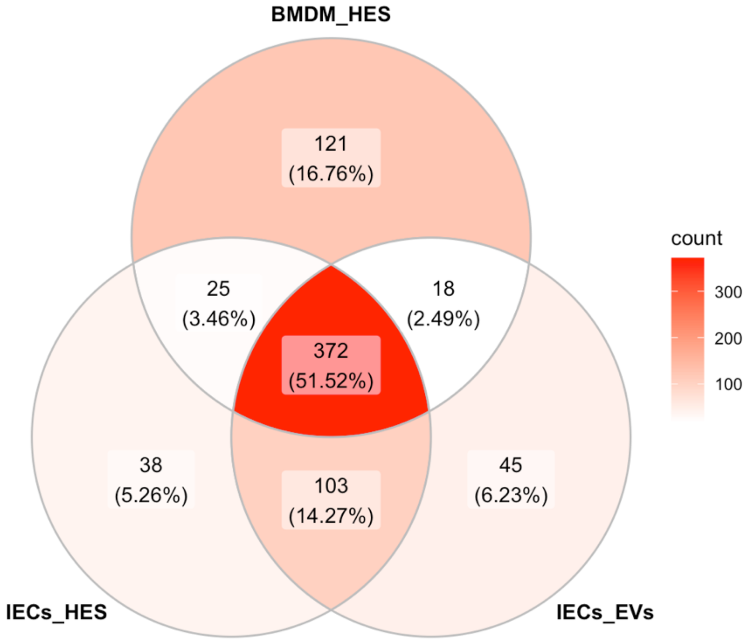


Figure 39. Comparison of best predicted targets for highly expressed *H. bakeri* sRNAs detected in cells.

We used different penalty score thresholds (3, 4, 5 and 6), as well as different portions of the highest and lowest expressed Hb-sRNAs (5%, 10% 25%) in order to look for a repression signal. The figures for all these results can be found in the supplementary material **(Supplementary Figure 12)**. Some combinations of parameters yield repression signal while others don't. In general, the penalty score 4 threshold provides repression signal for the best h-Hb-sRNA targets for macrophages or IECs (HES or EVs), regardless of the fraction of highest expressed Hb-sRNAs (5%, 10% or 25%). However higher or lower penalty scores result in variation of repression signal, which raises questions about the appropriate parameters to perform these analyses. There's still room for improvement for repression signal search, such as varying the fraction of best and worst predictions, as we used a fixed top and bottom 25% to set best and worst predictions.

The repression signal that we detect is subtle. In fact, it is so subtle that we had to use statistics in order to evaluate its significance, and we had to test different sets of parameters (penalty score, fraction of highly and lowly expressed Hb-sRNAs) in order to find it. It is completely valid to question the biological significance of this repression signal: would this subtle repression for many targets make any difference for *H. bakeri in vivo*? Maybe a slight repression for many targets does make a difference, as an expression buffer effect has been proposed for some miRNAs (Guo et al. 2010).

The high sRNA complexity of HES and EV contents can hinder the assessment of Hb-sRNAs relevance. It may be the case that there are only a few Hb-sRNAs that are critical for an immuno-modulatory effect, but this signal could be diluted due to the high numbers of unique Hb-sRNAs. We detected more than 16K sequences in mouse cells and this might still be an underestimation. Finally, there may be more relevant factors for infection other than Hb-sRNAs, such as a reported TGF-β analog inducing Treg differentiation (Grainger et al. 2010), or some other protein or lipid present in HES or EVs. A relevant additional example with immunomodulation is that of Osbourn et. al. (Osbourn et al. 2017). In this study, researchers found a secreted protein present in HES, ***H.** polygyrus* **A**larmin **R**elease **I**nhibitor (HpARI), that binds to IL-33 cytokine and inhibits its binding with the IL-33 receptor. HpARI contains three Sushi domains (PFAM00084), typically involved in protein-protein interaction. However, the activity of HpARI may not easily explain our observation of IL-33 transcript decrease (**Figure 35**), so there's room for Hb-sRNAs to perform this drop of transcript level.

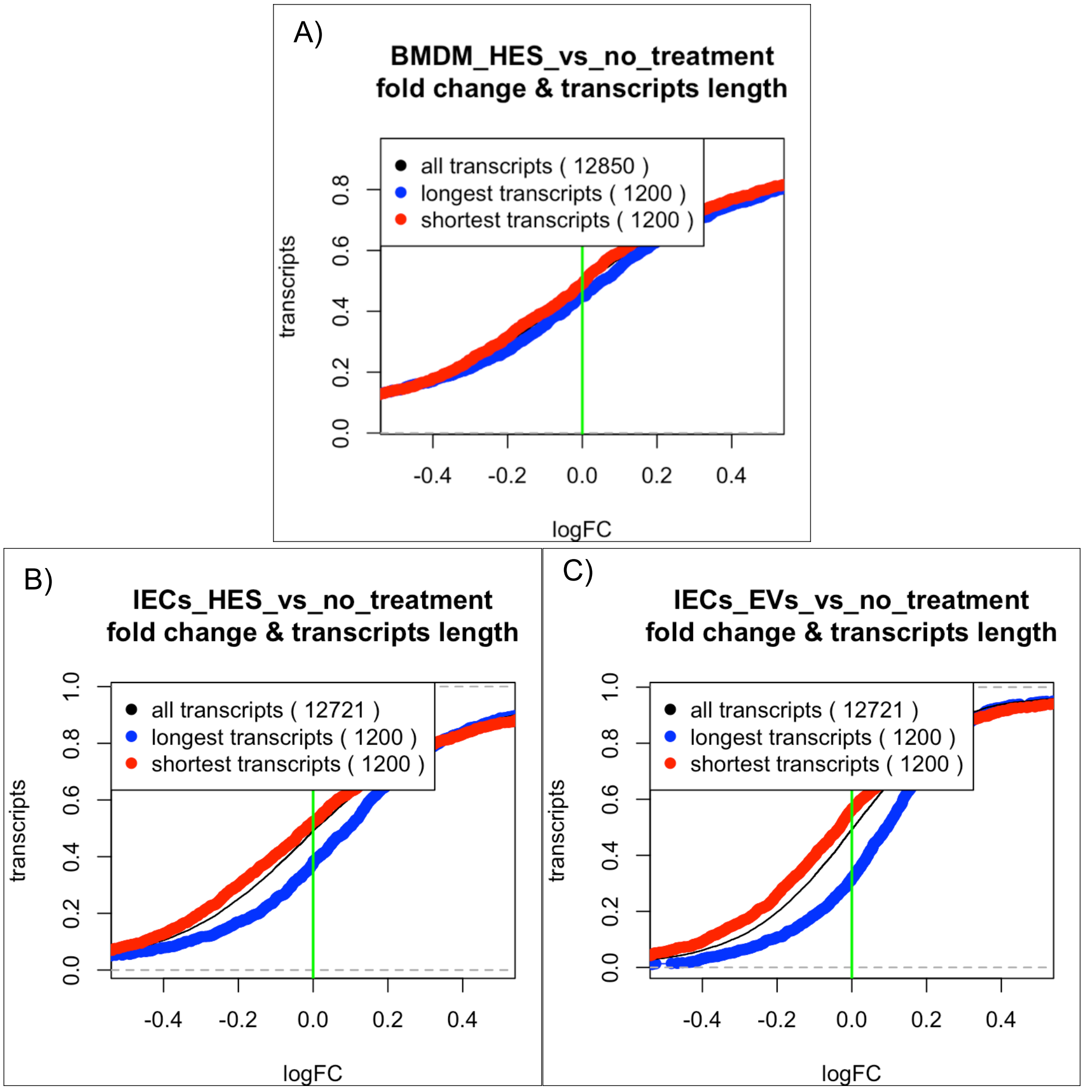


Figure 40. IECs differential expression analyzes have transcript lengths biases. Fold change comparison between 1200 longest (blue), 1200 shortest transcripts (red) and all transcripts length distributions. A) BMDM with HES. B) IECs with HES. C) IECs with vesicle (EVs).

Conclusions

HES treatment results in downregulation of proteins implicated in intracellular vesicular trafficking in macrophages such as Rab7b and RILP. The downregulation of these components may have implications such as inhibiting *H. bakeri* EVs lysosome degradation, further work is need to test these speculations.

We identified differential gene responses in macrophages and intestinal epithelial cells to HES or EVs, as was the case of IL-33 and its receptor, which were repressed by HES but not EV treatment. To the best of my knowledge, this would be the first observation of differential effects of HES and EVs into relevant genes for infection. Further work is needed in order to determine which components of HES are responsible for the downregulation of IL-33 and IL-33R transcripts.

We detected a subtle, but detectable, effect for HES and EV *H. bakeri* sRNAs in macrophages and intestinal epithelial cells. Splitting our Hb-sRNAs and target predictions into four different sets allowed us to relate the repression effects due to target prediction score and to Hb-sRNA expression. Additional controls such as Hb-sRNAs permutations and target prediction would rule out nucleotide compositional effects.

Perspectives

1. Include additional controls in our search for downregulation of Hb-sRNA targets. Two additional negative controls could be: First make Hb-sRNAs sequence permutations and perform target predictions on the transcriptome. Second, make target predictions on both strands of transcripts, the predictions made on the reverse-complement strand would still follow the trend reported in Chapter 3 of longer transcripts having more binding sites, but for most genes the reverse-complement strand will not be transcribed, or at least not to high enough levels.

2. Choose the best Hb-sRNAs candidates for individual experimental validation. A viable experimental approach would be to load exWAGO with some of our preferred Hb-sRNAs and transfect mouse cells to test for repression of their predicted targets via qPCR or RNA-Seq.

3. *In vivo* isolation from infected mice to rule out if the effect we found are exclusive of *in vitro* cell cultures.

References

Abad, Pierre, and Valerie M Williamson. 2010. *Plant Nematode Interaction : A Sophisticated Dialogue. Advances in Botanical Research-Volume 53*. First Edit. Vol. 53. Elsevier. https://doi.org/10.1016/S0065-2296(10)53005-2.

Almeida, Miguel Vasconcelos, and Miguel A Andrade-navarro. 2019. “Function and Evolution of Nematode RNAi Pathways.” *Non-Coding RNA* 5 (8): 1–24. https://doi.org/10.3390/ncrna5010008.

Andrews, S. (2010). 2010. “FastQC: A Quality Tool Control for High Throughput Sequence Data.” 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Anthony, Robert M, Joseph F Urban, Farhang Alem, Hossein A Hamed, Cristina T Rozo, Jean-luc Boucher, Nico Van Rooijen, and William C Gause. 2006. “Memory T H 2 Cells Induce Alternatively Activated Macrophages to Mediate Protection against Nematode Parasites.” *Nature Medicine* 12 (8): 955–60. https://doi.org/10.1038/nm1451.

Armisen, Javier, Michael J Gilchrist, Anna Wilczynska, Nancy Standart, and Eric A Miska. 2009. “Abundant and Dynamically Expressed MiRNAs, PiRNAs, and Other Small RNAs in the Vertebrate Xenopus Tropicalis.” *Genome Research* 10: 1766–75. https://doi.org/10.1101/gr.093054.109.1766.

Artis, D, and R K Grencis. 2008. “The Intestinal Epithelium: Sensors to Effectors in Nematode Infection.” *Mucosal Immunology* 1 (4): 252–64. https://doi.org/10.1038/mi.2008.21.

Atkin-smith, Georgia K, and Ivan K H Poon. 2016. “Disassembly of the Dying : Mechanisms and Functions.” *Trends in Cell Biology*, 1–12. https://doi.org/10.1016/j.tcb.2016.08.011.

Axtell, Michael J. 2013. “ShortStack: Comprehensive Annotation and Quantification of Small RNA Genes.” *Bioinformatics* 19 (6): 740–51. https://doi.org/10.1261/rna.035279.112.

Bansemir, Anne D, and Michael V K Sukhdeo. 2016. “The Food Resource Of Adult Heligmosomoides Polygyrus In The Small Intestine.” *The Journal of Parasitology* 80 (1): 24–28.

Batista, Pedro J, J Graham Ruby, Julie M Claycomb, Rosaria Chiang, Noah Fahlgren, Kristin D Kasschau, Daniel A Chaves, et al. 2008. “PRG-1 and 21U-RNAs Interact to Form the PiRNA Complex Required for Fertility in C . Elegans.” *Molecular Cell* 31 (1): 67–78. https://doi.org/10.1016/j.molcel.2008.06.002.

Bauer, David F. 1972. “Confidence Sets Using Rank Statistics Constructing Confidence Sets Using Rank Statistics.” *Journal of the American Statistical Association Constructing* 67 (339): 687–90. https://doi.org/10.1080/01621459.1972.10481279.

Behnke, J M, D M Menge, and H Noyes. 2009. “Heligmosomoides Bakeri : A Model for Exploring the Biology and Genetics of Resistance to Chronic Gastrointestinal Nematode Infections.” *Parasitology* 136 (12): 1565–80. https://doi.org/10.1017/S0031182009006003.

Berger, Abi. 2000. “Science Commentary : Th1 and Th2 Responses : What Are They ?” *British Medical Journal* 321 (August): 5500.

Bermúdez-Barrientos, José Roberto, Obed Ramírez-Sánchez, Franklin Wang-Ngai Chow, Amy H. Buck, and Cei Abreu-Goodger. 2020. “Disentangling SRNA-Seq Data to Study RNA Communication between Species.” *Nucleic Acids Research* 48 (4). https://doi.org/10.1093/nar/gkz1198.

Bieren, Julia Esser-von, Ilaria Mosconi, Romain Guiet, Alessandra Piersgilli, Beatrice Volpe, Fei Chen, William C Gause, Arne Seitz, J Sjef Verbeek, and Nicola L Harris. 2013. “Antibodies Trap Tissue Migrating Helminth Larvae and Prevent Tissue Damage by Driving IL-4R a -Independent Alternative Differentiation of Macrophages.” *PLoS Pathogens* 9 (11). https://doi.org/10.1371/journal.ppat.1003771.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Brown, Gordon D, Janet A Willment, and Lauren Whitehead. 2018. “C-Type Lectins in Immunity and Homeostasis.” *Nature Reviews Immunology* 18 (June). https://doi.org/10.1038/s41577-018-0004-8.

Brown, Kristen C, and Taiowa A Montgomery. 2017. “Transgenerational Inheritance: Perpetuating RNAi.” *Current Biology* 27 (10): R383–85. https://doi.org/10.1016/j.cub.2017.03.061.

Buck, Amy H., Gillian Coakley, Fabio Simbari, Henry J. McSorley, Juan F. Quintana, Thierry Le Bihan, Sujai Kumar, et al. 2014. “Exosomes Secreted by Nematode Parasites Transfer Small RNAs to Mammalian Cells and Modulate Innate Immunity.” *Nature Communications* 5: 5488. https://doi.org/10.1038/ncomms6488.

Buck, Amy H, and Mark Blaxter. 2013. “Functional Diversification of Argonautes in Nematodes: An Expanding Universe.” *Biochemical Society Transactions* 41 (4): 881–86. https://doi.org/10.1042/BST20130086.

Buckley, Bethany A, Kirk B Burkhart, Sam Guoping Gu, George Spracklin, Aaron Kershner, Heidi Fritz, Judith Kimble, Andrew Fire, and Scott Kennedy. 2012. “A Nuclear Argonaute Promotes Multigenerational Epigenetic Inheritance and Germline Immortality” 489: 447–51. https://doi.org/10.1038/nature11352.

Bushmanova, Elena, Dmitry Antipov, Alla Lapidus, and Andrey D Prjibelski. 2019. “RnaSPAdes: A de Novo Transcriptome Assembler and Its Application to RNA-Seq Data.” *GigaScience* 8: 1–13. https://doi.org/10.1093/gigascience/giz100.

Cable, J, P D Harris, J W Lewis, and J M Behnke. 2006. “Molecular Evidence That Heligmosomoides Polygyrus from Laboratory Mice and Wood Mice Are Separate Species.” *Parasitology*, no. 133:

111–22. https://doi.org/10.1017/S0031182006000047.

Cai, Qiang, Lulu Qiao, Ming Wang, Baoye He, Feng Mao Lin, Jared Palmquist, Sienna Da Huang, and Hailing Jin. 2018. "Plants Send Small RNAs in Extracellular Vesicles to Fungal Pathogen to Silence Virulence Genes." *Science* 360 (6393): 1126–29. https://doi.org/10.1126/science.aar4142.

Chatila, Talal A., F. Blaeser, N. Ho, H. M. Lederman, C. Voulgaropoulos, C. Helms, and A. M. Bowcock. 2000. "JM2, Encoding a Fork Head–Related Protein, Is Mutated in X-Linked Autoimmunity–Allergic Disregulation Syndrome." *The Journal of Clinical Investigation* 106 (12): 75–81.

Chavez-Montes, Ricardo A., Flor de Fatima Rosas-Cardenas, Emanuele De Paoli, Monica Accerbi, Blake C Meyers, Linda A Rymarquis, Gayathri Mahalingam, Nayelli Marsch-martı, Pamela J Green, and Stefan De Folter. 2014. "Sample Sequencing of Vascular Plants Demonstrates Widespread Conservation and Divergence of MicroRNAs." *Nature Communications* 5: 1–15. https://doi.org/10.1038/ncomms4722.

Chen, Gang, Alexander C Huang, Wei Zhang, Gao Zhang, Min Wu, Wei Xu, Zili Yu, et al. 2018. "Exosomal PD-L1 Contributes to Immunosuppression and Is Associated with Anti-PD-1 Response." *Nature*. https://doi.org/10.1038/s41586-018-0392-8.

Chen, Ho-ming, Li-teh Chen, Kanu Patel, Yi-hang Li, David C Baulcombe, and Shu-hsing Wu. 2010. "22-Nucleotide RNAs Trigger Secondary SiRNA Biogenesis in Plants." *Proceedings of the National Academy of Sciences* 107 (34): 15269–74. https://doi.org/10.1073/pnas.1001738107.

Chiou, Ni-ting, Robin Kageyama, K Mark Ansel, Ni-ting Chiou, Robin Kageyama, and K Mark Ansel. 2018. "Selective Export into Extracellular Vesicles and Function of TRNA Fragments during T Cell Activation Article Selective Export into Extracellular Vesicles and Function of TRNA Fragments during T Cell Activation." *Cell Reports* 25 (12): 3356-3370.e4. https://doi.org/10.1016/j.celrep.2018.11.073.

Chow, Franklin Wang-Ngai, Georgios Koutsovoulos, Cesaré Ovando-Vázquez, Kyriaki Neophytou, Dominik R Laetsch, Jose R Bermúdez-Barrientos, Sujai Kumar, et al. 2019. "Secretion of an Argonaute Protein by a Parasitic Nematode and the Evolution of Its SiRNA Guides." *Nucleic Acids Research*, 343772. https://doi.org/10.1101/343772.

Claycomb, Julie M. 2014. "Ancient Endo-SiRNA Pathways Reveal New Tricks." *Current Biology* 24 (15): R703–15. https://doi.org/10.1016/j.cub.2014.06.009.

Coakley, Gillian, Jana L. McCaskill, Jessica G. Borger, Fabio Simbari, Elaine Robertson, Marissa Millar, Yvonne Harcus, Henry J. McSorley, Rick M. Maizels, and Amy H. Buck. 2017. "Extracellular Vesicles from a Helminth Parasite Suppress Macrophage Activation and Constitute an Effective Vaccine for Protective Immunity." *Cell Reports* 19 (8): 1545–57. https://doi.org/10.1016/j.celrep.2017.05.001.

Cocucci, Emanuele, and Jacopo Meldolesi. 2015. "Ectosomes and Exosomes : Shedding the Confusion between Extracellular Vesicles." *Trends in Cell Biology* 25 (6): 364–72. https://doi.org/10.1016/j.tcb.2015.01.004.

Cocucci, Emanuele, Gabriella Racchetti, and Jacopo Meldolesi. 2009. "Shedding Microvesicles : Artefacts No More." *Trends in Cell Biology* 19 (2): 43–51. https://doi.org/10.1016/j.tcb.2008.11.003.

Coombes, Janine L, and Fiona Powrie. 2008. "Dendritic Cells in Intestinal Immune Regulation." *Nature Reviews Immunology* 8: 435–46. https://doi.org/10.1038/nri2335.

Davis, Matthew P a, Stijn van Dongen, Cei Abreu-Goodger, Nenad Bartonicek, and Anton J. Enright. 2013. "Kraken: A Set of Tools for Quality Control and Analysis of High-Throughput Sequence Data." *Methods* 63 (1): 41–49. https://doi.org/10.1016/j.ymeth.2013.06.027.

Delorme-axford, Elizabeth, Rogier B Donker, Jean-francois Mouillet, Tianjiao Chu, and Avraham Bayer. 2013. "Human Placental Trophoblasts Confer Viral Resistance to Recipient Cells." *Proceedings of the National Academy of Sciences* 110 (29): 12048–53. https://doi.org/10.1073/pnas.1304718110.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Durinck, Steffen, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. 2005. “BioMart and Bioconductor : A Powerful Link between Biological Databases and Microarray Data Analysis.” *Bioinformatics* 21 (16): 3439–40. https://doi.org/10.1093/bioinformatics/bti525.

Eichhorn, Stephen W, Huili Guo, Sean E Mcgeary, Ricard A Rodriguez-mias, and Chanseok Shin. 2010. “MRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues.” *Molecular Cell* 56: 104–15. https://doi.org/10.1016/j.molcel.2014.08.028.

Entwistle, Lewis J., Victoria S. Pelly, Stephanie M. Coomes, Yashaswini Kannan, Jimena Perez-Lloret, Stephanie Czieso, Mariana Silva dos Santos, et al. 2017. “Epithelial-Cell-Derived Phospholipase A2 Group 1B Is an Endogenous Anthelmintic.” *Cell Host and Microbe* 22 (4): 484-493.e5. https://doi.org/10.1016/j.chom.2017.09.006.

Fahlgren, Noah, and James C Carrington. n.d. *Plant MicroRNAs: Methods and Protocols. Chapter 4 MiRNA Target Prediction in Plants*. Edited by Blake C Meyers and Pamela J Green. Vol. 592. Springer US. https://doi.org/10.1007/978-1-60327-005-2.

Faulkner, Geoffrey J, Alistair R R Forrest, Alistair M Chalk, Kate Schroder, Yoshihide Hayashizaki, Piero Carninci, David A Hume, and Sean M Grimmond. 2008. “A Rescue Strategy for Multimapping Short Sequence Tags Refines Surveys of Transcriptional Activity by CAGE.” *Genomics* 91: 281–88. https://doi.org/10.1016/j.ygeno.2007.11.003.

Fei, Qili, Rui Xia, and Blake C Meyers. 2013. “Phased, Secondary, Small Interfering RNAs in Posttranscriptional Regulatory Networks.” *The Plant Cell* 25: 2400–2415. https://doi.org/10.1105/tpc.113.114652.

Filbey, Kara J, John R Grainger, Katherine A Smith, Louis Boon, Nico Van Rooijen, Yvonne Harcus, Stephen Jenkins, James P Hewitson, and Rick M Maizels. 2014. “Innate and Adaptive Type 2 Immune Cell Responses in Genetically Controlled Resistance to Intestinal Helminth Infection.” *Immunology and Cell Biology*, no. November 2013: 436–48. https://doi.org/10.1038/icb.2013.109.

Fire, A, S Xu, M K Montgomery, S A Kostas, S E Driver, and C C Mello. 1998. “Potent and Specific Genetic Interference by Double-Stranded RNA in Caenorhabditis Elegans.” *Nature* 391 (6669): 806–11. https://doi.org/10.1038/35888.

Fofanov, Yuriy, Yi Luo, Charles Katili, and Jim Wang. 2004. “How Independent Are the Appearances of n -Mers in Different Genomes ?” *Bioinformatics* 20 (15): 2421–28. https://doi.org/10.1093/bioinformatics/bth266.

G Raposo, H W Nijman, W Stoorvogel, R Liejendekker, C V Harding, C J Melief, and H J Geuze. 1996. “B Lymphocytes Secrete Antigen-PresentingVesicles.” *The Journal of Experimental Medicine* 183 (March): 1161–72. https://doi.org/10.1084/jem.183.3.1161.

Gent, Jonathan I, Ayelet T Lamm, Derek M Pavelec, Jay M Maniar, Poornima Parameswaran, Li Tao, Scott Kennedy, and Andrew Z Fire. 2010. “Distinct Phases of SiRNA Synthesis in an Endogenous RNAi Pathway in C . Elegans Soma.” *Molecular Cell* 37 (5): 679–89. https://doi.org/10.1016/j.molcel.2010.01.012.

Gerbe, François, Emmanuelle Sidot, Danielle J Smyth, Makoto Ohmoto, Ichiro Matsumoto, Valérie Dardalhon, Pierre Cesses, et al. 2016. “Intestinal Epithelial Tuft Cells Initiate Type 2 Mucosal Immunity to Helminth Parasites.” *Nature* 529 (7585): 226–30. https://doi.org/10.1038/nature16527.

Gordon, Siamon. 2003. “Alternative Activation of Macrophages.” *Nature Reviews Immunology* 3 (January): 23–35. https://doi.org/10.1038/nri978.

Gould, Stephen Jay; Vrba, Elisabeth S. 1982. “Paleontological Society Exaptation-A Missing Term in the Science of Form Exaptation-a Missing Term in the Science of Form.” *Paleobiology* 8 (1): 4–15.

Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn a Thompson, Ido Amit, Xian Adiconis, et al. 2011. “Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome.” *Nature Biotechnology* 29 (7): 644–52. https://doi.org/10.1038/nbt.1883.

Grainger, John R., Katie A. Smith, James P. Hewitson, Henry J. McSorley, Yvonne Harcus, Kara J. Filbey, Constance A.M. Finney, et al. 2010. “Helminth Secretions Induce de Novo T Cell Foxp3

Expression and Regulatory Function through the TGF-β Pathway.” *The Journal of Experimental Medicine* 207 (11): 2331–41. https://doi.org/10.1084/jem.20101074.

Guerra, Flora, and Cecilia Bucci. 2016. “Multiple Roles of the Small GTPase Rab7.” https://doi.org/10.3390/cells5030034.

Guo, Huili, Nicholas T Ingolia, Jonathan S Weissman, and David P Bartel. 2010. “Mammalian MicroRNAs Predominantly Act to Decrease Target MRNA Levels.” *Nature* 466 (7308): 835–40. https://doi.org/10.1038/nature09267.

Haber, Adam L., Moshe Biton, Noga Rogel, Rebecca H. Herbst, Karthik Shekhar, Christopher Smillie, Grace Burgin, et al. 2017. “A Single-Cell Survey of the Small Intestinal Epithelium.” *Nature* 551 (7680): 333–39. https://doi.org/10.1038/nature24489.

Hamilton, Andrew J, and David C Baulcombe. 1999. “A Species of Small Antisense RNA in Posttranscriptional Gene Silencing in Plants.” *Science* 286 (5441): 950–52.

Hardcastle, Thomas J, Krystyna A Kelly, and David C Baulcombe. 2012. “Identifying Small Interfering RNA Loci from High-Throughput Sequencing Data.” *Bioinformatics* 28 (4): 457–63. https://doi.org/10.1093/bioinformatics/btr687.

Havaux, X., A. Zeine, A. Dits, and O. Denis. 2005. “A New Mouse Model of Lung Allergy Induced by the Spores of Alternaria Alternata and Cladosporium Herbarum Molds.” *Clinical Experimental Immunology* 139 (2): 179–88.

He, Lin, and Gregory J Hannon. 2004. “MicroRNAs : SMALL RNAs WITH A BIG ROLE IN GENE REGULATION.” *Nature Reviews Genetics* 5: 522–31. https://doi.org/10.1038/nrg1379.

Herbert, De Broski R, Jun-qi Yang, Simon P Hogan, Kathryn Groschwitz, Marat Khodoun, Ariel Munitz, Tatyana Orekov, et al. 2009. “Intestinal Epithelial Cell Secretion of RELM-Beta Protects against Gastrointestinal Worm Infection.” *Journal of Experimental Medicine* 206 (13): 2947–57. https://doi.org/10.1084/jem.20091268.

Hewitson, James P, Al C Ivens, Yvonne Harcus, Kara J Filbey, Henry J Mcsorley, Janice Murray, Stephen Bridgett, David Ashford, Adam A Dowle, and Rick M Maizels. 2013. “Secretion of Protective Antigens by Tissue-Stage Nematode Larvae Revealed by Proteomic Analysis and Vaccination-Induced Sterile Immunity.” *PLoS Pathogens* 9 (8): e100349. https://doi.org/10.1371/journal.ppat.1003492.

Ivashuta, Sergey, Yuanji Zhang, B Elizabeth Wiggins, Partha Ramaseshadri, Gerrit C. Segers, Steven Johnson, Steve E. Meyer, et al. 2015. “Environmental RNAi in Herbivorous Insects.” *Rna* 5 (21): 1–11. https://doi.org/10.1261/rna.048116.114.2.

Jia, Dong, Lun Cai, Housheng He, Geir Skogerbø, Tiantian Li, Muhammad Nauman Aftab, and Runsheng Chen. 2007. “Systematic Identification of Non-Coding RNA 2 , 2 , 7-Trimethylguanosine Cap Structures in Caenorhabditis Elegans.” *BMC Molecular Biology* 9: 1–9. https://doi.org/10.1186/1471-2199-8-86.

Johansson, Malin E. V., and Gunnar C. Hansson. 2016. “The Mucins.” In *Encyclopedia of Immunology*, 381–88. Elsevier.

Johnson, Nathan R., Claude W. de Pamphilis, and Michael J. Axtell. 2019. “Compensatory Sequence Variation between Trans-Species Small Rnas and Their Target Sites.” *ELife* 8: 1–17. https://doi.org/10.7554/eLife.49750.

Johnston, Chris J C, Elaine Robertson, Yvonne Harcus, John R Grainger, Gillian Coakley, Danielle J Smyth, Henry J Mcsorley, and Rick Maizels. 2015. “Cultivation of Heligmosomoides Polygyrus : An Immunomodulatory Nematode Parasite and Its Secreted Products.” *Journal of Visualized Experiments: JoVE*, no. April: 1–10. https://doi.org/10.3791/52412.

Kaiser, Bettina, Gerd Vogg, Ursula B Fürst, Markus Albert, and James H Westwood. 2015. “Parasitic Plants of the Genus Cuscuta and Their Interaction with Susceptible and Resistant Host Plants.” *Frontiers in Plant Science* 6: 1–9. https://doi.org/10.3389/fpls.2015.00045.

Kashif, M, S Pietilä, K Artola, Agricultural Sciences, and R A C Jones. 2012. “Detection of Viruses in Sweetpotato from Honduras and Guatemala Augmented by Deep-Sequencing of Small-RNAs.” *Plant Disease* 96 (10): 1430–37.

Kim, John K, John K Kim, Harrison W Gabel, Ravi S Kamath, Joshua M Kaplan, Marc Vidal, and Gary Ruvkun. 2005. “Functional Genomic Analysis of RNA Interference in C . Elegans.” *Science* 308: 1164–67. https://doi.org/10.1126/science.1109267.

Kindt, Thomas J, Richard A Goldsby, Barbara Anne Osborne, and Janis Kuby. 2007. *Kuby Immunology*. Sixth edit. New York: W.H. Freeman, ©2007.

Kiontke, Karin; Fitch, David H. A. 2005. “The Phylogenetic Relationships of Caenorhabditis and Other Rhabditids.” WormBook, Ed. The C. Elegans Research Community, WormBook. 2005. https://doi.org/doi/10.1895/wormbook.1.11.1.

Kolaczkowska, Elzbieta, and Paul Kubes. 2013. “Neutrophil Recruitment and Function.” *Nature Reviews Immunology* 13: 159–75. https://doi.org/10.1038/nri3399.

Kowal, Joanna, Mercedes Tkach, and Clotilde Thery. 2014. “Biogenesis and Secretion of Exosomes.” *Current Opinion in Cell Biology* 29 (1): 116–25. https://doi.org/10.1016/j.ceb.2014.05.004.

Kowalski, Madzia P, and Torsten Krude. 2015. “Functional Roles of Non-Coding Y RNAs.” *The International Journal of Biochemistry & Cell Biology* 66: 20–29. https://doi.org/10.1016/j.biocel.2015.07.003.

Koyasu, Shigeo, Richard M Locksley, Andrew N J Mckenzie, Reina E Mebius, and Fiona Powrie. 2018. “Review Innate Lymphoid Cells : 10 Years On.” *Cell* 174 (5): 1054–66. https://doi.org/10.1016/j.cell.2018.07.017.

Lambert, Marine, Abderrahim Benmoussa, and Patrick Provost. 2019. “Small Non-Coding RNAs Derived from Eukaryotic Ribosomal RNA.” *Non-Coding RNA* 5 (16): 1–19. https://doi.org/10.3390/ncrna5010016.

Langmead, B, C Trapnell, M Pop, and SL Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biol.* 10 (3): R25. https://doi.org/10.1186/gb-2009-10-3-r25.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lawrence, Michael, Wolfgang Huber, Herve Pages, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. 2013. “Software for Computing and Annotating Genomic Ranges.” *PLoS Computational Biology* 9 (8): 1–10. https://doi.org/10.1371/journal.pcbi.1003118.

Levenshtein. 1966. “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals.” *Soviet Physics Doklady* 10 (8): 707–10.

Li, Bo, Victor Ruotti, Ron M Stewart, James A Thomson, and Colin N Dewey. 2010. “RNA-Seq Gene Expression Estimation with Read Mapping Uncertainty.” *Bioinformatics* 26 (4): 493–500. https://doi.org/10.1093/bioinformatics/btp692.

Li, Yang, Chaoqun Li, Guohui Ding, and Youxin Jin. 2011. “Evolution of MIR159 / 319 MicroRNA Genes and Their Post-Transcriptional Regulatory Link to SiRNA Pathways.” *BMC Evolutionary Biology* 11 (122): 1–18.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

Maizels, Rick M., James P. Hewitson, Janice Murray, Yvonne M. Harcus, Blaise Dayer, Kara J. Filbey, John R. Grainger, Henry J. McSorley, Lisa A. Reynolds, and Katherine A. Smith. 2012. “Immune Modulation and Modulators in Heligmosomoides Polygyrus Infection.” *Experimental Parasitology* 132 (1): 76–89. https://doi.org/10.1016/j.exppara.2011.08.011.

Marçais, Guillaume, and Carl Kingsford. 2011. “A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of k-Mers.” *Bioinformatics* 27 (6): 764–70. https://doi.org/10.1093/bioinformatics/btr011.

McCarthy, Davis J, Yunshun Chen, and Gordon K Smyth. 2012. “Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation.” *Nucleic Acids Research* 40 (10): 4288–97. https://doi.org/10.1093/nar/gks042.

Melo, Sonia A, Linda B Luecke, Christoph Kahlert, Agustin F Fernandez, Seth T Gammon, Judith Kaye, Valerie S Lebleu, et al. 2015. “Glypican-1 Identifies Cancer Exosomes and Detects Early Pancreatic Cancer.” *Nature* 523: 177–82. https://doi.org/10.1038/nature14581.

Millar, Anthony A, Allan Lohe, and Gigi Wong. 2019. “Biology and Function of MiR159 in Plants.”

Möller, Mareike, and Eva H Stukenbrock. 2017. “Evolution and Genome Architecture in Fungal Plant Pathogens.” *Nature Publishing Group* 15 (12): 756–71. https://doi.org/10.1038/nrmicro.2017.76.

Mortazavi, Ali, Brian A Williams, Kenneth Mccue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

Munitz, Ariel, Amanda Waddell, Luqman Seidu, Eric T Cole, Richard Ahrens, Simon P Hogan, and Marc E Rothenberg. 2008. “Resistin-like Molecule a Enhances Myeloid Cell Activation and Promotes Colitis.” *Journal of Allergy and Clinical Immunology* 122 (6): 1200–1208. https://doi.org/10.1016/j.jaci.2008.10.017.

Nadjsombati, Marija S, John W Mcginty, Miranda R Lyons-cohen, Richard M Locksley, Daniel Raftery, Jakob Von Moltke, Marija S Nadjsombati, et al. 2018. “Detection of Succinate by Intestinal Tuft Cells Triggers a Type 2 Innate Immune Circuit.” *Immunity* 49 (1): 33-41.e7. https://doi.org/10.1016/j.immuni.2018.06.016.

Nair, Meera G, Yurong Du, Jacqueline G Perrigoue, Colby Zaph, Justin J Taylor, Michael Goldschmidt, Gary P Swain, et al. 2009. “Alternatively Activated Macrophage-Derived RELM-Alpha Is a Negative Regulator of Type 2 Inflammation in the Lung.” *The Journal of Experimental Medicine* 206 (4): 937–52. https://doi.org/10.1084/jem.20082048.

Napoli, Carolyn, Christine Lemieux, and Richard Jorgensen. 1990. “Lntroduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes Ín Trans.” *The Plant Cell* 2 (April): 279–89.

Neill, Luke A J O, Douglas Golenbock, and Andrew G Bowie. 2013. “The History of Toll-like Receptors: Redefining Innate Immunity.” *Nature Reviews Immunology* 13: 453–60. https://doi.org/10.1038/nri3446.

Nicolas, Francisco E, Santiago Torres-Martinez, and Rosa M Ruiz-Vazquez. 2013. “Loss and Retention of RNA Interference in Fungi and Parasites.” *PLoS Pathogens* 9 (1): 1–4. https://doi.org/10.1371/journal.ppat.1003089.

Nunes, Cristiano C., and Ralph A. Dean. 2012. “Host-Induced Gene Silencing: A Tool for Understanding Fungal Host Interaction and for Developing Novel Disease Control Strategies.” *Molecular Plant Pathology* 13 (5): 519–29. https://doi.org/10.1111/j.1364-3703.2011.00766.x.

Obbard, Darren J, Karl H J Gordon, Amy H Buck, and Francis M Jiggins. 2009. “The Evolution of RNAi as a Defence against Viruses and Transposable Elements.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, no. 364: 99–115. https://doi.org/10.1098/rstb.2008.0168.

Osbourn, Megan, Dinesh C Soares, Francesco Vacca, Alasdair C Ivens, Rick M Maizels, Henry J Mcsorley, Megan Osbourn, et al. 2017. “HpARI Protein Secreted by a Helminth Parasite Suppresses Interleukin-33.” *Immunity* 47 (4): 739–51. https://doi.org/10.1016/j.immuni.2017.09.015.

Owhashi, Makoto, Hiroyuki Arita, and Naoko Hayai. 2000. “Identification of a Novel Eosinophil Chemotactic Cytokine (ECF-L) as a Chitinase Family Protein.” *The Journal of Biological Chemistry* 275 (2): 1279–86.

Paily, K. P., S. L. Hoti, and K. P Das. 2009. “A Review of the Complexity of Biology of Lymphatic Filarial Parasites.” *Journal of Parasitic Diseases* 33: 3–12.

Pan, Bin-tao, and Rose M Johnstone. 1983. “Fate of the Transferrin Receptor during Maturation of Sheep Reticulocytes In Vitro : Selective Externalization of the Receptor.” *Cell* 33 (July): 967–77.

Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19. https://doi.org/10.1038/nmeth.4197.

Pegtel, D Michiel, and Stephen J Gould. 2019. “Exosomes.” *Annual Review of Biochemistry* 88: 487–

514.

Pertea, Mihaela, Geo M Pertea, Corina M Antonescu, Tsung-cheng Chang, Joshua T Mendell, and Steven L Salzberg. 2015. “StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads.” *Nature Biotechnology* 33 (3): 290–95. https://doi.org/10.1038/nbt.3122.

Phipson, B., Stanley Lee, Ian J Majewski, Warren S. Alexander, and Gordon K. Smyth. 2016. “Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression.” *Annals of Applied Statistics* 10 (2): 946–63.

Quintana, Juan F, Sujai Kumar, Alasdair Ivens, Franklin W N Chow, M Hoy, Alison Fulton, Paul Dickinson Id, et al. 2019. “Comparative Analysis of Small RNAs Released by the Filarial Nematode Litomosoides Sigmodontis in Vitro and in Vivo.” *PLoS Neglected Tropical Diseases* 13 (11): 1–29.

Quintana, Juan F, Benjamin L Makepeace, Simon A Babayan, Alasdair Ivens, Kenneth M Pfarr, Mark Blaxter, Alexander Debrah, et al. 2015. “Extracellular Onchocerca -Derived Small RNAs in Host Nodules and Blood.” *Parasites & Vectors* 8 (58): 1–11. https://doi.org/10.1186/s13071-015-0656-1.

Register, James C;, and Roger N Beachy. 1988. “Resistance to TMV in Transgenic Plants Results Frdm Interference with an Early Event in Infection.” *Virology* 166: 524–32.

Ren, Bo, Xutong Wang, Jingbo Duan, and Jianxin Ma. 2019. “Rhizobial TRNA-Derived Small RNAs Are Signal Molecules Regulating Plant Nodulation.” *Science* 8907 (July): eaav8907. https://doi.org/10.1126/science.aav8907.

Reynolds, Lisa A., Kara J. Filbey, and Rick M. Maizels. 2012. “Immunity to the Model Intestinal Helminth Parasite Heligmosomoides Polygyrus.” *Seminars in Immunopathology* 34 (6): 829–46. https://doi.org/10.1007/s00281-012-0347-3.

Robertson, Gordon, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D Jackman, Karen Mungall, et al. 2010. “De Novo Assembly and Analysis of RNA-Seq Data” 7 (11). https://doi.org/10.1038/nmeth.1517.

Robinson, James T, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26. https://doi.org/10.1038/nbt0111-24.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2009. “EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Robinson, Mark D, and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” *Genome Biology* 11 (3): R25. https://doi.org/10.1186/gb-2010-11-3-r25.

Rodríguez-Vivas, Roger Ivan; Grisi, Laerte; Perez de León, Adalberto Angel; 2017. “Potential Economic Impact Assessment for Cattle Parasites in Mexico.” *Revista Mexicana de Ciencias Pecuarias* 8 (1): 61–74.

Sarkies, Peter, and Eric A Miska. 2013. “Is There Social RNA?” *Science* 341 (6145): 467–68. https://doi.org/10.1126/science.1243175.

Schmidt-Weber, C.B. 2008. “Th17 and Treg Cells Innovate the Th1 / Th2 Concept and Allergy Research.” *Chemical Immunology and Allergy* 94: 1–7.

Schulz, Marcel H, Daniel R Zerbino, Martin Vingron, and Ewan Birney. 2012. “Oases: Robust de Novo RNA-Seq Assembly across the Dynamic Range of Expression Levels.” *Bioinformatics* 28 (8): 1086–92. https://doi.org/10.1093/bioinformatics/bts094.

Scott, Ian C, Jayesh B Majithiya, Caroline Sanden, Peter Thornton, Philip N Sanders, Tom Moore, Molly Guscott, Dominic J Corkill, Jonas S Erjefält, and E Suzanne Cohen. 2018. “Interleukin-33 Is Activated by Allergen- and Necrosis-Associated Proteolytic Activities to Regulate Its Alarmin Activity during Epithelial Damage.” *Scientific Reports* 8 (3363). https://doi.org/10.1038/s41598-018-21589-2.

Segura, Mariela, Zhong Su, Ciriaco Piccirillo, and Mary M Stevenson. 2007. “Impairment of Dendritic Cell Function by Excretory-Secretory Products : A Potential Mechanism for Nematode-Induced Immunosuppression.” *European Journal of Immunology* 37 (7): 1887–1904.

https://doi.org/10.1002/eji.200636553.

Sela, Michael. 1998. “Antigens.” In *Encyclopedia of Immunology (Second Edition)*, 201–7.

Sela, Noa, Neta Luria, and Aviv Dombrovsky. 2012. “Genome Assembly of Bell Pepper Endornavirus from Small RNA.” *Journal of Virology* 86 (14): 7721. https://doi.org/10.1128/JVI.00983-12.

Shabalina, Svetlana A., and Eugene V. Koonin. 2008. “Origins and Evolution of Eukaryotic RNA Interference.” *Trends in Ecology and Evolution* 23 (10): 578–87. https://doi.org/10.1016/j.tree.2008.06.005.

Shahid, Saima, and Michael J Axtell. 2014. “Identification and Annotation of Small RNA Genes Using ShortStack ShortStack.” *Methods* 67 (1): 20–27. https://doi.org/10.1016/j.ymeth.2013.10.004.

Shahid, Saima, Gunjune Kim, Nathan R. Johnson, Eric Wafula, Feng Wang, Ceyda Coruh, Vivian Bernal-Galeano, et al. 2018. “MicroRNAs from the Parasitic Plant Cuscuta Campestris Target Host Messenger RNAs.” *Nature* 553 (7686): 82–85. https://doi.org/10.1038/nature25027.

Sijen, Titia, Florian A Steiner, Karen L Thijssen, and Ronald H A Plasterk. 2007. “Secondary SiRNAs Result from Unprimed RNA Synthesis and Form a Distinct Class.” *Science (New York, N.Y.)* 315 (5809): 244–47. https://doi.org/10.1126/science.1136699.

Son, Dong Ju, Sandeep Kumar, Wakako Takabe, Chan Woo Kim, Chih-wen Ni, Noah Alberts-grill, In-hwan Jang, et al. 2013. “The Atypical Mechanosensitive MicroRNA-712 Derived from Pre-Ribosomal RNA Induces Endothelial Inflammation and Atherosclerosis.” *Nature Communications* 4 (3000). https://doi.org/10.1038/ncomms4000.

Soneson, Charlotte, Michael I Love, Mark D Robinson, and Rob Patro. 2015. “Differential Analyses for RNA-Seq : Transcript-Level Estimates Improve Gene-Level Inferences.” *F1000 Research* 4 (1521): 1–18.

Soto-suárez, Mauricio, Patricia Baldrich, Detlef Weigel, and Ignacio Rubio-somoza. 2017. “The Arabidopsis MiR396 Mediates Pathogen-Associated Molecular Pattern-Triggered Immune Responses against Fungal Pathogens.” *Nature Publishing Group*, no. November 2016: 1–14. https://doi.org/10.1038/srep44898.

Srivastava, Prashant K, Taraka Ramji Moturu, Priyanka Pandey, Ian T Baldwin, and Shree P Pandey. 2014. “A Comparison of Performance of Plant MiRNA Target Prediction Tools and the Characterization of Features for Genome-Wide Target Prediction.” *BMC Genomics* 15 (348): 1–15.

Stepek, Gillian, David J Buttle, Ian R Duce, and Jerzy M Behnke. 2006. “Human Gastrointestinal Nematode Infections : Are New Control Methods Required ?” *International Journal of Experimental Pathology* 87 (5): 325–41. https://doi.org/10.1111/j.1365-2613.2006.00495.x.

Stocks, Matthew B, Irina Mohorianu, Matthew Beckers, Claudia Paicu, Simon Moxon, Joshua Thody, Tamas Dalmay, and Vincent Moulton. 2018. “Sequence Analysis The UEA SRNA Workbench (Version 4.4): A Comprehensive Suite of Tools for Analyzing MiRNAs and SRNAs.” *Bioinformatics* 34 (19): 3382–84. https://doi.org/10.1093/bioinformatics/bty338.

Sutherland, T. E., Dominik Ruckerl, N Logan, S Duncan, T. A. Wynn, and Judith E Allen. 2018. “Ym1 Induces RELMα and Rescues IL-4Rα Deficiency in Lung Repair during Nematode Infection.” *PLoS Pathogens* 14 (11): e1007423.

Talavera, Maria M, Sushma Nuthakki, Hongmei Cui, Yi Jin, and Yusen Liu. 2017. “Immunostimulated Arginase II Expression in Intestinal Epithelial Cells Reduces Nitric Oxide Production and Apoptosis.” *Frontiers in Cell and Developmental Biology* 5 (15): 1–10. https://doi.org/10.3389/fcell.2017.00015.

Théry, Clotilde, Laurence Zitvogel, Sebastian Amigorena, and Institut Gustave Roussy. 2002. “Exosomes: Composition, Biogenesis And Function.” *Nature Reviews Immunology* 2 (August): 569–79. https://doi.org/10.1038/nri855.

Tomoyasu, Yoshinori, Sherry C Miller, Shuichiro Tomita, Michael Schoppmeier, Daniela Grossmann, and Gregor Bucher. 2008. “Exploring Systemic RNA Interference in Insects : A Genome-Wide Survey for RNAi Genes in Tribolium.” *Genome Biology* 9 (1): 1–22. https://doi.org/10.1186/gb-2008-9-1-r10.

Tops, B B J, Ronald H A Plasterk, and R F Keiting. 2006. “The Caenorhabditis Elegans Argonautes

ALG-1 and ALG-2 : Almost Identical yet Different." *Cold Spring Harbor Symposia on Quantitative Biology* 71: 189–94.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. "Letters Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation." *Nature Biotechnology* 28 (5). https://doi.org/10.1038/nbt.1621.

Valadi, Hadi, Karin Ekström, Apostolos Bossios, Margareta Sjöstrand, James J. Lee, and Jan O. Lötvall. 2007. "Exosome-Mediated Transfer of MRNAs and MicroRNAs Is a Novel Mechanism of Genetic Exchange between Cells." *Nature Cell Biology* 2 1 (6): 654–59. https://doi.org/10.1038/ncb1596.

Vidal, Karine, Isabelle Grosjean, Jean-Pierre Revillard, Christian Gespach, and Dominique Kaiserlian. 1993. "Immortalization of Mouse Intestinal Epithelial Cells by the SV40-Large T Gene." *Journal of Immunological Methods* 166 (1): 63–73. https://doi.org/10.1016/0022-1759(93)90329-6.

Villalobos-Escobedo, José Manuel, Nohemí Carreras-Villaseñor, and Alfredo Herrera-Estrella. 2016. "The Interaction of Fungi with the Environment Orchestrated by RNAi." *Mycologia* 108 (3): 15-246-. https://doi.org/10.3852/15-246.

Vivier, Eric, Laurence Zitvogel, Lewis L Lanier, Wayne M Yokoyama, and Sophie Ugolini. 2011. "Innate or Adaptive Immunity? The Example of Natural Killer Cells." *Science* 331 (6013): 44–49. https://doi.org/10.1126/science.1198687.

Voehringer, David. 2013. "Protective and Pathological Roles of Mast Cells and Basophils." *Nature Reviews Immunology* 13: 362–75. https://doi.org/10.1038/nri3427.

Wang, Daniel Y, Sudhir Kumar, and S Blair Hedges. 1999. "Divergence Time Estimates for the Early History of Animal Phyla and the Origin of Plants, Animals and Fungi." *Proceedings of the Royal Society: Biological Sciences* 266 (1415): 163–71.

Waterhouse, Andrew M, James B Procter, David M A Martin, Michèle Clamp, and Geoffrey J Barton. 2009. "Jalview Version 2 — a Multiple Sequence Alignment Editor and Analysis Workbench." *Bioinformatics* 25 (9): 1189–91. https://doi.org/10.1093/bioinformatics/btp033.

Weiberg, Arne, Marschal Bellinger, and Hailing Jin. 2015. "Conversations between Kingdoms: Small RNAs." *Current Opinion in Biotechnology* 32: 207–15. https://doi.org/10.1016/j.copbio.2014.12.025.

Weiberg, Arne, Ming Wang, Feng-mao Lin, Hongwei Zhao, Zhihong Zhang, Isgouhi Kaloshian, Hsien-Da Huang, and Hailing Jin. 2013. "Fungal Small RNAs Suppress Plant Immunity by Hijacking Host RNA Interference Pathways." *Science (New York, N.Y.)* 342 (6154): 118–23. https://doi.org/10.1126/science.1239705.

Winston, Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2020. "Shiny: Web Application Framework for R. R Package Version 1.4.0.2." 2020. https://cran.r-project.org/package=shiny.

Winston, William M, Christina Molodowitch, and Craig P Hunter. 2002. "Systemic RNAi in C. Elegans Requires the Putative Transmembrane Protein SID-1." *Science (New York, N.Y.)* 295 (5564): 2456–59. https://doi.org/10.1126/science.1068836.

Winston, William M, Marie Sutherlin, Amanda J Wright, Evan H Feinberg, and Craig P Hunter. 2007. "Caenorhabditis Elegans SID-2 Is Required for Environmental RNA Interference." *Proceedings of the National Academy of Sciences of the United States of America* 104 (25): 10565–70. https://doi.org/10.1073/pnas.0611282104.

Witwer, Kenneth W, and Clotilde Théry. 2019. "Extracellular Vesicles or Exosomes? On Primacy, Precision, and Popularity Influencing a Choice of Nomenclature." *Journal of Extracellular Vesicles* 8 (1). https://doi.org/10.1080/20013078.2019.1648167.

Xie, Yinlong, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, et al. 2014. "Sequence Analysis SOAPdenovo-Trans : De Novo Transcriptome Assembly with Short RNA-Seq Reads." *Bioinformatics* 30 (12): 1660–66. https://doi.org/10.1093/bioinformatics/btu077.

Xie, Zhixin, Edwards Allen, April Wilken, and James C Carrington. 2005. “DICER-LIKE 4 Functions in Trans-Acting Small Interfering RNA Biogenesis and Vegetative Phase Change in Arabidopsis Thaliana.” *Proceedings of the National Academy of Sciences* 102 (36): 12984–1289.

Yates, Andrew D, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-jarreta, M Ridwan Amode, et al. 2020. “Ensembl 2020.” *Nucleic Acids Research* 48 (November 2019): 682–88. https://doi.org/10.1093/nar/gkz966.

Youngman, Elaine M., and Julie M. Claycomb. 2014. “From Early Lessons to New Frontiers: The Worm as a Treasure Trove of Small RNA Biology.” *Frontiers in Genetics* 5 (NOV): 1–13. https://doi.org/10.3389/fgene.2014.00416.

Yuana, Yuana, Auguste Sturk, and Rienk Nieuwland. 2013. “Extracellular Vesicles in Physiological and Pathological Conditions.” *Blood Reviews* 27 (1): 31–39. https://doi.org/10.1016/j.blre.2012.12.002.

PhD paper contributions

A. Rougon-Cardoso et al., “The genome, transcriptome, and proteome of the nematode Steinernema carpocapsae: evolutionary signatures of a pathogenic lifestyle,” Sci. Rep., vol. 6, no. October, p. 37536, 2016.

J. R. Bermúdez-Barrientos, O. Ramírez-Sánchez, F. W.-N. Chow, A. H. Buck, and C. Abreu-Goodger, “Disentangling sRNA-Seq data to study RNA communication between species,” Nucleic Acids Res., vol. 48, no. 4, 2020.

A. N. Espino-Vázquez et al., “Narnaviruses: novel players in fungal – bacterial symbioses,” ISME J., vol. 14, pp. 1743–1754, 2020.

José Roberto Bermúdez Barrientos

Born in August 18th, 1989. Mexican. Civil Status: Single

Corredores 120# col. Punto Verde, Leon, Gto. Mexico 37298
T: +52 477 1129137 E: roberto89bermudez@gmail.com

Bioinformatics Skills

Eight years of experience working in Bioinformatic projects. Confortable in Unix operative systems and the command line interface. Coding in Perl and R languages. Accustomed to handle high throughput data. Experience using cluster computing. Analyzed Illumina RNA-Seq data to identify gene expression and discover differentially expressed genes. Expertise working with sRNA-Seq data of multiple organisms. Familiarized with tools such as fastQC, STAR aligner, samtools, HTSeq and R bioconductor.

Working Experience

Research Assistant at RNA Computational Genomics Lab [2012]. Under leadership of PhD Cei Abreu-Goodger. National Laboratory of Genomics for Biodiversity (LANGEBIO), CINVESTAV Irapuato, Mexico.

Teaching Experience

An introduction of RNA-Seq data analysis BSc level [2014].
Eight hours course with both theory and hands on practices. Given at ENES Leon, Mexico. From the National Autonomous University of Mexico. Part of the Bioinformatics subject.

An introduction of RNA-Seq data analysis MSc level [2018-2020].
A twelve hours semester-wise course with both theory and hands on practices. Given at Centro de Biotecnología Genómica, from the Instituto Politécnico Nacional. Part of the Bioinformatics subject.

Formation

PhD in Integrative Biology [2020], at LANGEBIO CINVESTAV, Irapuato, Mexico

MSc in Integrative Biology [2016], at LANGEBIO CINVESTAV, Irapuato, Mexico

BSc in Genomics Sciences [2012], graduated with honors from the National Autonomous University of Mexico (UNAM).

Publications

A. N. Espino-Vázquez et al., “Narnaviruses: novel players in fungal – bacterial symbioses,” ISME J., vol. 14, pp. 1743–1754, 2020

J. R. Bermúdez-Barrientos, O. Ramírez-Sánchez, F. W.-N. Chow, A. H. Buck, and C. Abreu-Goodger, “Disentangling sRNA-Seq data to study RNA communication between species,” Nucleic Acids Res., vol. 48, no. 4, 2020.

F. W.-N. Chow et al., “Secretion of an Argonaute protein by a parasitic nematode and the evolution of its siRNA guides,” Nucleic Acids Res., p. 343772, 2019.

A. Rougon-Cardoso et al., “The genome, transcriptome, and proteome of the nematode Steinernema carpocapsae: evolutionary signatures of a pathogenic lifestyle,” Sci. Rep., vol. 6, no. October, p. 37536, 2016.

M. Angel Vences-Guzman, et al., “Discovery of a bifunctional acyltransferase responsible for ornithine lipid synthesis in Serratia proteamaculans,” Environ. Microbiol., vol. 17, no. 5, pp. 1487–1496, 2015.

M. Angel Vences-Guzman, et al., “Agrobacteria lacking ornithine lipids induce more rapid tumour formation,” Environ. Microbiol., vol. 15, no. 3, pp. 895–906, 2013.

Languages

Spanish (mother language)
English, fluently

References

PhD Cei Abreu-Goodger. MSc thesis co-advisor & PhD thesis advisor.
Cei.abreu@cinvestav.mx
PhD Laila Partida-Martínez. MSc thesis co-advisor.
laila.partida@ira.cinvestav.mx
PhD Christian Sohlenkamp. BSc thesis advisor.
chsohlen@ccg.unam.mx

Declaración de Independencia

Por este medio declare que yo he preparado este trabajo de tesis de forma independiente y sin ayuda externa. Especialmente declare que he citado de forma correcta y explícita a los autores y trabajos en los que esta tesis se apoya, así como las contribuciones de las personas que coadyuvaron en su desarrollo.

Lugar: Irapuato, Guanajuato, México
Fecha: 17 de Agosto de 2020 Firma: