Molecular Characterisation of Virulence in *Entamoeba histolytica*

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Kanok Preativatanyou

April 2015

UNIVERSITY OF LIVERPOOL

Acknowledgements

My sincerest gratitude and thanks go to Prof. Neil Hall and Prof. Steve Paterson for their supervision, knowledge, care and support throughout this PhD. I am very delighted to have been given the opportunity to learn how to think integratively about Biology and explain wisely in an evolutionary sense. Over the past four years, Neil has taught me different ways of how to comprehensively approach and solve the problems, changing my learning attitude into the better direction than ever. This makes me very impressed and I will be very fortunate if I have a chance to collaborate with them in the long run!

Heartfelt thanks also go to Dr. Gareth Weedall who has provided me with a lot of knowledge, guidance, inspiration as well as warm encouragement. He often advises me to analyse the data carefully and professionally, upgrading my research skills and performance. Every time I can answer his difficult questions, I get more self confident and relaxed. I have always been very grateful for all his special attention and warm kindness. Absolutely, I can say that Gareth has contributed a lot of success to me in this PhD!

To Dr. Xuan Liu and Dr. Yongxiang Fang, I really appreciate their wonderful Linux command lines and statistical scripts. As Xuan said 'You are in the Big Data Era', this sentence really motivated me to explore more about Bioinformatics. Thanks for their teachings and great encouragement! In addition, I wish to express sincere gratitude to Dr. Linda D'Amore, Dr. John Kenny, Dr. Lisa Olohan, Dr. Margaret Hughes and all CGR members. They gave me a clear understanding of RNA Biology and provided me with good practices in RNA-Seq library preparation and NanoString analysis. I admire their professional laboratory skills very much!

Particular thanks and appreciation go to Dr. Graham Clark of the London School of Hygiene and Tropical Medicine for his kind advice and help on amoeba cultures and high-quality figures. I have learned a lot from his collection of published papers! In addition, sincere thanks also go to my PhD thesis examining committee: Dr. Alistair Darby of the University of Liverpool and Dr. Mark Van Der Giezen of the University of Exeter for revising my dissertation and giving me nice suggestions!

For all of my friends, I would like to convey my special thanks to Dr. Ian Wilson for his kind support and help in writing some perl scripts; Dr. Ian Goodhead for his qPCR software demonstration; Dr. Jennifer Kelly and Sarah Forrester for their warm encouragement and advice on grammar corrections.

I wish to gratefully acknowledge the support from Dr. Padet Siriyasatien of the Department of Parasitology, Faculty of Medicine, Chulalongkorn University, who encouraged me to write up and offered some photomicrographs included in my introductory chapter. Also, appreciation and gratitude are extended to the Faculty of Medicine and Chulalongkorn University for their sponsorship and great support throughout the years!

Finally, I would like to dedicate all my knowledge in this thesis to my parents, grandparents and beloved family. They gave a life and all things to me. Inexpressibly, I am extremely grateful to them and promise that I will take care of them until my last breath.

Abstract

Entamoeba histolytica is a parasitic protozoan that infects the human digestive tract. Infection results from ingestion of cyst-contaminated food or water. To date, *E. histolytica* infection remains a major worldwide public health problem in worldwide endemic areas. The spectrum of clinical manifestations ranges from asymptomatic carrier to mucous and bloody diarrhea or even extraintestinal amoebiasis, usually amoebic liver abscess. Several molecular studies have been carried out to reveal novel aspects of *E. histolytica* infection. However, the studies focused on genomic-wide analysis comparing between *E. histolytica* strains are still limited. Thus, the aims of this project are to comprehensively study the comparative analysis of the whole and small RNA transcriptomes amongst nonvirulent and virulent strains of laboratory cultured *E. histolytica* trophozoites as well as to integrate such transcriptomic findings with the genomic data for advanced understanding of the molecular pathogenesis and virulence in amoebiasis.

In this study, genome-wide transcriptome analysis using Illumina RNA-Seq technology can illustrate significant expression differences between nonvirulent and virulent *E. histolytica* strains. Differential gene expression analysis between nonvirulent Rahman strain and other three virulent strains (i.e. PVBM08B, HM-1:IMSS and IULA:1092:1) reveals that transcripts involved in host cell killing and mucosal invasion, nucleic acid interaction and response to oxidative stress are notably upregulated in the virulent trophozoites. InterProScan results show the upregulation of genes encoding proteolysis-related domains and the co-upregulation of cytoskeleton and actin-modulating domains in the virulent strains. Also, process ontologies related to protein degradation, cellular biosynthesis, DNA metabolism, repair and recombination, mitotic cell division, actin dynamics and response to stress are highly enriched as a core metabolism in the virulent strains, indicating the rapid growing and active metabolic state are the main drivers of virulence. However, the striking underrepresentation of ontologies involved in signaling and regulatory processes was observed in the virulent parasites. It could be inferred that reduced regulation of sensing and correctly responding to the environmental stimuli potentially enable the parasites to become virulent and subsequently cause the invasive infection. Also, NanoString validation reveals the spectrum of virulence-associated gene expression among these four strains, reflecting their different degrees of virulence.

Gene copy number variation (CNV) is widespread among the genomes of the *E. histolytica* strains, reflecting genomic plasticity and variability in gene family content. Herein, this present data show that patterns of CNV contribute to differential expression profiles, therefore it can be extrapolated that differences in gene copy number between genomes could contribute to the variation in phenotypic attributes, including virulence, among *E. histolytica* strains. Also, genome plasticity can also be seen in *Trypanosomes* and *Leishmania*, suggesting that CNV is a potentially important mechanism in generating genetic diversity and regulating gene expression levels in almost exclusively asexual parasite group.

For small RNA transcriptomics, the size-fractionated sRNA sequencing data demonstrate the inverse relationship between antisense sRNA abundance and target gene expression levels, strongly suggesting the sRNA-mediated regulation. Differential sRNA regulation in virulence-associated gene expression was found among strains, indicating that sRNA-mediated post-transcriptional regulation may be important in shaping the parasite virulence. In addition, this study identified the novel putative miRNA from the sRNA sequencing data using the biogenesis-based bioinformatic analysis and qPCR validation, implying that miRNA potentially play a regulatory role in *E. histolytica*. In summary, it can be inferred that genomic plasticity and sRNA-mediated regulation are important mechanisms of virulence modulation in *E. histolytica*.

Table of Contents

Figure Legends

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Table Legends

Chapter 2

Chapter 3

Chapter 5

List of abbreviations

AGO	Argonaute protein
AIG1	Avirulence induced gene 1
ALA	Amoebic liver abscess
BLAST	Basic Local Alignment Search Tool
BWA	Burrows-Wheeler Aligner
CDC	Centers for Disease Control and Prevention
CGR	Centre for Genomic Research
CH	Calponin homology
CHO	Chinese hamster ovary
CNV	Copy number variation
Cp	Crossing point
CP	Cysteine protease
CPM	Count per million
CRD	Carbohydrate recognition domain
DE	Differentially expressed
DGE	Differential gene expression
dsRNA	Double-stranded RNA
ERE2	*Entamoeba* repeat element 2
FDR	False discovery rate
FPKM	Fragments per kilobase of transcript per million fragments mapped
Gal/GalNAc	Galactose/N-acetylgalactosamine
glm	Generalised linear model
GM-CSF	Granulocyte-macrophage colony-stimulating factor
GO	Gene ontology
GOA	Gene Ontology Annotation
GPI	Glycosylphosphatidylinositol
GX	Gene expression
Hgl	Heavy chain of the Gal/GalNAc lectin
HP	Hypothetical protein
HR	Homologous recombination
IGV	Integrative Genomics Viewer
IL	Interleukin
IR	Inverted repeat
kb	Kilobase
KERP	Lysine and glutamic acid rich protein

Lgl	Light chain of the Gal/GalNAc lectin
LIM	Lin-11, Isl-1 & Mec-3
LINE	Long interspersed nuclear element
LR	Likelihood ratio
LRRP	Leucine-rich repeat protein
miRNA	MicroRNA
MRE	Myb recognition element
mRNA	Messenger RNA
MUC2	Mucin 2
MYB DBD	MYB DNA-binding domain
NB	Negative binomial
NCBI	National Center for Biotechnology Information
NO	Nitric oxide
nt	Nucleotide
PBS	Phosphate buffer saline
PCA	Principal component analysis
PI	Phosphatidylinositol
PK	Protein kinase
qPCR	Quantitative polymerase chain reaction
r	Pearson's correlation coefficient
RdRp	RNA-dependent RNA polymerase
REVIGO	Reduce and Visualise Gene Ontology
RhoGAP	GTPase-activator protein for Rho/Rac/Cdc42-like GTPases
RhoGEF	Guanine nucleotide exchange factor for Rho/Rac/Cdc42-like GTPases
RIN	RNA integrity number
RISC	RNA-induced silencing complex
RNAi	RNA interference
RNA-Seq	RNA sequencing
ROS	Reactive oxygen species
rRNA	Ribosomal RNA
SCAR	Suppressor of cAMP receptor
sIgA	Secretory immunoglobulin A
SINE	Short interspersed nuclear element
siRNA	Small interfering RNA
SNP	Single nucleotide polymorphism
SOD	Superoxide dismutase
SOLiD	Sequencing by Oligonucleotide Ligation and Detection

sRNA	Small RNA
SSU rRNA	Small subunit ribosomal RNA
STARP	Sporozoite threonine-asparagine-rich protein
STIRP	Serine-threonine-isoleucine-rich protein
STR	Short tandem repeat
TE	Transposable element
TF	Transcription factor
TMK	Transmembrane kinase
TNF-α	Tumor necrosis factor-alpha
tRNA	Transfer RNA
tRNA-linked STR	Transfer RNA-linked short tandem repeat
TTO	Terminal-tagging oligo
URE3-BP	Upstream regulatory element 3-binding protein
WASP	Wiskott-Aldrich syndrome protein
WH2	Wiskott-Aldrich syndrome homology region 2
Wm	Wortmannin

1.1 Amoebiasis

Human amoebiasis is caused by *Entamoeba histolytica*, a parasitic protozoan that infects the human intestinal tract. Infection results from ingestion of cyst-contaminated food or water. Progression of disease occurs by multiplication and tissue invasion of trophozoites into the colon mucosa. Mostly, trophozoites commensally colonise and feed on bacteria but can invade the mucosal epithelium, typically resulting in 'flask-shaped' ulcers. In some cases, parasites penetrate into the intestinal portal vein and spread to other extraintestinal organs, including liver, lungs and brain [1]. *E. histolytica* infection remains a major worldwide public health problem and is endemic in many developing countries. As transmission is via a feacal-oral route, communities with poor sanitation and nutrition are at higher risk. Also, *E. histolytica* infection has been widely documented in travelers returning from amoebiasis-endemic areas and in men who have sex with men [2]. As estimated in 1986, *E. histolytica* affected approximately 10% of the world's population with an associated mortality rate estimated between 40,000 to 110,000 deaths per year [3]. The spectrum of disease severity can be manifested from asymptomatic colonisation to mucous and bloody diarrhea (dysentery) or even a complication of invasive amoebiasis, usually amoebic liver abscess. Interestingly, the majority of cases are asymptomatic carriers whilst invasive amoebiasis is rare [3]. There are still many unanswered questions concerning amoebic pathogenesis as well as differences of virulence among strains of parasite. So far, the studies, focused on genomic and transcriptomic data comparing nonvirulent and virulent strains of *E. histolytica*, are still limited.

1.2 The life cycle of *Entamoeba histolytica*

The life cycle of *E. histolytica* consists of an infective cyst stage and a pathogenic multiplying trophozoite stage. Typically, infection occurs via the faecal-oral route by ingestion of stool-contaminated food and water, or even transmitted from heterosexual and homosexual activity. After ingestion into the upper gastrointestinal tract, the excystation is triggered by exposure of the encysted parasite to water, bicarbonate and bile [4]. The infection occurs in the human colon where trophozoite emerges from the mature cyst. In general, especially in asymptomatic individuals, trophozoites commensally colonise the colon mucosa by phagocytosing enteric bacteria and multiply by binary fission as explained in Figure 1.1. To complete the life cycle, trophozoites re-encyst and are finally released to the environment via the stool but the stimuli for this process of encystation in *E. histolytica* is still unknown. These cysts can remain viable and infective in the environment for several weeks to months [5,6].

Most cases (90%) of the *E. histolytica* infection are asymptomatic cyst shedders [3,6]. The invasive trophozoites which are capable of penetrating the colon and even hematogenously spreading to infect other organs are the rare form. Degrees of disease severity range from colonic invasion, ulcerative colitis, bloody mucus dysentery to extraintestinal spread [1,3,6]. As clinically reported, the invasive trophozoites can spread to almost all human body tissues such as liver, lungs, brain, pericardium, peritoneum, cutaneous tissue, genitourinary tract and even bone [6]. Most commonly, trophozoites can be disseminated to the liver by vascular invasion via the hepatic portal venous system, resulting in apoptosis of hepatic immune cells and inflammation. As a result, amoebic liver abscess (ALA) is the most common complication of extraintestinal infection. However, devastating tissue invasion is not the essential part of the life cycle of the parasite since those invasive trophozoites could not develop cysts and complete their life cycle outside the colonic mucosa. Therefore, this virulent behaviour is likely to reduce the parasite's fitness because invasive trophozoites have no ability to cause the new infection and therefore are no longer to contribute their genetic content to the gene pool of the next generation.

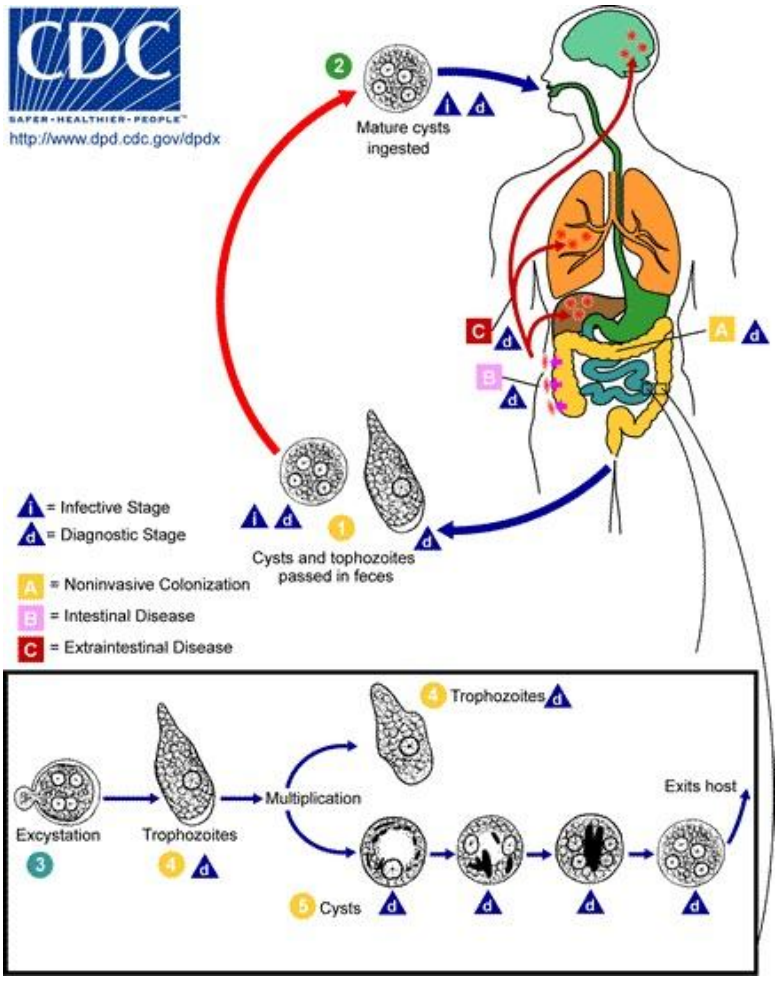


Figure 1.1: Life cycle of *Entamoeba histolytica*. Infection occurs by ingestion of food and water contaminated with mature cysts or even transmission from heterosexual and homosexual activity. Parasites undergo excystation by compound stimuli in the small bowel and then proliferate by mitotic binary fission in the luminal mucosa of the colon. To complete the life cycle, trophozoites encyst again and are finally passed through the stool in the cyst form to infect the new host. The figure is reproduced from the Centers for Disease Control and Prevention (CDC; 2010) [5].

1.3 Mechanisms of pathogenesis

In invasive amoebiasis, virulent trophozoites, triggered by unknown stimuli, can express several pathogenic factors for mucosal invasion. Degradation of the tissue matrix and cytolysis are the hallmarks of parasite invasion. The protective colonic mucous layer, consisting mainly of mucin 2 (MUC2) protein, is the first host target for the parasite to adhere to and degrade [7]. As recent findings reviewed by Lejeune *et al.*, 2009, the parasite uses the Galactose/N-acetylgalactosamine (Gal/GalNAc) lectin for interaction with terminal Gal and GalNAc of the MUC2 polymer and secretes cysteine proteinases (EhCPs) for depolymerisation of MUC2, resulting in weakening of the mucous layer and providing the way for interaction with the host cell [8-10]. Also, the serine-threonine-isoleucine rich protein, EhSTIRP, may have a role in concert with the Gal/GalNAc lectin in host cell adherence and contact-induced apoptosis [11].

E. histolytica cysteine proteinases also cleave the tight junctional complex between the enterocytes, resulting in the detachment of colonic mucosal epithelium [12]. The leukocyte recruitment, especially macrophages, monocytes and neutrophils, occurs due to proinflammatory cytokines released from damaged enterocytes, i.e. tumor necrosis factor-alpha (TNF-α), interleukin (IL)-1α, IL-6, IL-8 and granulocyte-macrophage colony-stimulating factor (GM-CSF) [13]. These activated immune cells play important roles in preventing parasite invasion by producing reactive oxygen species (ROS), nitric oxide (NO) and cytotoxic enzymes, e.g. cathepsinG, to damage trophozoites [14]. To resist these host defences, trophozoites can reduce the toxicity of these reactive molecules by surface-bound peroxiredoxin, superoxide dismutase (SOD) and NADPH:flavin oxidoreductase [15-19]. However, ROS and other cytotoxic substances from immune cells cause nonspecific destruction and apoptosis in surrounding tissue, resulting in clinical symptoms of diarrhea and/or dystentery [8]. Characteristically in colonoscopic and pathological findings, the flask-shaped ulcers could be found in the colonic mucosa of patients with amoebic colitis due to phagocytic activity of *E. histolytica* trophozoites as shown in Figure 1.2 [8].

Rarely, systemic invasion of trophozoites can occur and develop eventually extraintestinal amoebiasis in almost all body tissues as mentioned before. The most common manifestation of extraintestinal infection is ALA. After reaching the liver by hematogenous spread via the hepatic portal vein system, trophozoites trigger periportal inflammation and then rapidly lyse acute inflammatory cells and surrounding host hepatocytes by release of lytic enzymes. Recently, some virulence factors have been characterised for their roles in the pathogenesis of ALA [20-22]. Amoebapore-A, a pore

forming peptide, has a crucial role in ALA formation by nonspecific insertion into the host cell and pore formation, causing cytolysis [20,23]. More recently, a lysine and glutamic acid rich protein (KERP1) has been proposed as a pathogenic factor for its associated upregulation in ALA, however it remains to be confirmed [22]. As a consequence of liver infection and following inflammation, abscesses are formed with collection of necrotic debris and trophozoites could be found, if present, at the rim of the abscess capsule.

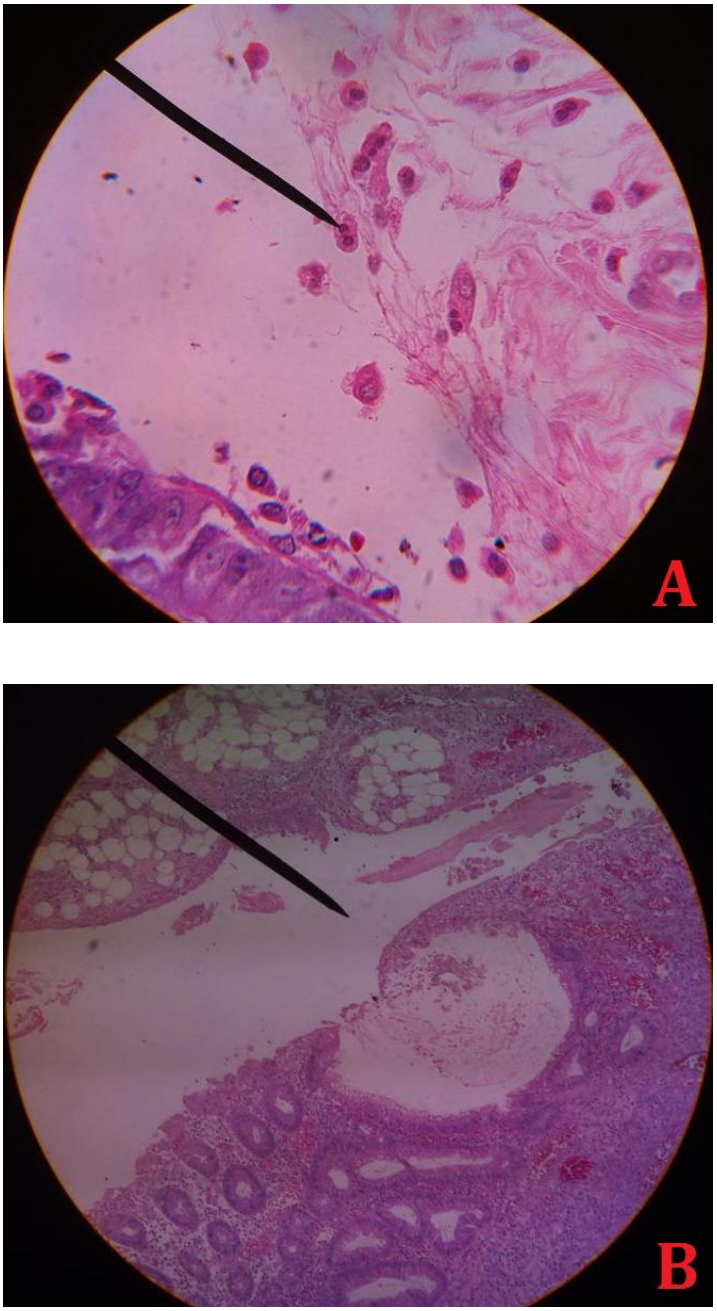


Figure 1.2: **Histopathological preparation of colonic biopsy from the patient with amoebic colitis.** *E. histolytica* trophozoites and a characteristic flask-shaped ulcer are identified by the arrows (A) and (B), respectively. The neighboring inflammation of colonic mucosa can be identified by vasodilatation as well as red blood cell and neutrophil extravasation.

This histological section was kindly offered by Associate Professor Dr. Padet Siriyasatien (MD, PhD), Department of Parasitology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.

1.4 Genomic structure and organisation

The genome organisation of *E. histolytica* has been extensively documented [24,25]. The genome sequence of virulent *E. histolytica* HM-1:IMSS was published in 2005 [24]. After genome reassembly and reannotation it was re-published in 2010, its genomic features consist of 20.80 megabases in 1,496 scaffolds (data available at http://amoebadb.org/amoeba/) [25,26]. The sequence of the genome is AT rich (approximately 75%) and contains 8,333 predicted genes [25]. Uniquely, *Entamoeba* transfer RNA (tRNA) genes are organised in arrays, separated by DNA which consists of tRNA-linked short tandem repeats (tRNA-linked STRs), possibly acting as telomeres [27]. It is interesting that variable arrangements of tRNA gene arrays in *Entamoeba* species are associated with their evolution of species divergences as shown in Figure 1.3A [28]. Moreover, these unique tRNA organisations with variable STRs suggest that tRNA genes are likely to be the 'hotspots' of recombination and genetic diversity in this parasite [28,29]. Due to high polymorphisms in their sequences, i.e. number of repeats and arrangement pattern, the tRNA-linked STR loci have been used as genetic markers to study the *E. histolytica* population structure as well as the relationship between the parasite lineages and their geographical regions [30]. Also, these unusual features have been used to study the possible correlation between the parasite genotypes and the clinical outcomes [31-34]. However, only few associations with disease outcomes were reported in limited geographical regions and not entirely related to the virulence variability.

Transposable elements (TEs) are abundant in *Entamoeba* genome, including EhLINEs, EhSINEs and *Entamoeba*-specific repetitive elements [35]. These TEs can affect the expression of neighboring genes by several mechanisms, e.g. heterochromatin formation and alternative 3' splice site or promoters [35]. As such, it can be implied that genomic location of these TEs may determine the virulence phenotype.

In addition, the *E. histolytica* genome reveals remarkable evolutionary characteristics concerning secondary gene loss and lateral gene transfer from prokaryotes for metabolic adaptation to an anaerobic environment. Its metabolism resembles two other amitochondrial parasite, *Giardia lamblia* and *Trichomonas vaginalis* in terms of catabolism and biosynthesis [36]. However, some points remain to be further investigated. For instance, ploidy, haploid chromosome number and chromosome size are variable between strains, suggesting considerable genomic size plasticity in *E. histolytica* [37].

Previously, *E. histolytica* has been believed to be a clonal or asexual organism since no genotypic change can be observed in *in vitro* cultivation and long-term animal passages.

The interesting question has been raised whether sexual reproduction occurs in multiploid *Entamoeba* species. However, it has recently been shown that genes involved in meiosis and homologous recombination have been identified, implying that sexual reproduction could possibly occur and may contribute to their genetic diversity [38,39].

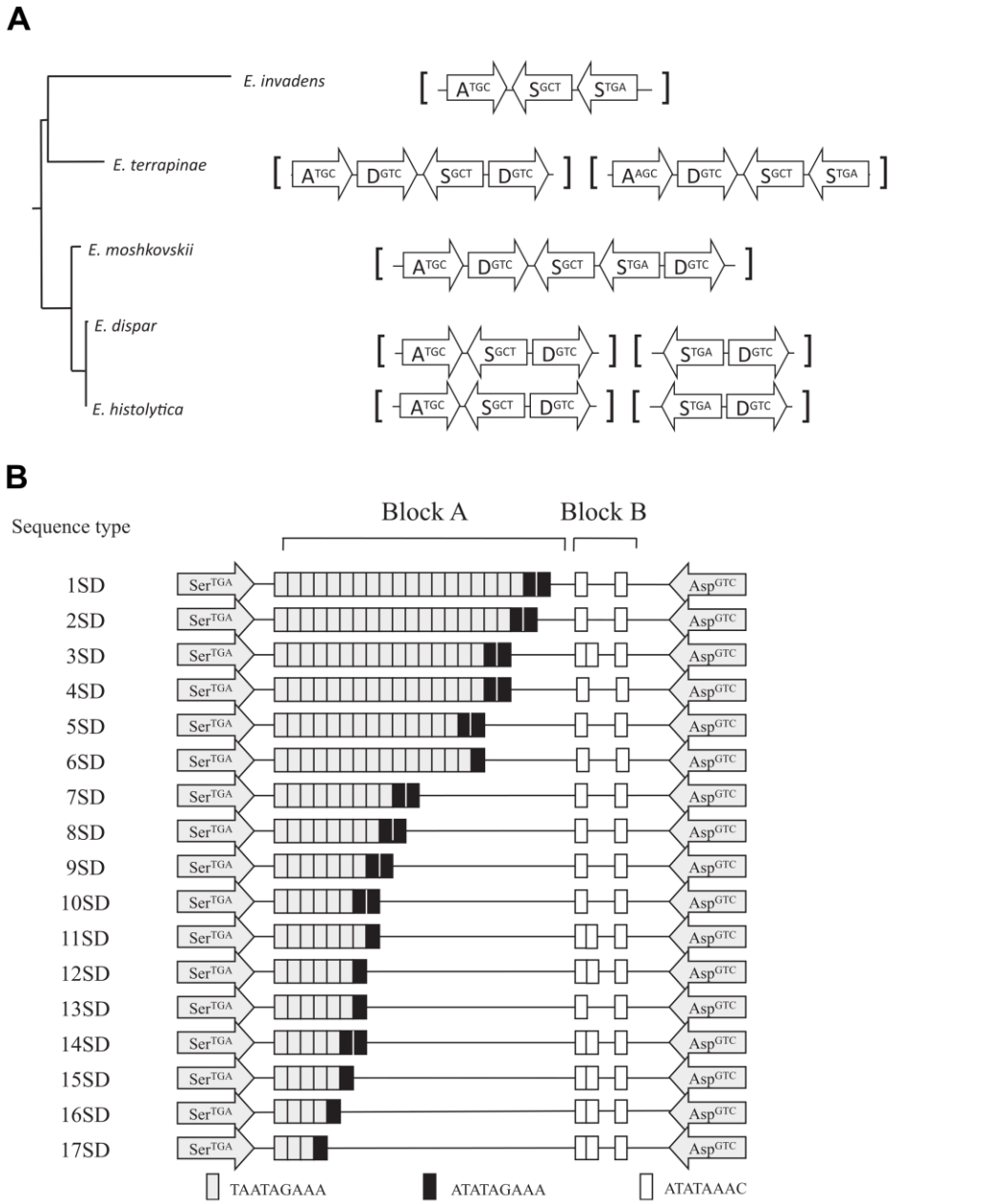


Figure 1.3: Variable arrangements of tRNA gene arrays. Unique array unit organisations of tRNA-Ala, tRNA-Ser and tRNA-Asp in five *Entamoeba* species (A). The arrows refer to the tRNA gene orientation and the amino acid with corresponding anticodon is designated inside. The array-based relationships among species were also shown. Of these polymorphisms, 17 patterns of tRNA-linked STR organisation in the S^{TGA}-D intergenic region of *E. histolytica* were illustrated (B). This figure is reproduced with permission from Tawari *et al.*, 2008 [28].

1.5 Closely related *Entamoeba* species relevant to human amoebic research

There are other two closely related species with identical microscopic morphology: *Entamoeba dispar* and *Entamoeba moshkovskii*. Although these three *Entamoeba* species share common morphological features, there are certain genetic divergences among these three species. Based on small subunit ribosomal RNA (SSU rRNA) gene sequences retrieved from 18 taxa, *E. histolytica*, *E. dispar* and *E. moshkovskii* are phylogenetically clustered together within the same group of species producing tetranucleate cysts as shown in Figure 1.4 [28]. *E. histolytica* and *E, dispar* are closely related each other while *E. moshkovskii* is more distantly related.

Entamoeba dispar

E. dispar has been initially proposed by Emile Brumpt in 1925 that it lacked the ability to cause the disease in humans and experimental animals [40]. However, Brumpt's nomenclature was disregarded because there is no morphological difference between *E. dispar* and *E. histolytica*. Also, *Entamoeba* species clinically isolated from asymptomatic individuals could trigger the disease in experimental subjects [41]. In 1973, different lectin agglutination profiles between clinical samples isolated from patients and asymptomatic individuals were reported as the first biochemical evidence, referring to subgroups within *E. histolytica*. Several evidences showing the existence of *E. dispar* have been reported including isoenzyme analysis [42], antigenic differences [43] and genetic markers [44]. In 1993, Diamond and Clark validated and redescribed 'non-pathogenic *E. histolytica*' as *E. dispar,* formerly named by Brumpt in 1925. Reported so far, this species has been isolated from a wide range of primate hosts including old world monkeys, new world monkeys and human [45].

Compared to the *E. histolytica* genome, the genome of *E. dispar* strain SAW760 contains slightly greater size of 22.96 Mbp in 3,312 scaffolds with 8,748 genes in total [46]. Its genomic AT content is high, about 76.5% , very similar to the *E. histolytica* genome [46]. As mentioned previously, the similarities of genetic characteristics between these two morphologically identical species support that these two sibling species share a recent common ancestor as shown in the phylogenetic tree of 18 *Entamoeba* species using SSU rRNA gene sequences (see Figure 1.4). When comparing the transcriptome between *E. histolytica* HM-1:IMSS and *E. dispar*, a key difference is the lack of members of cysteine proteinase gene family (i.e. EhCP1 and EhCP5) and downregulated expression of EhCP8 in *E. dispar* [47-49]. Also, the *KERP1* gene encoding surface-associated protein involved in host cell adhearence and ALA formation is present in *E. histolytica* (EHI_098210) but absent in *E.*

dispar [50]. Furthermore, the activity of the pore-forming peptide amoebapore A, implicated in the killing of engulfed bacteria and the host cytolytic reaction, is 25 times more active in *E. histolytica* HM-1:IMSS than *E. dispar* SAW142 [23].

E. dispar has been characterisd as non-pathogenic commensal in the human colon and the non-pathogenic *E. dispar* SAW760 strain has been extensively used for the experimental study of differential virulence among *Entamoeba* species. However, it was recently reported by *Dolabella et al.*, 2012 that *E. dispar* xenic strain ICB-ADO, clinically isolated from a non-dysenteric Brazilian patient could cause liver necrosis and liver abscesses in a hamster model [51]. Therefore, these findings in xenic strain ICB-ADO suggests that *E. dispar* could potentially exhibit virulent phenotype, resulting in tissue destruction, inflammation and even abscess formation in human.

Previous studies have attempted to elucidate the virulence potential across *Entamoeba* species in relation to the bacterial interplay [52-54]. As published many years ago, a change in the zymodeme patterns was found in *E. histolytica* strain CDC:0784:4 trophozoites after interacting with bacteria, and associated with their increased capability to trigger the destruction of cultured monolayer cells and abscess formation in hamster model [54]. Recently, Galván-Moroyoqui *et al.*, 2008 have reported that co-culture of *E. histolytica* HM-1:IMSS with enteropathogenic bacteria increased the expression of Gal/GalNAc lectin, the cysteine proteinase activity as well as the cytopathic effect whereas *E. dispar* SAW760 did not show any significant change [52]. In contrast to the findings of Dorabella *et al*., 2012, *E. dispar* ICB-ADO which was cultured under xenic condition with bacterial flora showed the pathogenicity both *in vitro* and *in vivo*, suggesting that it is possible that the interaction between *E. dispar* and gut bacteria in the host colon can lead to alteration in the regulation of virulence and eventually cause the disease [51].

Entamoeba moshkovskii

E. moshkovskii is another *Entamoeba* species morphologically indistinguishable from *E. histolytica* and *E. dispar* in both trophozoite and cyst forms. Originally, it has been identified to be a free-living and non-pathogenic amoeba in sewage in Moscow, Russia in 1941 [55]. The Laredo strain of *Entamoebic moshkovskii* was initially isolated as the first case of human infection in Laredo, Texas in 1961 [56]. However, due to its morphology identical to *E. histolytica*, the first isolate was named *E. histolytica* Laredo strain. Different from typical *E. histolytica*, the Laredo strain can grow at the room temperature, survive in osmotic stress conditions and is commonly found in polluted water [57]. Clark and

Diamond, 1991 demonstrated different profiles of the polymerase chain reaction-restriction fragment length polymorphism analysis of the SSU rRNA genes or 'riboprinting' between the Laredo strain and *E. histolytica,* indicating that the '*E. histolytica*-like' Laredo strain is truly a strain of *E. moshkovskii* , based on this DNA marker [57].

E. moshkovskii infection could be found ranging from 1% to 50% of the *Entamoeba* complex parasites (*E. histolytica*/ *E. dispar*/ *E. moshkovskii*) detected in worldwide collected stool samples [58]. Recently, Shimokawa *et al.*, 2012 have found that susceptible mice inoculated intracaecally with $1x10^6$ trophozoites of *E. moshkovskii* exhibited diarrhea, colitis and weight loss and also reported a longitudinal study in Bangladesh that 42 of 1,426 diarrheal cases in infants were associated with *E. moshkovskii* infection [58]. This therefore implies that *E. moshkovskii* has potential pathogenic capacity to cause disease in human.

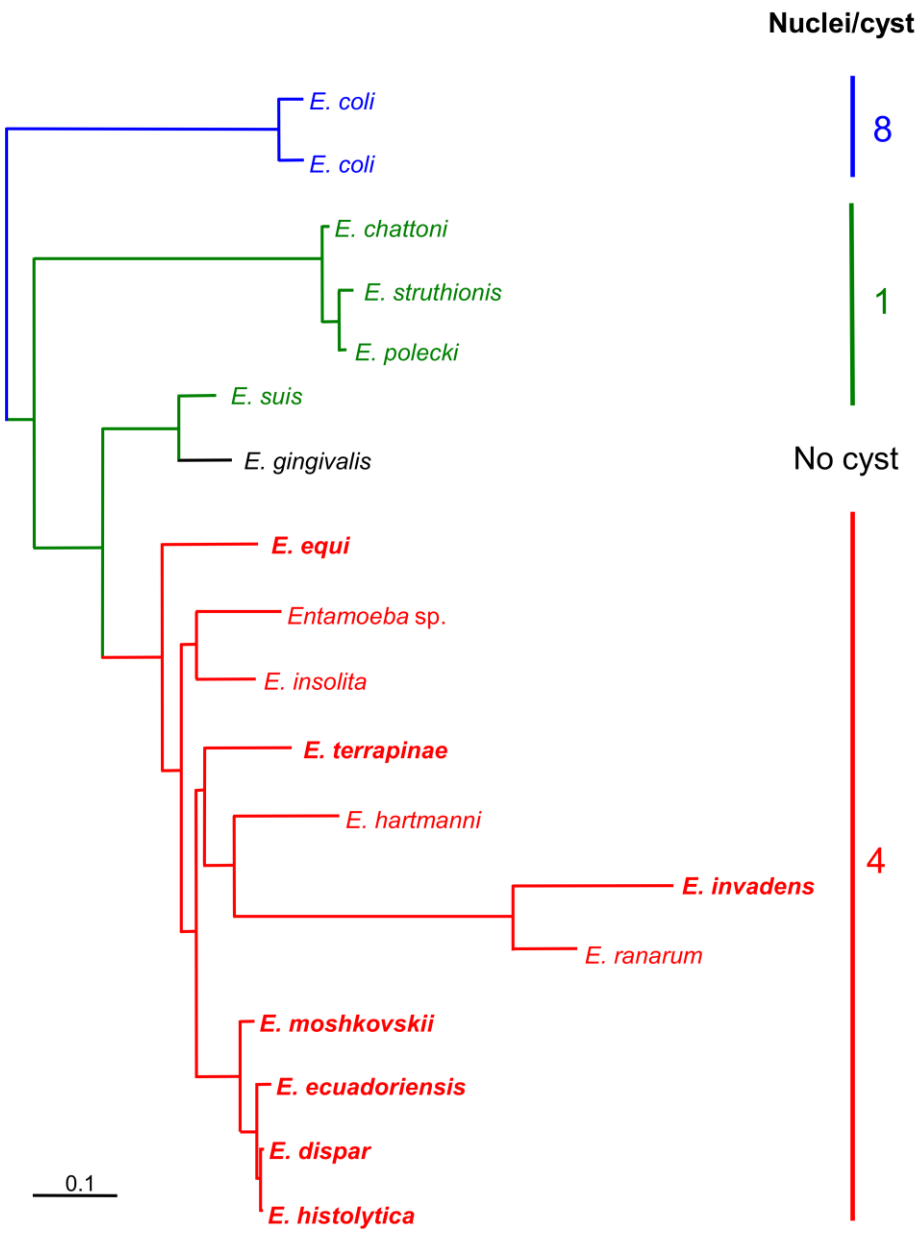


Figure 1.4: Molecular phylogeny of 18 *Entamoeba* species based on SSU rRNA gene sequences across 1,572 nucleotide positions. This tree is rooted with two sequences of *E. coli* and the scale bar represents 0.1 changes per nucleotide position. This tree figure was kindly offered by Dr. Graham Clark, Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom.

1.6 Differential virulence of amoebiasis across *E. histolytica* strains

A wide spectrum of clinical manifestations has been reported in individuals infected with *E. histolytica* ranging from asymptomatic carriers to extraintestinal invasive diseases [3]. Nevertheless, most cases, approximately 90%, showed clinical histories of asymptomatic infections or mild intestinal symptoms, suggesting that not all *Entamoeba* infections exhibit equally the virulence [3]. Clinical observations have raised the question which factors are responsible for a degree of virulence variability of the disease. This variation in clinical symptoms could be explained that many cases are indeed infected by microscopically indistinguishable *E. dispar* and/or different strains of *E. histolytica* as well as influenced by different host susceptibility [8,14,59].

For the host conditions, the growth of parasites could be affected by intestinal micro-environments variable between individual patients including the bacterial flora, ROS, protective mucus barrier and secretory immunoglobulin A (sIgA) secretions [14,59,60]. The intestinal bacterial flora is a direct nutrient source for the trophozoites and also controls the pH and redox potential of the colon. As previously mentioned, releases of nitric oxide and ROS including oxygen ions and peroxides by polymorphonuclear cells, monocytes and macrophages cause harmful effects to the trophozoites [14]. The sIgA can block the adhesion of the trophozoites to the intestinal mucosal cells by neutralising the parasite surface molecules and also recruit the complement proteins to promote the opsonisation and the lytic pathway [60-62]. The host gender could influence to the virulence of the disease since amoebic dysentery and ALA are more frequently found in men than women for unknown reasons [63]. Moreover, malnutritional status and leptin receptor mutant are associated with the increased susceptibility to the invasive infection [64,65]. Taken together, these host conditions vary from individual to individual, resulting in variable degree of host susceptibility to the parasite invasion.

Regarding the parasites, the relative virulence between different cultured strains/clinical isolates can be determined by certain phenotypic parameters of virulence such as the rate of destruction of MDCK cell monolayer by the cytopathic effect [66], the ability to cause the ALAs in a hamster model [67], the erythrocyte hemolysis and phagocytosis rate of the parasites [68] as well as the resistance to complement-mediated lysis [62]. A number of *E. histolytica* strains were isolated and well characterised, i.e. HM-1:IMSS strain obtained from the colonic biopsy of the dysenteric patient in Mexico in 1971; Rahman strain isolated from the feces of asymptomatic sailor in UK in 1964; PVBM08B and PVBM08F strains isolated from colonic biopsy and feces of the same patient in Italy in 2007;

IULA:1092:1 strain isolated from a symptomatic patient in Venezuela in 1992 [69,70]. The genealogic relationships amongst well-characterised *E. histolytica* strains are estimated using 3,696 polymorphic sites as illustrated in Figure 1.5.

HM-1:IMSS is a virulent strain extensively used as a genomic reference strain and characterised as the most virulent strain since it could produce hepatic lesions in 19% of newborn hamsters injected with just 20 trophozoites [71] . Conversely, Rahman is considered as a 'nonvirulent' strain due to its defects in phagocysis and cytopathic activity as well as its inability to cause abscess lesions when inoculating a large number of trophozoites into hamsters whilst other *E. histolytica* strains exhibit the virulent phenotypes, resulting in amoebic colitis and/or ALAs [72]. Also, the transcriptomic and proteomic evidences reported its reduced expression of antioxidative proteins in comparison to HM-1:IMSS, indicating the decreased resistance to oxidative and nitrosative stresses in Rahman [16,73].

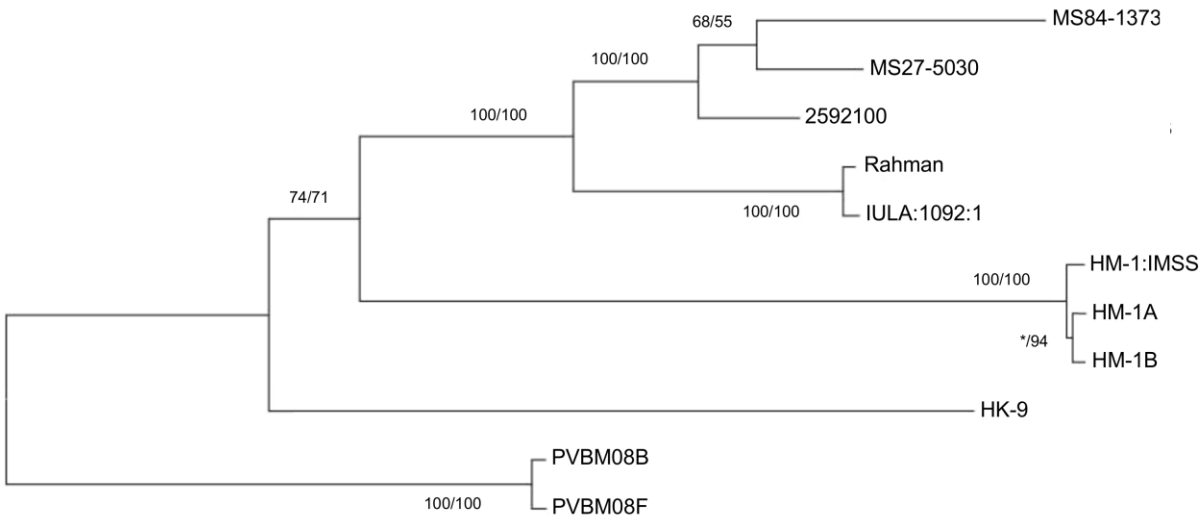


Figure 1.5: Phylogenetic relationship of 11 well-characterised *E. histolytica* strains based on 3,696 polymorphic sites. The tree was constructed using distance-based method and maximum likelihood with shown bootstrap values in respective order: Distance/ML. The bootstrap value less than 50% is designated as an asterisk. This tree figure is reproduced with permission from Weedall *et al.*, 2012 [70].

Over a decade ago, after launcing of the complete genome sequence and annotation of *E. histolytica*, it has highlighted the amoebic virulence research in genomic and transcriptomic scales. Several genes responsible in amoebic pathogenesis have been extensively investigated among *E. histolytica* strains and *E. dispar* by genome-wide analyses [73-77]. MacFarlane and Singh, 2006 applied the DNA microarray containing 2,110 genes to unravel the transcriptional differences among *E. histolytica* HM-1:IMSS, Rahman and *E. dispar* SAW760 and found that 415 genes in *E. dispar*, 32 genes in *E. histolytica* Rahman and 29 genes in both *E. dispar* and *E. histolytica* Rahman were downregulated relative to *E. histolytica* HM-1:IMSS [74]. Among these, 29 lower expressed genes in both nonvirulent strain/species involved in stress response and virulence, for instance: Fe hydrogenase, peroxiredoxin, lysozyme, sphingomyelinase, a protein with domains homologous to a *Plasmodium* sporozoite threonine-asparagine-rich protein (STARP) and a hemagglutinin.

Using a 70mer DNA microarray covering 6,242 genes, Davis *et al.*, 2007 showed key transcriptomic differences in the expression of virulence-associated genes, i.e. CPs, the light chains of the Gal/GalNAc lectin (Lgls) and calmodulin between HM-1:IMSS and Rahman [76]. As the sensitivity of this array was increased due to its coverage of more genes accounting for ~80% of the current genomic database than the previous study of MacFarlane and Singh and its ability to differentiate the paralogous transcripts of the same gene family, 353 putative differentially expressed (DE) genes (with fold change greater than 2) between HM-1:IMSS and Rahman were identified [76]. In this microarray study, a number of DE genes involved in pathogenesis and virulence were upregulated in HM-1:IMSS including cysteine proteinases (EhCP4, EhCP6 and EhCP7), bacterial interaction/killing proteins (AIG1-like family proteins and lysozyme), members of protein kinase family, Rho and Ras family GTPases, 70 kDa heat shock protein, BspA-like leucine-rich repeat protein and calmodulin [76]. In addition, comparative proteomic study by Davis *et al.*, 2006 identified two antioxidative proteins, i.e. peroxiredoxin and superoxide dismutase, as important virulence determinants in HM-1:IMSS and found that peroxiredoxin overexpression in Rahman resulted in increased resistance to the oxidative stress [16].

Another interesting protein is the light subunit (35 kDa) of the Gal/GalNAc lectin complex which functions as a primary adhesive molecule for host cell adhesion and killing [78]. As previously published, cDNA representational difference analysis identified the under-representation of Lgl1 transcripts in Rahman relative to HM-1:IMSS [79]. Also, downregulation of Lgl1 by antisense inhibition and dominant negative N-truncated Lgl1 expression in HM-1:IMSS were associated with reduced erythrophagocytosis [79,80]. Contrastedly, this microarray study showed no significant difference of Lgl1 between two

strains but found conversely that Lgl3 was significantly upregulated 22-fold in Rahman compared to HM-1:IMSS, leading the interesting question that over-represented Lgl3 in Rahman might be associated with its reduced phagocytosis and virulence [76].

As noted above, most previous publications were experimentally designed for *in vitro* transcriptomic studies using axenically cultured strains. However, it was reported before that axenisation and long term *in vitro* cultivation potentially decrease the virulence of the parasites as well as reduce their ability to resist the complement lysis [81]. Therefore, it is indeed worth studying their gene expression changes during interaction with the colon mucosa, both *in vivo* and *ex vivo*, to reflect virulence determinants directly responsible for their different phenotypes.

The first transcriptome *in vivo* study of the HM-1:IMSS trophozoites isolated from the infected CBA/J mice colon compared to those from axenic culture was performed using an 25mer Affymetrix array platform covering 9,435 open reading frames by Gilchrist *et al.*, 2006 [77]. Similar to the previous study of Davis *et a*l., 2007, signaling genes (i.e. transmembrane kinases, Ras and Rho family GTPase), EhCP4, AIG1-like family proteins and calcium-binding proteins were found to be upregulated in *in vivo* condition. However, this study focused on the transcriptomic responses of the same strain between *in vivo* and *in vitro* conditions to see the impact of mucosal colonisation and invasion rather than exploring the *in vivo* key differences of transcriptomic profiles between virulent and nonvirulent strains.

More recently, Thibeaux *et al.*, 2013 investigated the comparative transcriptomic profiles between *E. histolytica* Rahman and HM-1:IMSS strains in axenic culture and upon *ex vivo* contact with the intestinal mucus on human colon explants, using whole genome microarray analysis [82]. It was found that upon contact with the mucus, a number of genes involved in glycolysis (i.e. triosephosphate isomerase and glucose-6-phosphate isomerase) and carbohydrate catabolism (i.e. starch-binding protein, β-amylase, β-galactosidase, β-N-acetylhexosaminidase, 4-α-glucanotransferase and oligosaccharide-glycosyltransferase) were exclusively upregulated in HM-1:IMSS relative to Rahman.

Interestingly, such upregulated transcripts encodes enzymes, i.e. β-galactosidase and β-N-acetylhexosaminidase, that play a crucial role in MUC2 degradation in conjunction with cysteine proteinases by cleaving the oligosaccharide from the protective MUC2 mucin layer into UDP-glucose [82]. Surprisingly, the upregulated β-amylase absent in human might participate in hydrolysing oligosaccharide released from the degraded MUC2 layer into glucose-1-phosphate. Altogether, both UDP-glucose and glucose-1-phosphate could be

utilised for energy production by glycolytic pathway. This is consistent with the upregulation of genes in glycolytic pathway previously mentioned. To prove the possible role of β-amylase in mucosal invasion, double-stranded RNA (dsRNA)–based knock down experiment and histological study of mucus layer degradation were done [82]. The results showed significantly reduced β-amylase abundance and intact protective mucus layer in dsRNA-treated HM-1:IMSS trophozoites, inferring that β-amylase deficient parasites could not invade the physical mucus barrier and not utilise the MUC2-associated oligosaccharides as a carbon source for energy production. The authors also suggested to the possibility to develop *E. histolytica* β-amylase as a potential specific therapeutic candidate in invasive amoebiasis due to the absence of this enzyme in the human genome [82].

To date, endogenous small non-coding RNAs have been reported in many human protozoan parasites such as *Giardia lamblia*, *Trichomonas vaginalis*, *Toxoplasma gondii*, *Trypanosoma brucei* and *E. histolytica* [83-86]. Typically, these small RNAs (sRNAs) can modulate the gene expression at post-transcriptional level by complementary base-pairing to the target mRNA transcripts and subsequently causing translation repression and mRNA cleavage [87-91].

Two major classes of sRNAs, small interfering sRNAs (siRNAs) and microRNAs (miRNAs) have been previously reported for their regulatory roles in *E. histolytica* [85,86,92-94]. Zhang *et al*., 2008 demonstrated the presence of 27 nt sRNAs with 5'-polyphosphorylated termini which were associated with an Argonaute protein and play a role in the siRNA pathway in *E. histolytica* [85]. Also, these 5'-polyphosphate sRNAs have been identified for their roles in silencing of gene expression at both transcriptional and post-transcriptional levels [85,86,92,93]. Recently, the sRNA sequencing data in Rahman and HM-1:IMSS strains revealed that these siRNAs regulate the expression of certain genes including two virulent *EhSTIRP* genes in a strain-specific manner [92]. For miRNAs, the seventeen putative miRNA candidates were firstly predicted using the computational method by De *et al*., 2006 after the complete genome sequencing of *E. histolytica* published in 2005 [36,95]. Recently, the deep sequencing data of sRNA transcriptome in *E. histolytica* HM-1:IMSS strain revealed a total of 199 potential miRNA candidates predicted from the hairpin-forming precursor sequences as well as 66 potential target genes [94]. However, biological significance of miRNAs towards the differential virulence among *E. histolytica* strains needs to be elucidated.

As explained so far, transcriptomic differences in relevance to virulence variability between virulent and nonvirulent strains of *E. histolytica* have been explored. A number of

genes implicated for amoebic pathogenicity and virulence have been identified in different experimental conditions. To complete the jigsaws of the knowledge, transcriptomic networks controlling the degree of parasite virulence in each strain as well as their transcriptional regulation need to be investigated more thoroughly.

1.7 Treatment of amoebiasis

Asymptomatic individuals who were found to have *E. histolytica* cysts in their stool specimens are recommended for amoebicide medication [6,61,96]. Luminal amoebicides such as diloxanide furoate (Furamide), quinodocholor (Entero Quinol), iodochlorhydroxyquin (Vioform) and paromomycin (Humantin) are commonly used to treat asymptomatic amoebiasis [61,96]. For invasive cases, metronidazole (Flagyl) is a drug of choice and provides the most effective treatment for amoebic colitis and extraintestinal amoebiasis including ALA, pleuropulmonary amoebiasis and brain abscesses [61,96]. Tinidazole (Tindamax) could be used as an alternative tissue amoebicide for such invasive amoebiasis. Also, a course of luminal amoebicide, i.e. diloxanide furoate, is usually prescribed in combination with metronidazole to completely eradicate the infection [61,96]. For hepatic abscess, surgical drainage might be required to get the clinical improvements and reduce the severe systemic complications including rupture into the pleura and the pericardial cavity. Emergency drainage is mandatory in case of abscess rupture into the pericardial cavity, resulting in cardiac tamponade [6].

In addition to such above amoebicide medication, oral rehydration is also essential in patients with colitis symptoms since the water reabsorption is affected due to colonic mucosal inflammation. For cases with severe dehydration, intravenous fluid replacement will be considered.

1.8 Project objectives and methodology

In this present study, I hypothesised that there should be certain differences in expression of genes between virulent and nonvirulent strains of *E. histolytica*, contributing to their virulence phenotype. RNA sequencing (RNA-Seq) is a novel technology for transcriptomic studies by using next-generation sequencing technologies to sequence cDNA reverse transcribed from RNA. RNA-Seq can provide genome-wide transcriptomic data so that researchers can understand biological implications of gene expression. Additionally, no previous research has been done for RNA-Seq analyses in axenically cultured strains of *E. histolytica* in relation to their clinical phenotype.

Therefore, I applied this RNA-Seq technology for whole transcriptomics to explore differences in gene expression in relevance to virulence among the four laboratory-cultured and well-characterised strains of *E. histolytica* as described in Chapter 2. Functional characterisation and annotation of protein domains found in each set of differentially expressed genes between transcriptomes of nonvirulent and virulent parasites were done to reveal the biological functions and implications in relevance to virulence and pathogenesis. Also, gene ontology (GO) enrichment analysis and summarisation by REVIGO software were performed to show comprehensive networks of both overrepresented and underrepresented gene ontologies in the transcriptomes of virulent strains. Taken together, these transcriptomic analyses can fulfil knowledge of the molecular basis of virulence in *E. histolytica* infection.

In Chapter 3, novel gene expression analysis system, the NanoString nCounter® technology was applied to validate the accuracy of RNA-Seq experiments previously done, using a representative set of putative virulence-associated genes. In addition, the obtained nCounter data was used to compare transcriptional profiles of such representative genes between strains in relation to their differential virulence.

Single nucleotide polymorphisms (SNPs) and genomic plasticity including gene gain or gene loss and gene copy number variation could be found among the genome of *E. histolytica* strains [70]. As such, this genomic variability can potentially cause variation in transcriptional levels as well as flexibility in transcriptional regulation, contributing to difference in virulence. To prove this assumption, I correlated the obtained RNA-Seq data with the genomic data previously published. The association between sequence polymorphisms and transcriptional variation among the strains was explained in Chapter 2. Also, the impact of copy number variation on gene expression levels was demonstrated in Chapter 4.

In addition, I hypothesised that expression of virulence-associated genes might be regulated by sRNAs, potentially miRNAs, leading to different clinical phenotypes among strains. The miRNA, a small non-coding RNA with 21-23 nucleotides (nt) in length, is well-conserved among eukaryotic organisms and functions via complementary base-pairing with messenger RNA (mRNA) molecules, resulting in mRNA degradation and translation repression. As mentioned before, the bioinformatics-based predictions of novel miRNA candidates have been previously reported in HM-1:IMSS strain but functional studies towards the virulence regulation are still lacking [94,95]. Also, putative miRNAs have been recently identified in deep-branching unicellular flagellate parasites, e.g. *G. lamblia* and *T. vaginalis* [97-100]. So, it is possible for the presence of miRNA regulation system in *E. histolytica*. Hence, in Chapter 5, I designed the experiments using size-selected small RNA-Seq to investigate whether miRNAs could be found and play a role in regulating the parasite transcriptome. The overall methodology used in this study was schematically outlined in Figure 1.6.

Thus, the main aims of this present study are to comprehensively study the comparative analysis of the whole and small RNA transcriptomes amongst nonvirulent and virulent strains of axenically cultured *E. histolytica* trophozoites as well as to integrate such transcriptomic findings with the genomic data for better understanding of the pathogenesis and virulence in amoebiasis.

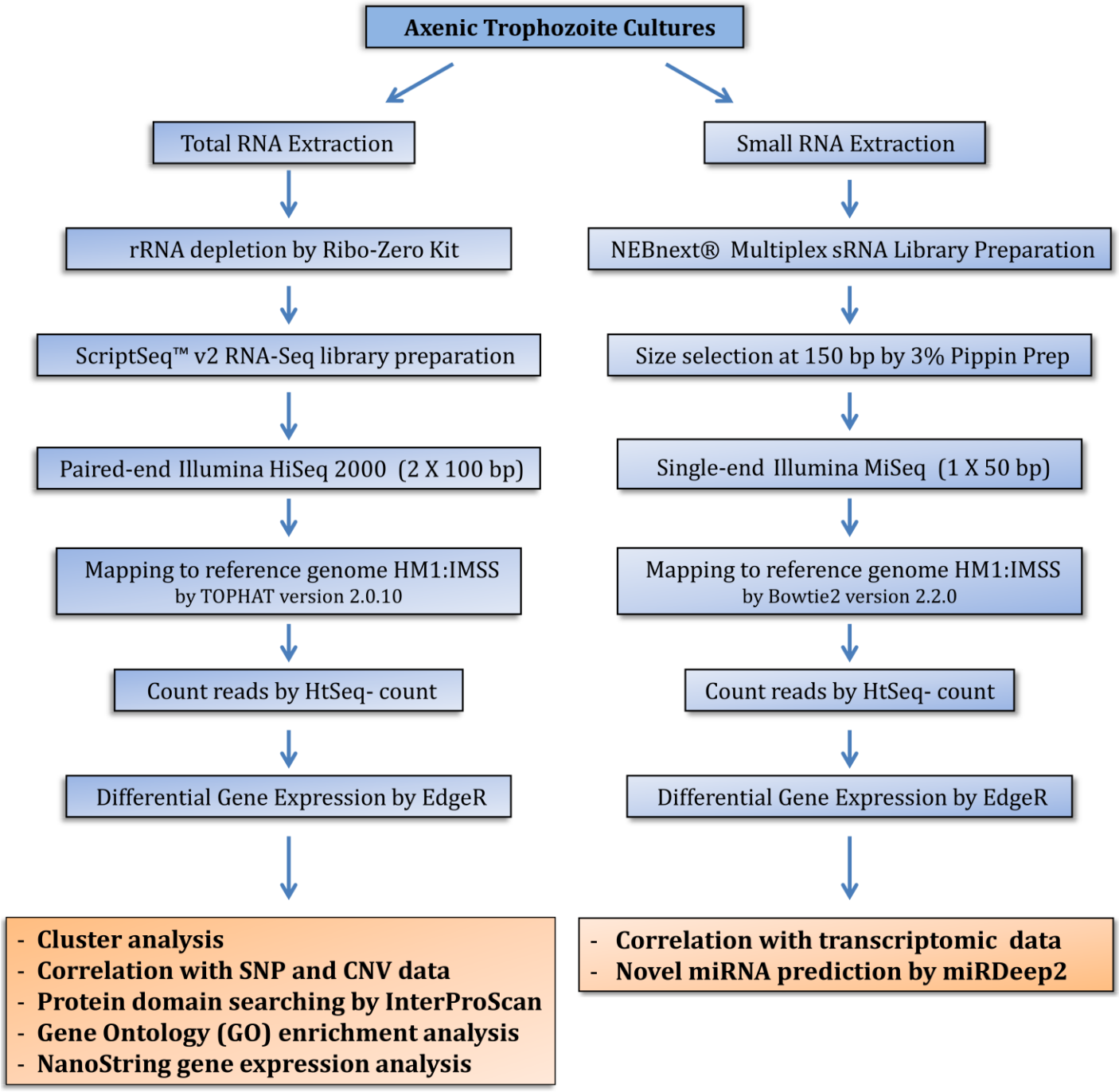


Figure 1.6: Overall methodology for transcriptomic characterisation of virulence in *E. histolytica* in this present study.

Chapter 2: Exploration of the transcriptomes in the four laboratory-adapted strains of *E. histolytica* to identify genes responsible for virulence

2.1 Introduction

After new assembly and reannotation of the *E. histolytica* genome republished in 2010, the big genomic data of this parasite have revolutionised and made both transcriptomic and proteomic analyses easier and more comprehensive than ever [25]. As reviewed in the previous chapter, many microarray-based expression profilings have characterised both *in vitro* and *in vivo* molecular differences between *E. histolytica* strains in relevance to their differential virulence [74,76,77,82]. Generally, hybridisation-based transcriptomic method, i.e. DNA microarray, requires the synthesis of fluorescently labelled probes to detect expressed transcripts. Even though this high-throughput method can be applied to quantify the gene expression levels on a genome-wide scale, there are still some limitations in transcript detection. The major disadvantages are high background noises due to cross hybridisation as well as a narrow dynamic range of quantification caused by signal saturation [101,102].

RNA-Seq technology has recently been developed to explore transcriptomic profiles in the samples of interest. Basicly, the next generation DNA sequencing technologies that have been previously developed, e.g. SOLiD™ system (Applied Biosystems), 454 pyrosequencing (Roche) and Illumina sequencing, can be applied for RNA-Seq to sequence a population of cDNA fragments reverse-transcribed from RNA and provide high-throughput data for downstream analysis [103-108]. This application provides many benefits over other transcriptomic profiling methods [109]. Firstly, RNA-Seq can analyse gene expression without limitation of probe design or reliance on genomic reference, required for hybridisation-based methods [104,109]. Secondly, there is very low background signal in RNA-Seq since most obtained DNA sequence can be mapped to the reference sequence.

Essentially, RNA-Seq has a very high sensitivity and broad dynamic range greater than 9,000-fold, indicating its capability to precisely measure the expression levels of both rare and abundant transcripts [109]. In principle, the number of sequenced transcripts would represent the level of gene expression. In contrast to RNA-Seq, DNA microarrays show low sensitivity and have much limited dynamic range not greater than 150-fold as a result of signal saturation, therefore the microarray method is not appropriate in detecting rare or very highly expressed transcripts [109].

Besides a purpose of transcript quantification, RNA-Seq can be used to reveal novel transcripts, novel isoforms, alternative splicing, alleleic expression as well as RNA editing [110]. Typically, RNA-Seq requires low amount of initial RNA for library preparation and also shows high reproducibility [103,106]. Ultimately, sequencing cost in the post-genomic era has continuously decreased. Taken together, such above advantages make RNA-Seq very popular for current transcriptomic researches.

For whole transcriptomics using RNA-Seq, large-sized RNAs such as poly(A)$^+$ RNA need to be fragmented into smaller sizes prior to steps of reverse transcription and amplification to get highly qualified reads with high Phred quality score [109,110]. Raw short reads obtained from RNA-Seq must be trimmed for adaptor sequence and filtered out for low-quality reads. Then, high-quality reads after quality assessment will be used for transcriptome assembly and alignment to the reference sequence to estimate their abundance. Bioinformatic tools, e.g. Cufflinks, R Bioconductor (edgeR, DEseq), etc., have been applied to analyse the mapping results for differential expression analysis [111-115]. Conclusively, the key superiority of this sequencing-based method is its capability to reveal whole transcriptomic profile of the interested cells or tissues with quantitative and accurate measurements.

As reviewed in Chapter 1, previous studies mainly focused on transcriptomic differences between two well characterised *E. histolytica* strains, i.e. nonvirulent Rahman and virulent HM-1:IMSS. However, clinical case reports of *E. histolytica* infection revealed difference in virulence and various pictures of clinical manifestations. Such broad spectrum of disease severity likely reflects the diversity of this parasite in both genomic and transcriptomic levels, resulting in varied phenotypes. Recently, Weedall *et al*., 2012 revealed genome diversity among axenically cultured *E. histolytica* strains, suggesting the differences in transcriptomic profiles among such re-sequenced strains [70]. Therefore, it is hoped that integration of knowledge in genomics and transcriptomics would give us fresh understanding in amoebic virulence better than before.

In this chapter, I approached the whole transcriptomic analysis of the four axenically cultured *E. histolytica* strains by using the Illumina HiSeq RNA-Seq technology to explain the molecular basis of transcriptomic differences among *E. histolytica* strains. Additionally, protein domain signatures of genes with transcriptomic modulation were characterised to reveal the biological functions and implications towards virulence and pathogenesis. Furthermore, GO enrichment analysis and comprehensive summarisation by the REVIGO software were performed to display biologically relevant interconnections of modulated gene ontology terms in the transcriptomes of virulent strains. Therefore, these functional

and global transcriptomic analyses can provide us the new insights into the molecular and evolutionary basis of virulence and pathogenesis in *E. histolytica* infection.

2.2 Materials and Methods

2.2.1 Strains of *E. histolytica* used in this whole transcriptomic study

Four strains of *E. histolytica* detailed in Table 2.1 were available in my laboratory and used for this transcriptomic study [70]. The experiment was designed in triplicate to prevent bias of measurements so 12 samples in total (3 replicate lines for four strains) were collected for analysis. Firstly, *E. histolytica* trophozoites were axenically cultured from cryopreserved stocks kept in liquid nitrogen. Trophozoites were then subcultured in 13 ml tube of LYI-S-2 medium twice per week. After inoculation in LYI-S-2 medium, the trophozoites were cultured at 36 ^{0}C and evaluated for the mid-log phase growth (50-70 % confluency) under Nikon Diaphot 200 inverted microscope. When they had reached the appropriate confluency at 60 hrs of culture, the mid-log phase trophozoites were collected by centrifugation and washed in phosphate buffer saline (PBS) solution. Then, these harvested trophozoites were immediately used for total RNA isolation.

Table 2.1: *Entamoeba histolytica* strains used in this study, including country of origin, year of collection and clinical manifestation. For Rahman, the UK (*) patient was a sailor, so the infection was unknown in origin, probably contracted elsewhere. The PVBM08B strain was isolated from an Italian (**) who had a travel history possibly from Liberia or Columbia.

Strain	Country of origin	Year of collection	Clinical manifestation
Rahman	United Kingdom*	1964	Asymptomatic
HM-1:IMSS	Mexico	1971	Intestinal amoebiasis (Amoebic liver abscess in inoculated hamster)
IULA:1092:1	Venezuela	1992	Intestinal amoebiasis
PVBM08B	Italy **(colonic biopsy)	2007	Intestinal amoebiasis

2.2.2 Total RNA isolation, quality assessment and ribosomal RNA depletion

Total RNA was extracted using the Trizol® plus RNA purification kit (Invitrogen, USA). The RNA integrity number (RIN) was verified to determine the quality of each sample using an Agilent 2100 Bioanalyser with the Eukaryotic RNA Pico chip (Agilent Technologies, USA). Qualified undegraded samples with RIN score greater than or equal to 6.0 were used for further steps. Ribosomal RNAs (rRNAs) were then removed by using the RiboZero™ magnetic gold rRNA removal kit. The rRNA-depleted samples were rechecked by the Eukaryotic RNA Pico chip to ensure that at least 95 % depletion of 18S and 28S rRNA species was successful. These samples were also assessed in quantity using the Qubit® fluorometric assay (Invitrogen). Then, the processed RNA samples were kept at -80 ^{o}C until used for RNA-Seq library preparation.

2.2.3 ScriptSeq™ v2 RNA-Seq library construction

Library construction is performed using total rRNA-depleted RNA as a template for reverse transcription to provide set of cDNA fragments with average size of 200-500 bp. In this work, I used ScriptSeq™ v2 RNA-Seq library preaparation kit (Epicentre, USA) for constructing adaptor-tagged RNA-Seq library, ready for deep sequencing technology, i.e. Illumina sequencing HiSeq 2000.

RNA-Seq library construction was performed following carefully the manufacturer's protocol (Epicentre), as overviewed in Figure 2.1. Briefly, 50 ng of rRNA-depleted sample obtained from the previous step was fragmented using the RNA fragmentation solution and annealed with random-sequence primers to synthesise cDNAs with 5' tagged end. After removal of RNA, terminal-tagging oligo (TTO) was used for annealing to the 3' end of the cDNAs to act as a template for cDNA extension by DNA polymerase. Then, di-tagged cDNAs were purified by Agencourt® AMPure XP system (Beckman Coulter, USA). The purified di-tagged DNAs were amplified by 15 cycles of PCR reaction using PCR primers specific to the tagging sequences. For this study, different index primers were used as reverse primers individually for each library. Then, obtained cDNA libraries were purified using Agencourt® AMPure XP magnetic beads prior to Qubit® fluorometric quantitation. Finally, cDNA libraries constructed from each strain were run in a High Sensitivity DNA chip (Agilent Technologies) to check the overall profile with range from 100 to 3,000 bp and pooled together in equimolar fractions for paired-end sequencing (2x 100 bp) on one lane of the Illumina HiSeq 2000 platform with version 3 chemistry at the Centre for Genomic Research (CGR), University of Liverpool.

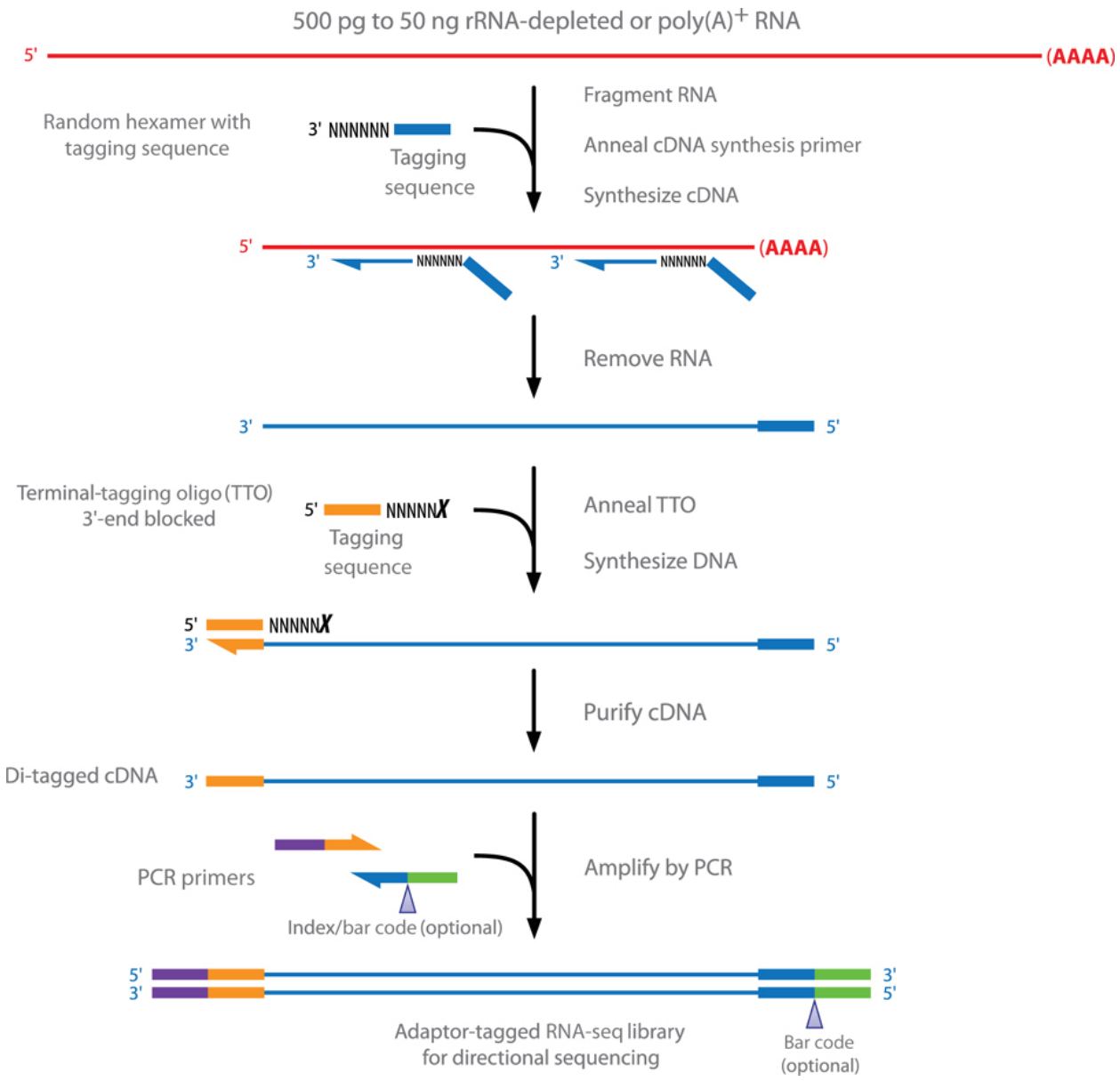


Figure 2.1: Comprehensive workflow of the ScriptSeq™ v2 RNA-Seq library preparation. (available online at https://www.epicentre.com).

2.2.4 Bioinformatics Pipeline

I. Read processing and quality assessment of the raw sequence data

For the whole transcriptome library, raw data were obtained in the form of Fastq formatted files. Cutadapt version 1.2.1 with option '**-O** 3' was used for trimming the 3' end of any reads matched with adaptor sequences for 3 bp or greater [116]. Then, bases with low quality scores were trimmed using Sickle version 1.200 with a minimum window quality score of 20. After quality trimming, short reads less than 10 bp were removed. For the paired-end library, if both paired–end reads passed the filter, they were designated as R1 and R2 reads for forward and reverse reads, respectively. The reads where one read was filtered out due to poor sequence quality or adaptor contamination were included as R0 reads. The total number of raw reads as well as the percentage of trimmed reads were summarised in Figure 2.2 and Table 2.2, respectively. The distribution of trimmed read lengths in all library samples was illustrated in Figure 2.3.

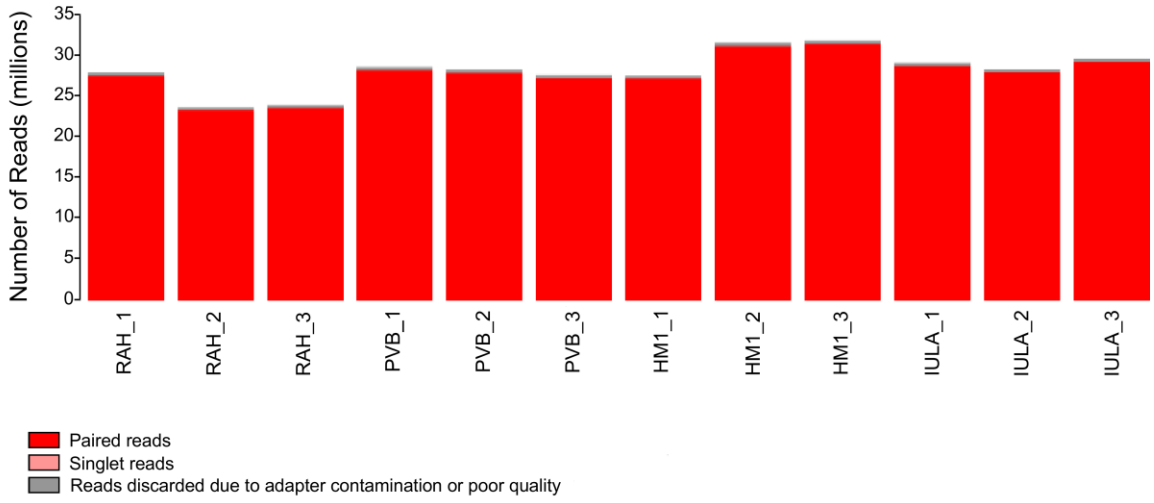


Figure 2.2: The total number of reads in millions retrieved from each library of the four strains.

Table 2.2: Summary of sequence read data before and after adapter removal and low Phred score trimming.

Sample	Raw reads	Trimmed reads	R1/R2 reads	R0 reads
Rahman_1	27,628,770	27,441,181 (99.32%)	13,628,234	184,713 (0.67%)
Rahman_2	23,372,776	23,206,434 (99.29%)	11,521,192	164,050 (0.70%)
Rahman_3	23,669,456	23,492,355 (99.25%)	11,659,162	174,031 (0.74%)
PVBM08B_1	28,307,440	28,118,858 (99.33%)	13,967,785	183,288 (0.65%)
PVBM08B_2	27,996,296	27,786,978 (99.25%)	13,790,563	205,852 (0.74%)
PVBM08B_3	27,327,270	27,131,353 (99.28%)	13,469,506	192,341 (0.70%)
HM-1:IMSS_1	27,249,994	27,071,846 (99.35%)	13,448,192	175,462 (0.64%)
HM-1:IMSS_2	31,279,360	31,068,562 (99.33%)	15,430,699	207,164 (0.66%)
HM-1:IMSS_3	31,543,100	31,328,023 (99.32%)	15,558,614	210,795 (0.67%)
IULA:1092:1_1	28,823,354	28,618,907 (99.29%)	14,210,852	197,203 (0.68%)
IULA:1092:1_2	28,027,418	27,847,442 (99.36%)	13,835,844	175,754 (0.63%)
IULA:1092:1_3	29,324,306	29,130,281 (99.34%)	14,471,195	187,891 (0.64%)

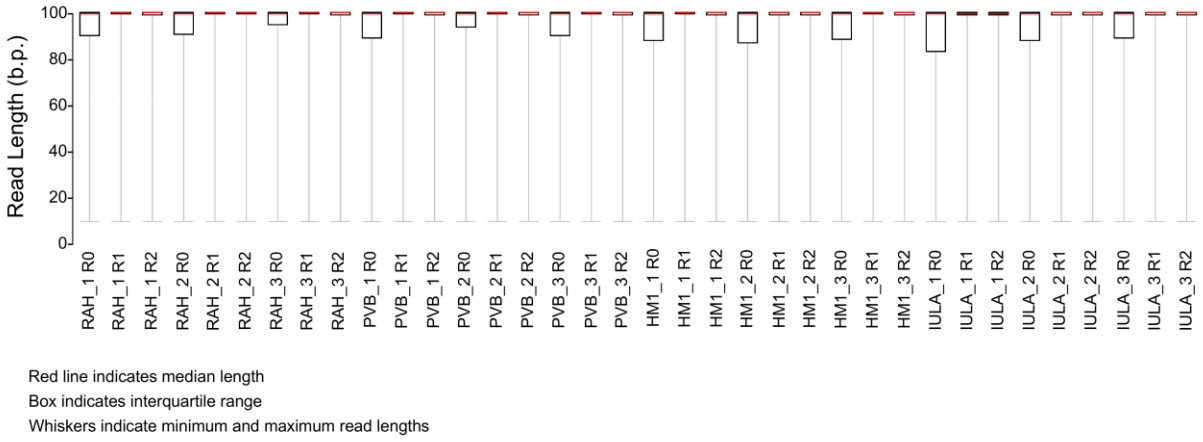


Figure 2.3: Read length distributions after adaptor and low base quality trimming. The forward, reverse and singlet unpaired reads are represented as R1, R2 and R0 reads respectively.

II. Mapping of reads to the reference genome sequence

After getting the quality trimmed reads, TopHat version 2.0.10 (http://tophat.cbcb.umd.edu) was used as a read alignment software to map R1/R2 reads to the *E. histolytica* HM-1:IMSS reference genome sequence (release 2.0, http://AmoebaDB.org/common/downloads/release2.0/EhistolyticaHM1IMSS/fasta/data/AmoebaDB-2.0_EhistolyticaHM1IMSS_Genome.fasta) [26,111,114,117]. The corresponding AmoebaDB-2.0_EhistolyticaHM1IMSS.gff file was used to annotate a total of 8,333 genes in the genome [26].

TopHat was set for paired-end data with following parameters: **-p** <number of threads> 8; **--library-type** fr-secondstrand; **–G** <GTF/GFF3 file> AmoebaDB-2.0_EhistolyticaHM1IMSS.gff; **–r** <mean inner distance> 445. The output 'accepted_hits.bam' file was then used for all downstream analysis.

Alignment statistics were calculated using SAMtools with 'flagstat' option. The number and percentage of total read mapping and uniquely mapped reads are shown in Table 2.3. Additionally, Cufflinks version 2.1.1 (http://cufflinks.cbcb.umd.edu) using the modified annotation file 'AmoebaDB-2.0_EhistolyticaHM1IMSS.exon.only.gff' as the reference was used to calculate the comparable normalised values, i.e. **F**ragments **P**er **K**ilobase of transcript per **M**illion fragments mapped (FPKM) from the dataset [111]. The number and percentage of genes with five different ranges of FPKM values in all four strains were summarised in Table 2.4 and Figure 2.4.

Table 2.3: **Summary of number and percentage of total and uniquely read alignments to the *E. histolytica* HM-1:IMSS reference genome using TopHat software version 2.0.10.**

Sample	Number of total reads generated	Number of properly paired reads mapped to reference	Percentage of total read mapping	Number of uniquely mapped reads	Percentage of uniquely mapped reads
Rahman_1	29,346,366	25,963,308	88.47 %	22,441,712	76.47 %
Rahman_2	24,930,442	21,902,690	87.86 %	19,325,670	77.52 %
Rahman_3	25,262,009	22,290,760	88.24 %	19,110,836	75.65 %
PVBM08B_1	26,526,824	22,595,244	85.18 %	19,881,804	74.95 %
PVBM08B_2	29,509,371	25,086,758	85.01 %	22,501,128	76.25 %
PVBM08B_3	26,157,558	22,371,072	85.52 %	19,702,578	75.32 %
HM-1:IMSS_1	28,331,798	25,174,890	88.86 %	20,578,122	72.63 %
HM-1:IMSS_2	35,210,633	31,709,340	90.06 %	25,319,948	71.91 %
HM-1:IMSS_3	32,887,888	29,029,556	88.27 %	24,237,370	73.70 %
IULA:1092:1_1	30,608,590	26,960,566	88.08 %	21,472,782	70.15 %
IULA:1092:1_2	24,874,054	21,613,244	86.89 %	17,959,660	72.20 %
IULA:1092:1_3	26,626,517	23,072,116	86.65 %	19,170,672	72.00 %

III. Differential gene expression (DGE) analysis

Differential gene expression (DGE) analyses were carried out using the processed HTSeq-count read data and the generalised linear model (glm) approach of the edgeR package [117]. Firstly, the alignment file 'accepted_hits.bam' of each library sample was converted from BAM to SAM file, using SAMtools with 'view -h' option. Using the accepted_hits.sam file as an input, HTSeq-count (release 0.6.1, http://www-huber.embl.de/users/anders/HTSeq/doc/count.html) was applied to count only raw sense reads per gene with following parameters: **-m** <mode> intersection-strict; **-i** <id attribute> Parent; **-t** <type> exon; **-s** <stranded> yes [119]. The HM-1:IMSS genome annotation file (release 2.0, AmoebaDB-2.0_EhistolyticaHM1IMSS.gff file) was used to count raw reads aligned to each gene [26]. The raw sense reads for each gene obtained from the HTSeq-count data would be used for DGE analysis.

Then, the edgeR Bioconductor package software (available at http://www.bioconductor.org/packages/release/bioc/html/edgeR.html) was used to explore differential expression profiles in pairwise comparison between two strains of *E. histolytica* [115]. Briefly, genes with zero HTSeq-count in all samples were filtered out and then the subset of only expressed genes in each library was analysed for both 'within-group' and 'between-group' variations in form of pairwise scatterplots as shown in Figures 2.5 and 2.6, respectively. A sample correlation heatmap was construced based on Pearson's correlation coefficients (*r*) to reveal transcriptomic variability within a sample group and between different sample groups in form of colour spectrum as shown in Figure 2.7. Also, the two-dimensional principal component analysis (2D-PCA) plots were constructed using the $\log_2$-transformed FPKM and HTSeq-count values to estimate the overall transcriptomic variation among all 12 library samples as demonstrated in Figure 2.8.

Normalisation factor was calculated for each library using calcNormFactors function to correct for differences of library sizes among samples after filtering all zero count. Then, the dispersion plot was constructed by fitting to a negative binomial (NB) model to show the values of common, trended and tagwise dispersions of all genes among all libraries as shown in Figure 2.9. Tagwise dispersion specific to each gene was applied for significance testing in differential gene expression analysis.

A model matrix was constituted with six pairwise contrasts as follows: Rahman vs PVBM08B; Rahman vs HM-1:IMSS; Rahman vs IULA:1092:1; PVBM08B vs HM-1:IMSS; PVBM08B vs IULA:1092:1; HM-1:IMSS vs IULA:1092:1. The estimated $\log_2$-transformed values of fold change ($\log_2$FC) for all genes in each contrast were determined for differential

expression using a likelihood ratio (LR) test [120]. *P*-values calculated for each gene were corrected for multiple comparisons using the false discovery rate (FDR, Benjamini-Hochberg) method [121]. Differentially expressed genes were considered statistically significant when an FDR-adjusted *P*-value is less than 0.05. The log_2FC values were plotted against the average expression levels, represented by log_2-transformed values of count per million mapped reads (log_2CPM) as shown in Figure 2.10. The distribution of *P*-values for each contrast was shown in Figure 2.11. The number of significantly DE genes with upregulation and downregulation in each contrast was summarised in Table 2.5.

Venn diagrams were constructed to show the numbers of upregulated and downregulated genes which were exclusively found in one strain or overlapping between strains as shown in Figures 2.13-2.16. The most frequent functionally annotated transcripts with upregulation and downregulation in the three virulent strains were listed in Tables 2.6 and 2.8, respectively. Common DE genes with upregulation and downregulation in the three virulent strains were categorised, based on their functional categories, as summarised in Tables 2.7, 2.9 and Figure 2.17. Also, the modulated transcripts with their functional gene annotations in all three virulent strains were listed in Table 2.10. The numbers of upregulated and downregulated functional genes with absolute log_2FC ≥ 2 in each strain were detailed in Appendix Tables 1.1-1.7 and 2.1-2.7, respectively.

To analyse the pattern of transcriptional differences across the strains, hierarchical clustering analyses were performed using the R script 'mymkheatmap', generously provided by Dr. Yongxiang Fang, a biostatistician of the Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool. All 7,024 DE genes showing significant expression differences in at least one or more contrast pairs from the previous DGE analysis were enrolled for heatmap construction. The package 'fields' installed from the bioconductor website (http://bioconductor.org/biocLite.R) was applied to cluster these 7,024 DE genes based on their relative expression pattern (log_2FC) across 6 contrast pairs as illustrated in Figure 2.18 [122]. Then, the 6th cluster with 98 DE genes showing remarkable fold change differences across contrasting pairs were further categorised into 5 subclusters as shown in Figure 2.19. The functional annotations, number of genes and AmoebaDB_IDs of these 98 DE genes were listed in Table 2.11.

In order to test the hypothesis that sequence divergence is associated with transcriptional variation between strains, the numbers of total SNPs found in 98 DE genes, retrieved from the previous 6th cluster, across all strains (see Appendix Table 3) were plotted against their transcriptional variability represented by log_2-transformed values of the ratio of maximum FPKM and minimum FPKM observed in the four strains, as shown in

Figure 2.20. Also, the correlation analysis between nucleotide polymorphisms and transcriptional variation of such 98 DE genes in Rahman relative to HM-1:IMSS was conducted as plotted in Figure 2.21.

IV. Protein domain searching by the InterProScan

Two sets of upregulated (n=1,162) and downregulated (n=997) genes commonly seen in all three virulent strains, i.e. PVBM08B, HM-1:IMSS and IULA:1092:1, relative to Rahman as shown in Venn diagrams (see Figures 2.13 and 2.15) were recruited for scanning their putative functional protein domain or motif. Briefly, protein functional analysis was performed against the Pfam database by the InterProScan program using the FASTA formatted protein sequences of all upregulated or downregulated genes [123]. The 30 most prevalent functionally annotated domains in upregulated and downregulated gene sets were ranked in order based on their frequency of proteins found, as shown in Figures 2.22 and 2.23, respectively.

V. Gene ontology (GO) enrichment analysis and interactive summarisation by the REVIGO software

The upregulated and downregulated gene sets previously used for InterProScan were further investigated for their ontologies and biological implications. Briefly, the upregulated gene set (n=1,162) and downregulated gene set (n=997) were individually applied for GO enrichment analysis in the AmoebaDB website (http://amoebadb.org/amoeba/) to explore overrepresentation or underrepresentation of ontologies in these two gene sets in comparison to the background. Enrichment was considered as statistically significant where an FDR-adjusted *P*-value is less than 0.05. Enrichment analyses for biological process, molecular function and cellular component ontologies were summarised in Appendix Tables 4, 5 and 6 for upregulated gene sets and Appendix Tables 7 and 8 for downregulated gene sets.

Summarisation and visualisation of the previous ontology analyses were performed using the online REVIGO software (http://revigo.irb.hr/) [124]. Then, ontologies for biological process, cellular component and molecular function with associated FDR-corrected *P*-values were analysed by simple clustering algorithm to reduce redundant GO terms as detailed in Appendix Tables 9, 10 and 11 for upregulated ontologies and in Appendix Tables 12 and 13 for downregulated ontologies. Also, the semantic relationships of ontology representatives in multidimensional scaling plot, graph-based visualisation and treemap were illustrated as shown in Figures 2.24-2.33. Cytoscape software version 3.2.0

(http://www.cytoscape.org/) was applied to view upregulated/downregulated biological networks in the REVIGO interactive graph (see Figures 2.25A, 2.27A, 2.29A, 2.31A and 2.33A) [125].

2.3 Results and Discussion

2.3.1 Transcriptomic profiling of the four *E. histolytica* strains from axenic culture

To identify the key differences between three virulent strains (i.e. PVBM08B, HM-1:IMSS and IULA:1092:1) and nonvirulent Rahman strain, total RNA samples were extracted from three biological replicates of the mid-log phase axenicly cultured trophozoites. Quality assessment, rRNA depletion as well as RNA-Seq library construction were conducted. Twelve samples (three replicates for each strain) with different reverse index primers were pooled together for single run sequencing. Paired-end RNA sequencing (2X100 bp) was performed by Illumina HiSeq 2000 platform. FPKM values and raw read counts were calculated using Cufflinks and HTSeq-count softwares, respectively. Then, comparative transcriptomics between strains were explored by DGE tests, GO enrichment analysis as well as protein domain searching, discussed later.

As illustrated in Figure 2.4, the pie charts show the percentage of expressed genes with five different ranges of FPKM values. In general, over 90% of all annotated 8,333 genes are transcribed since the fragments (FPKM > 0) could be mapped against the annotated reference genome. Notably, the majority of genes in all these four strains seem to have leaky expression with FPKM values ranging from > 0 to 50, accounting for approximately 70% of the whole transcriptome. Compared to RNA-Seq experiment done in higher organisms such as *Drosophila melanogaster,* only 9,995 genes on average from total 12,490 expressed genes were expressed in each stage of development (i.e. embryo, larva, pupa and adult), due to the tight transcriptional control [126]. Therefore, this finding suggests that *E. histolytica* possesses weak transcriptional control, resulting in leaky transcription that constitutes about 70% of the transcriptome. However, low abundance transcripts with FPKM less than 50 could be normally detected by RNA-Seq due to its very high sensitivity [109].

Table 2.4: Categorisation of all 8,333 *E. histolytica* genes into five groups based on their expression level. All of 8,333 genes are categorised into 5 groups: inactive, low, moderate, high, and very high expression levels reflected by different FPKM ranges: FPKM = 0, FPKM < 50, 50 < FPKM < 500, 500 < FPKM < 5,000 and FPKM > 5,000, respectively. The number of genes and corresponding percentages in each strain are shown below.

Strain / FPKM	Rahman		PVBM08B		HM-1:IMSS		IULA:1092:1	
	No. of genes	Percentage	No. of genes	Percentage	No. of genes	Percentage	No. of genes	Percentage
0	762	9.14 %	654	7.85 %	542	6.51 %	606	7.27 %
< 50	5,459	65.51 %	5,585	67.02 %	5,849	70.19 %	5,758	69.10 %
50 – 500	1,766	21.19 %	1,744	20.93 %	1,585	19.02 %	1,627	19.53 %
500 – 5,000	323	3.88 %	325	3.90 %	326	3.91 %	317	3.80 %
> 5,000	23	0.28 %	25	0.30 %	31	0.37 %	25	0.30 %

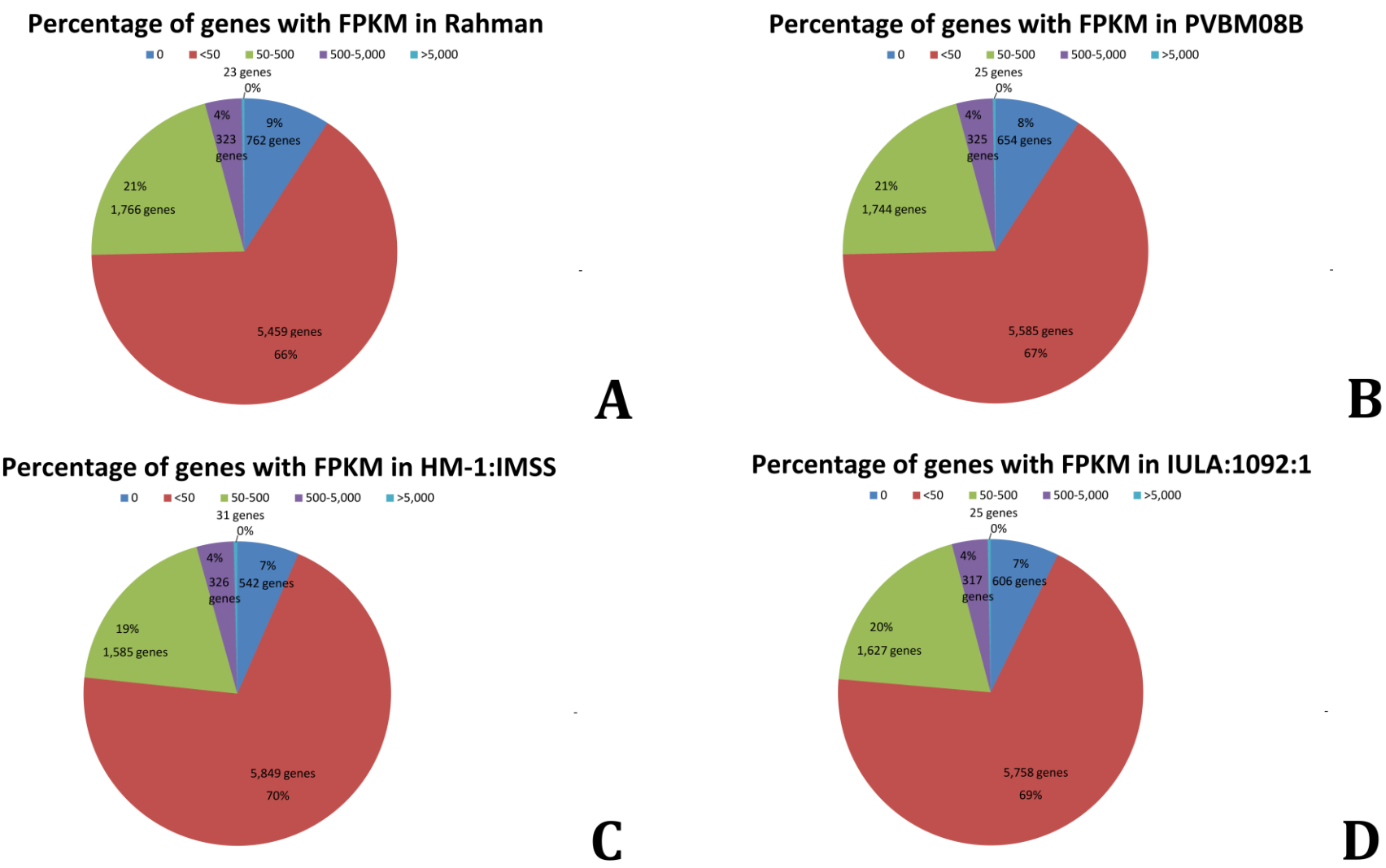


Figure 2.4: Percentage of genes with differential expression levels in Rahman (A), PVBM08B (B), HM-1:IMSS (C) and IULA:1092:1 (D). Most of the genes (approximately 70%) in all these four strains are low in expression with FPKM values ranging from > 0 to 50 as shown above.

2.3.2 Assessment of transcriptional variation in the RNA-Seq data

Firstly, a DGEList object was constructed using the DGEList function of the edgeR package, followed by removing genes which are not transcribed in all library samples to get the best performance of the edgeR when fitting the NB model. Then, filtered RNA-Seq data would be assessed for the variation of gene expression profiles between biological replicates to ensure that their average RNA-Seq data would precisely represent transcriptomic profile in each strain. To assess this 'within-group' variation, scatterplots were drawn using the $\log_{10}$-transformed values of raw read counts per gene to determine a degree of concordance among all biological replicate samples within each strain of *E. histolytica* as illustrated in Figure 2.5. Also, average read count per gene in each strain was calculated and plotted against each other between strains to show the 'between-group' variation as depicted in Figure 2.6. Comparing between Figures 2.5 and 2.6, the RNA-Seq data of the samples within the same strain have less variation than those between the different strains which contains more biological variation due to their dissimilar expression profiles. Strikingly, pairwise differences of average read count per gene between nonvirulent Rahman and other three virulent strains, i.e. PVBM08B, HM-1:IMSS and IULA:1092:1, are more obvious than those compared within these three virulent strains, implying that RNA-Seq has the power to distinguish differences of transcript abundance between nonvirulent and virulent strains of *E. histolytica*.

In addition to the pairwise scatterplots discussed above, a sample correlation heatmap was constructed using Pearson's correlation matrix based on raw read counts of all 8,333 *E. histolytica* genes as shown in Figure 2.7. A bar of colour spectrum represents a Pearson's correlation coefficients (*r*), ranging from 0.941 to 1.000, to show the degree of similarity of expression profiles between the samples. The correlation levels between replicate samples within the same group (reddish brown to orange colour spectrum) are notably higher than those between different sample groups (aqua blue to deep blue colour spectrum), meaning that variations of RNA-Seq data between strains are stronger than those within the same strain. Interestingly, the correlation scores are the lowest in two clusters comparing Rahman and HM-1:IMSS groups, suggesting that variations in relative transcript abundance likely account for different biological behavior of trophozoites between nonvirulent and virulent strains. Taken together, the scatterplots and the Pearson's correlation coefficient-based heatmap reveal that all the RNA-Seq data retrieved from total 12 samples of the four *E. histolytica* strains are of sufficient quality for further analysis.

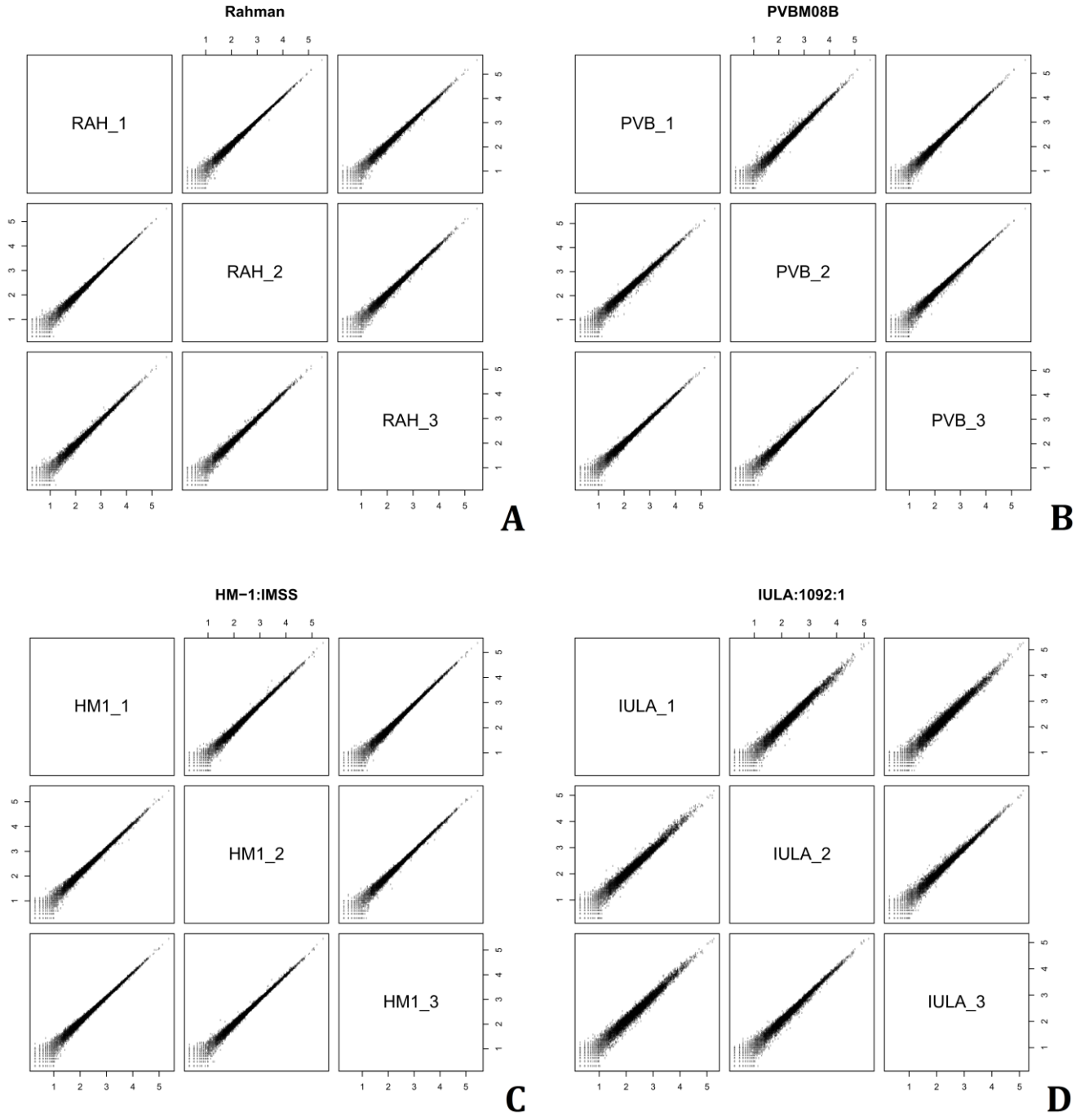


Figure 2.5: **'Within-group' transcriptomic variation among three biological replicate samples in each *E. histolytica* strain.** Both X and Y graph axes represent the logarithm (base 10) of raw read count per gene in each replicate.

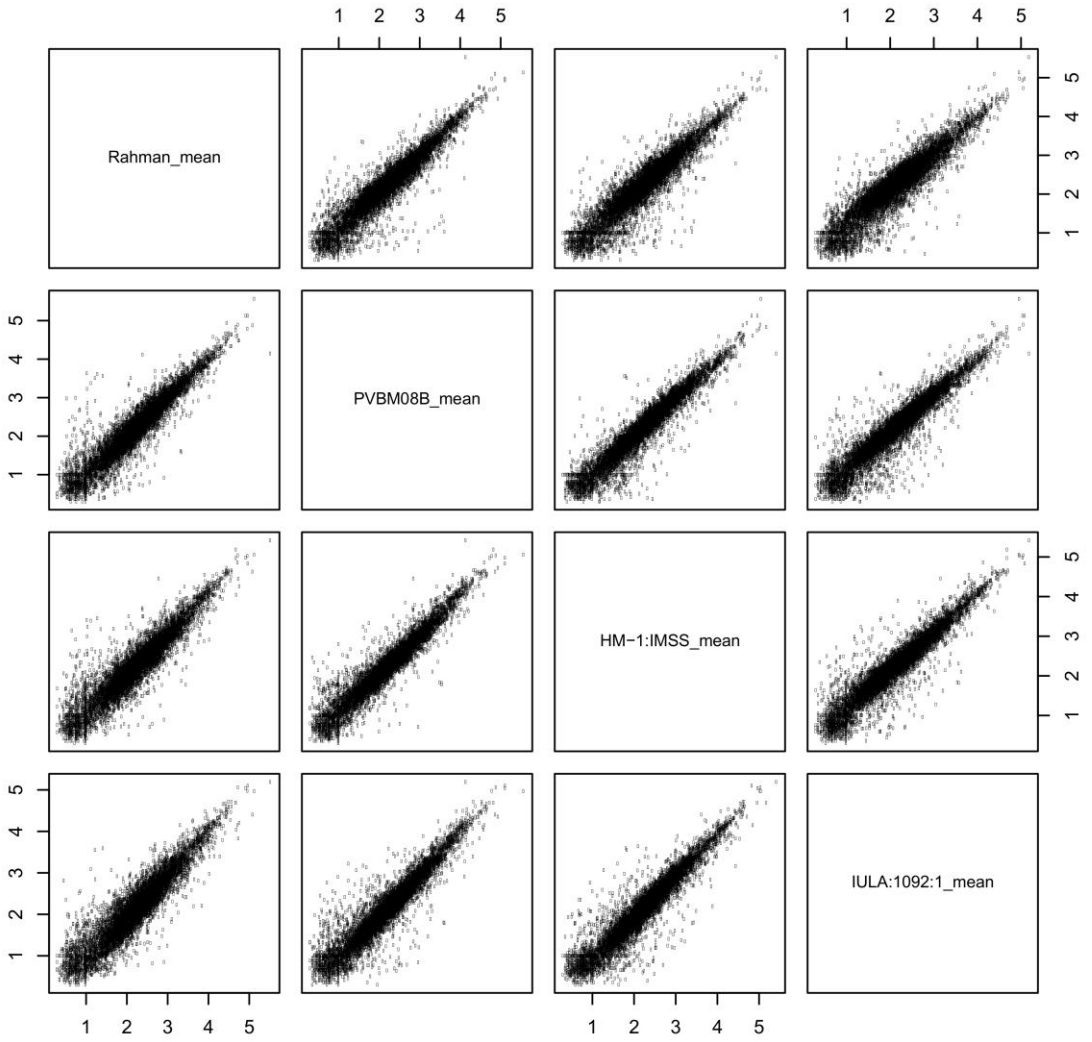


Figure 2.6: **'Between-group' transcriptomic variation among the four *E. histolytica* strains.** Both X and Y graph axes represent the logarithm (base 10) of average read count per gene in each group. In overall, variations between groups of samples are more remarkable than those within the same group previously illustrated in Figure 2.5. Stronger differences in average read count per gene are also observed between nonvirulent Rahman and the other three virulent strains (i.e. PVBM08B, HM-1:IMSS and IULA:1092:1) than between within the three virulent strains.

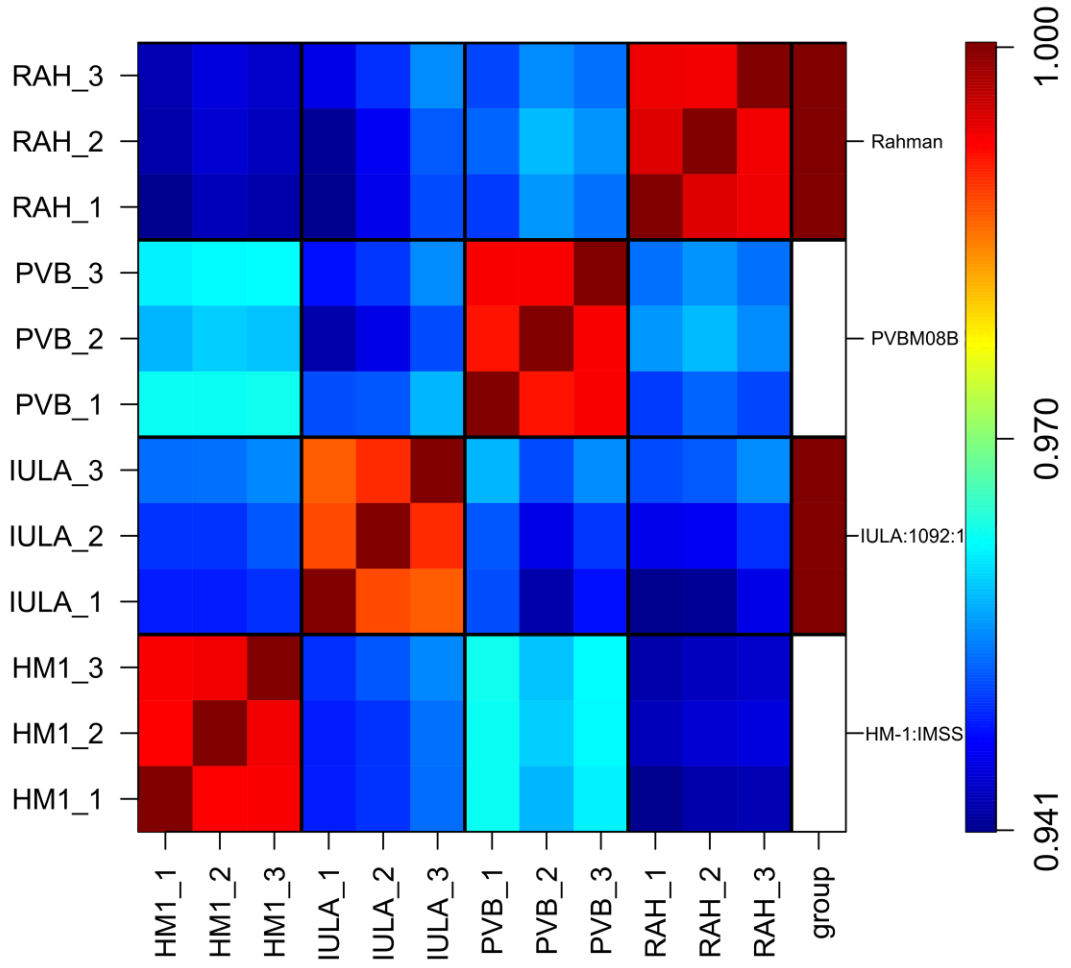


Figure 2.7: **Agglomerative hierarchical clustering of gene expression profiles within and among the four strain groups.** The pairwise correlation patterns are shown in 16 clusters between strains and in 9 sub-clusters among three biological replicates of the same strain. The colour spectrum represents the Pearson's correlation coefficients (*r*) scoring from 0.941 to 1.000. In overall, similarity of whole transcriptomic profiles is more pronounced within the same group than that between different sample groups. Intriguingly, the lowest correlation score could be observed in a pair of comparison between Rahman and HM-1:IMSS groups, suggesting that differences of expression profiles likely link to the differential degrees of virulence.

2.3.3 Principal component analysis reveals the variation map across all 12 samples of 4 *E. histolytica* strains as well as the similarity of transcriptomic profiling between two virulent strains: HM-1:IMSS and PVBM08B

To estimate the overall variation among all 12 samples, the raw HTSeq-count data for all expressed *E. histolytica* genes on a $\log_2$ scale were applied to plot each replicate of the four parasite strains in relation to all other samples as shown in Figure 2.8A. In this work, filtering out of genes with zero HTSeq-count data in all samples was previously performed to keep only expressed genes across samples for downstream analysis. However, HTSeq-count commands in this analysis were set with the option '**-m** intersection-strict' to eliminate any ambiguous read which shows partial alignment to the reference or can be assigned entirely to the two overlapping genes [119]. This HTSeq-count algorithm option was designed to reduce ambiguous reads which can interfere with differential expression analysis.

As such, HTSeq-count would generally provide the read counts lower than the FPKM values reported by other softwares such as Cufflinks. To see the correspondence between the read values generated from these two software packages with different algorithms, PCA plots were constructed individually using these two parameters, i.e. HTSeq-count and FPKM. Therefore, the FPKM values of all 8,333 *E. histolytica* genes previously calculated by Cufflinks were $\log_2$-transformed and then plotted on two dimensional PCA plots as illustrated in Figure 2.8B, in order to determine whether it is congruent with the former PCA plot using the normalised HTSeq-count. Herein, the variation across the samples was obviously demonstrated between the 2nd component and the 3rd component since the 1st component was influenced by differences in library sizes, not providing a clear segregation.

The 1st PCA plot using $\log_2$(HTSeq-count) shows a clustering of biological replicates within the same group as well as a obvious separation across the four strains of *E. histolytica,* indicating that there was no any unusual sample mixed within these expression data whereas the 2nd PCA plot using $\log_2$(FPKM) exhibits overlapping between one PVBM08B library (i.e. PVBM08B_1) and the cluster of three HM-1:IMSS libraries. On both plots, it is interesting that sample groups of PVBM08B and HM-1:IMSS were closely plotted relative to each other, suggesting the similarity of expression profiles between these two strains. Also, it could be interpreted that both Rahman and IULA:1092:1 show strong variation relative to each other since they were widely separated on the plot. Essentially, this comparison can point out that HTSeq-count is more suitable for differential gene expression analysis than FPKM values since this parameter is able to not only cluster the

biological replicates of the same group together but also discriminate between very similar two sample groups, i.e. Rahman and IULA:1092:1, which resemble genetically each other.

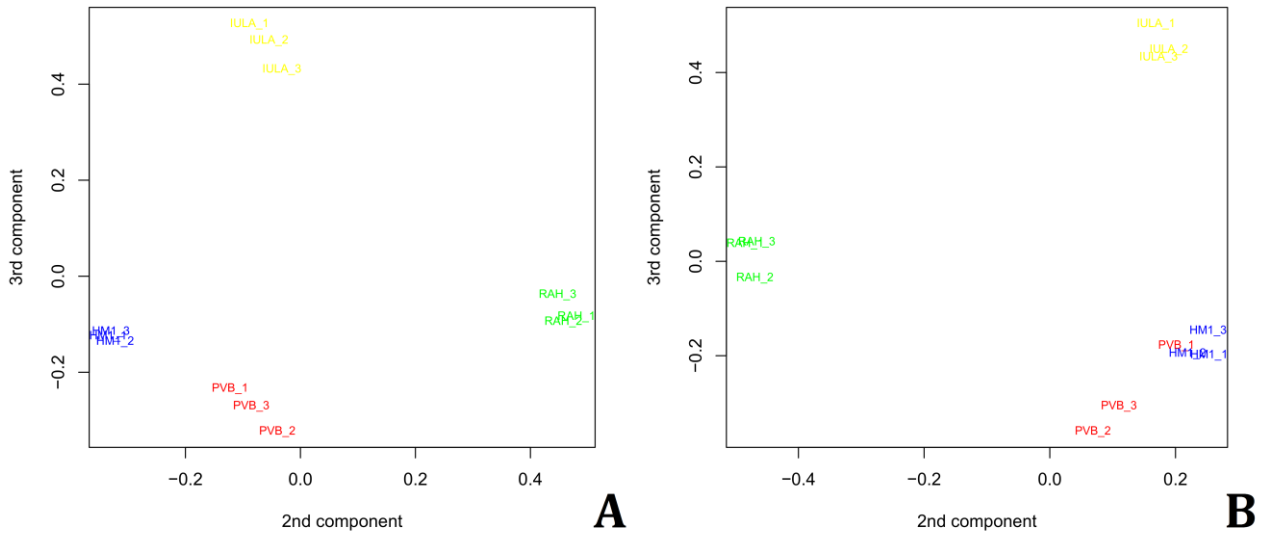


Figure 2.8: **Two dimensional principal component analysis of whole transcriptomes in the four *E. histolytica* strains.** The above two PCA plots (A and B) were generated using $\log_2$(HTSeq-counts) and $\log_2$(FPKM) values of all whole transcriptome library samples, respectively. It is obvious that HTSeq-count data is a better parameter in this study since it shows the clear segregation between PVBM08B and HM-1:IMSS groups, not overlapping as shown in the second plot using FPKM. From the plots, it also implies that the transcriptomes of two virulent strains, i.e. PVBM08B and HM-1:IMSS, are most similar to each other.

2.3.4 Normalisation and estimation of dispersions

To correct any bias due to sequencing depth and differences in library sizes after filtering, normalisation factors were created and applied individually to each library in the dataset using the calcNormFactors function. Prior to performing DGE testing, this normalised dataset with effective library size would be measured for inter-library variation which is an essential parameter for fitting the model and testing for the statistical significance.

Normally, the number of read counts obtained from each gene would likely follow a Poisson distribution [127]. However, all 12 RNA libraries in this study were prepared from a different axenic culture and for this reason the extra variability across samples was introduced in the dataset. This extra variation between the samples is mainly biological. In this case, the NB model was thus applied to analyse this RNA-Seq dataset which has both

biological and technical variations [127]. Additionally, this NB model could reduce the number of false-positive DE genes due to Type I Errors in the DGE analysis.

To assess the extra variability in this NB model, the dispersion parameter was calculated to estimate the degree of variation in tag counts between the libraries. Firstly, it could be assumed that all tags have the same relationship between mean and variance, referring to the common dispersion across all genes. The common dispersion parameter would reflect the average overall variability of the transcriptome across all samples, without regard to gene. However, this common dispersion parameter is not practical since in fact, dispersion of each tag can vary due to different expression levels. So, the tagwise dispersion was estimated using the empirical Bayes method to show gene-by-gene dispersion. Then, the tagwise dispersions for all genes were plotted against their transcript abundance, $\log_2$-transformed values of count per million (CPM) as represented in Figure 2.9.

Obviously, it can be observed in the dispersion plot (see Figure 2.9) that lowly expressed genes show higher tagwise dispersions than highly expressed genes, indicating greater variation in genes with low abundance. Taking into consideration, this finding implies that higher noise in rare transcripts can restrict the power of the RNA-Seq in revealing the real differential expression due to sampling variability, associated with their low sequencing depth, as well as biological variability among the samples.

2.3.5 Fitting the generalised linear model and statistical testing

After normalisation and estimation of the tagwise dispersions completed, the negative binomial models were fitted and then calculated for statistical parameters, i.e. *P*-value and FDR-adjusted *P*-value, using the glmFit and glmLRT functions of the edgeR package, respectively. As shown in Figure 2.11, all histograms for six contrasts show a tall peak of *P*-values approaching zero. This could be interpreted that the majority of enrolled genes in each contrast pair have significant differences in transcript levels between two contrast members. Venn diagrams were constructed to show the number of upregulated or downregulated DE genes, seen in each strain and overlapped between strains. In this study, marked upregulation or downregulation is considered when absolute $\log_2$FC is greater than or equal to 2 or 'more than or equal to 4-fold change'.

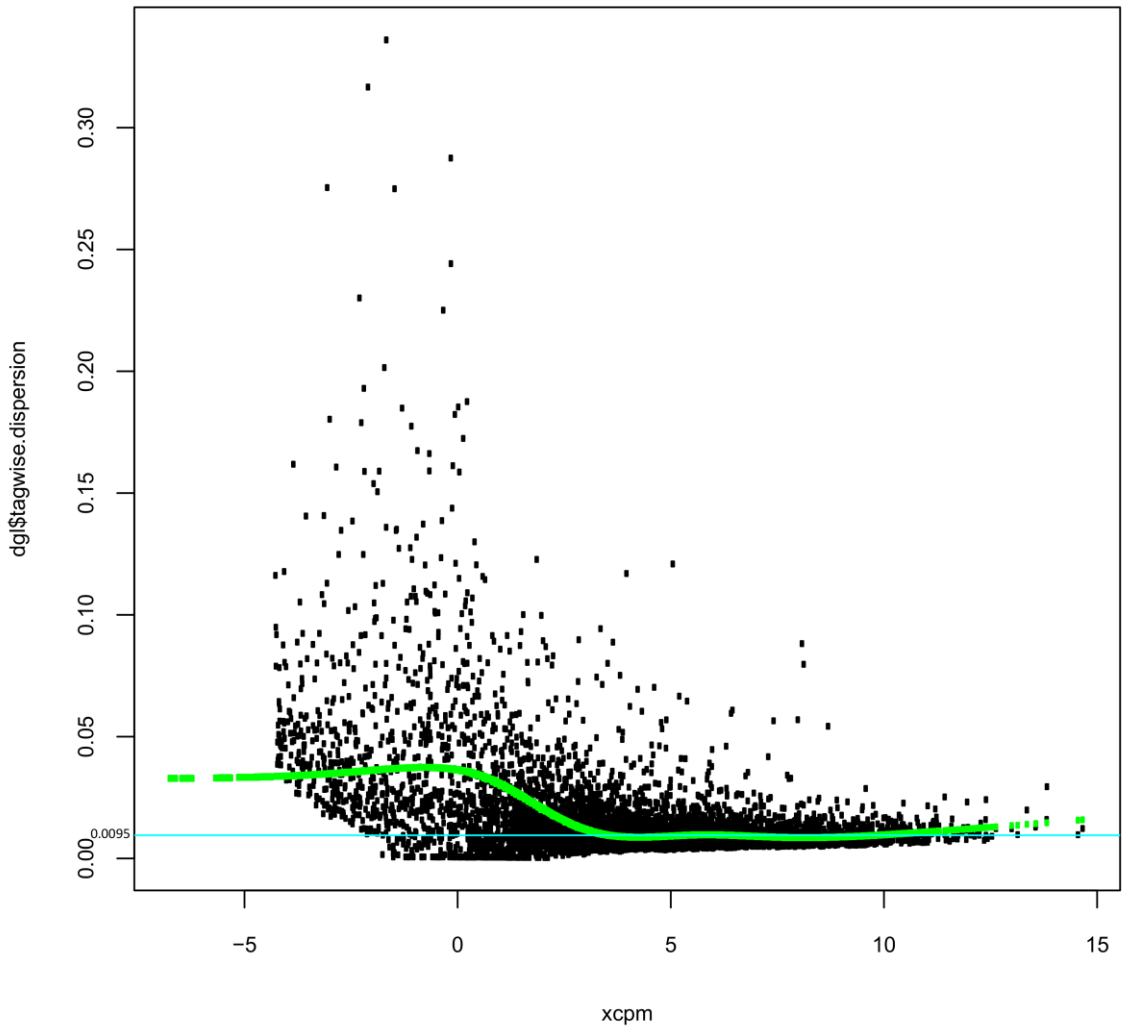


Figure 2.9: **Relationship of inter-library variation for each gene transcript and its corresponding abundance (log_2CPM).** The aqua blue horizontal line represents the common dispersion, equal to 0.0095 across all 12 samples, regardless of gene. The green curve line is the trended dispersion varied by transcript abundance. The black spots represent the gene-by-gene (tagwise) dispersions. Obviously, higher dispersions could be seen in genes with low abundance, implying that the power of RNA-Seq to investigate differential expression in rare transcripts can be affected by low sequencing depths and biological variation among the samples.

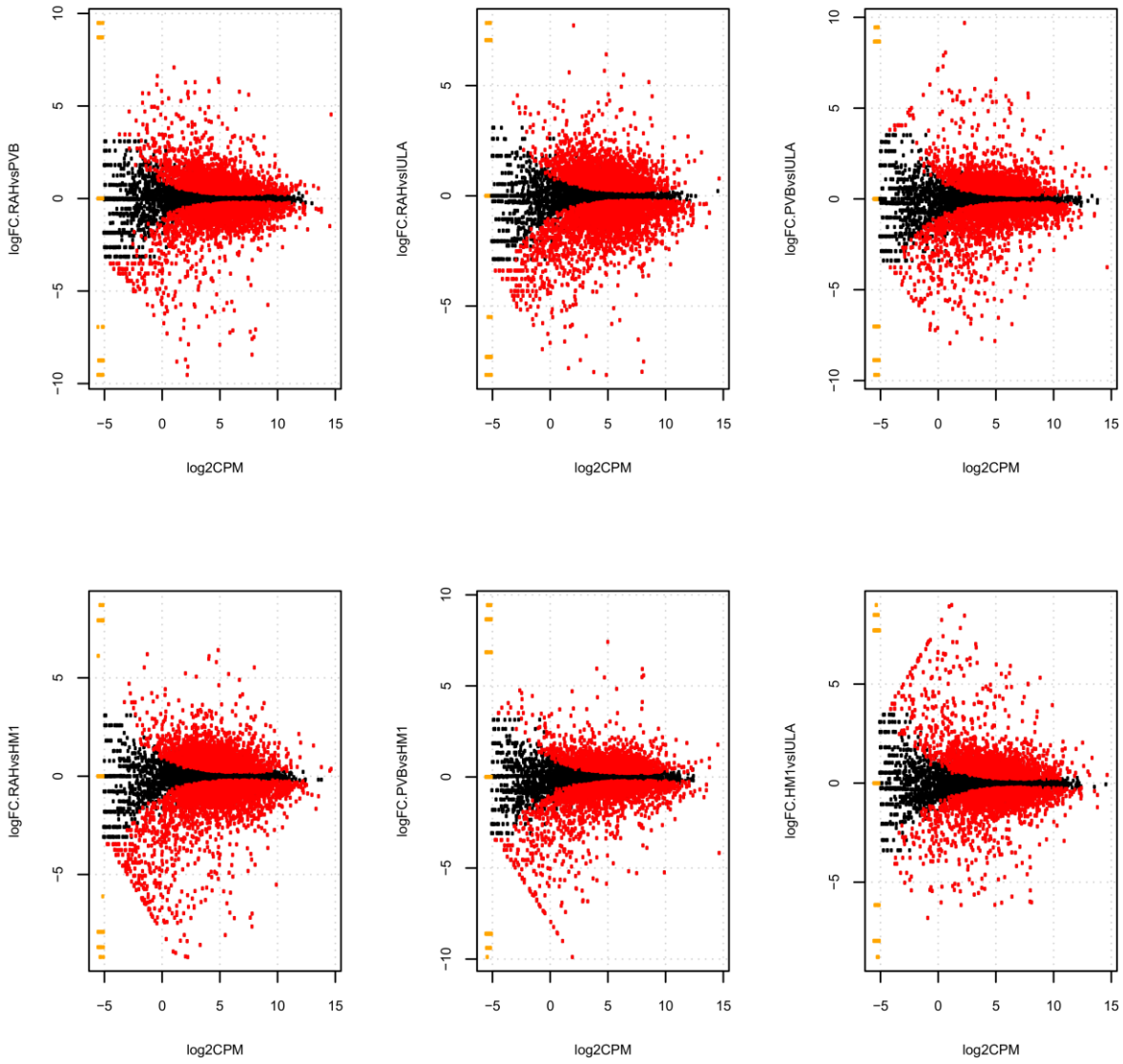


Figure 2.10: **Relationship of the fold change (log_2FC) and the level of expression, i.e. average count per million of mapped reads (log_2CPM), for each contrast pair.** Significant DE genes with FDR-adjusted *P*-value < 0.05 were highlighted in red. Black spots represent non-DE genes. Lowly expressed genes with the value of log_2CPM < -5 are shown in orange spots.

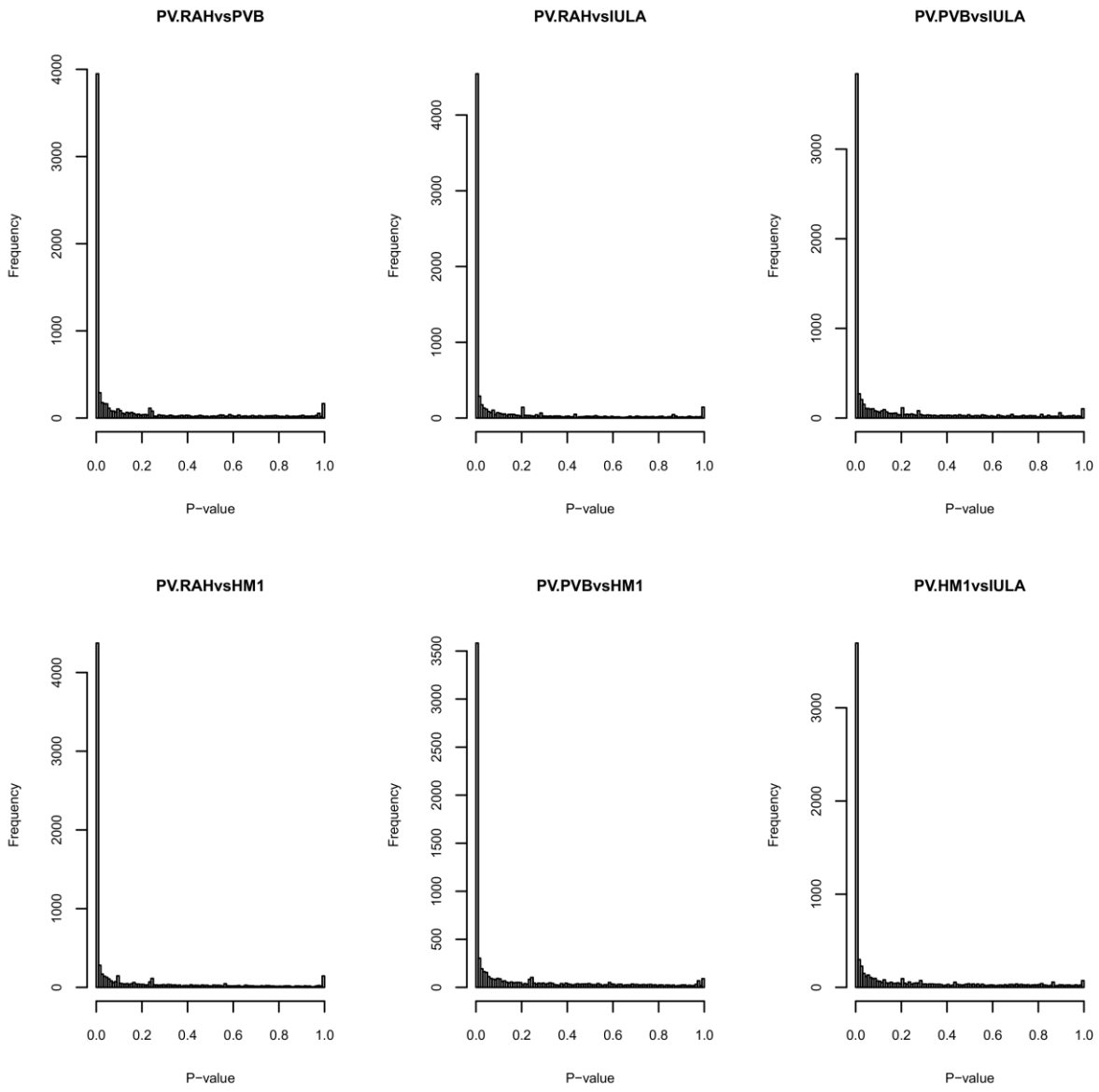


Figure 2.11: Distribution of the *P*-values for each contrast pair. Remarkably, a strong spike of very small *P*-values towards zero could be observed in all histograms, indicating several responding genes between strains. However, FDR-corrected *P*-value less than 0.05 will be considered instead for identifying the significantly DE genes.

Table 2.5: The number of significantly upregulated and downregulated DE genes for each specific contrast. Categories of 'DE, Up' and 'DE, Down' mean the upregulated DE genes and downregulated DE genes with significant FDR-adjusted *P*-value less than 0.05 in the 1st strain of each contrast pair, respectively. DE, Up $\log_2$FC ≥ 2 refers to the 'more than or equal to 4-fold' upregulated genes whilst DE, Down $\log_2$FC ≤ -2 refers to the 'more than or equal to 4-fold' downregulated genes.

Category	Rahman vs PVBM08B	Rahman vs HM-1:IMSS	Rahman vs IULA:1092:1	PVBM08B vs HM-1:IMSS	PVBM08B vs IULA:1092:1	HM-1:IMSS vs IULA:1092:1
DE	4376	4829	5035	3988	4226	4128
DE, Up	2157	2287	2378	1914	2068	2046
DE, Down	2219	2542	2657	2074	2158	2082
DE, Up $\log_2$FC ≥ 2	220	161	161	93	158	254
DE, Down $\log_2$FC ≤ -2	269	503	478	324	326	207

2.3.6 Transcriptomic profiles of the virulent *E. histolytica* strains show a core set of upregulated DE genes involved in host cell killing and mucosal invasion, nucleic acid interaction and oxidative stress response

Using ExactTest as a tool for DGE analysis, a number of significantly DE genes (FDR-adjusted *P*-value less than 0.05) and DE genes with marked upregulation or downregulation (FDR-adjusted *P*-value < 0.05 and absolute value of $\log_2$FC ≥ 2) between two strains of contrast were as listed in Table 2.5. Summary of DE genes with marked upregulation or downregulation in each virulent strain were also listed in Appendix Tables 1.1-1.7 and 2.1-2.7, respectively.

In this study, RNA Seq data of all four strains would be divided into two groups. PVBM08B, HM-1:IMSS and IULA:1092:1 strains are representatives of virulent group whereas the other strain, Rahman, is a nonvirulent group. Venn diagrams were constructed to show overlapping numbers of upregulated DE genes and more than/equal to 4-fold upregulated DE genes (with $\log_2$FC ≥ 2) in Figures 2.13 and 2.14, respectively for these three virulent strains when compared to nonvirulent Rahman. For downregulated DE genes and downregulated DE genes with more than or equal to 4-fold change ($\log_2$FC ≤ -2), Venn diagrams were performed in the same manner as illustrated in Figures 2.15 and 2.16 .

As shown in Figure 2.13, 1,162 upregulated DE genes have significantly higher levels of mRNA expression in all three virulent strains than those in Rahman. When exploring these 1,162 upregulated DE genes, it was found that only 108 genes have $\log_2$FC greater than or equal to 2. For downregulated DE genes, 997 genes are commonly seen in all these three strains but only 23 genes exhibit marked downregulation with $\log_2$FC ≤ -2 as shown in Figures 2.15 and 2.16. Then, these common upregulated and downregulated DE genes could be further assigned to their functional gene categories, i.e. surface-associated, host cell killing and mucosal invasion, oxidative stress response, bacterial killing, nucleic acid interaction, ribosomal structure, protein folding, signaling, protein degradation, miscellaneous and hypothetical, as detailed in Tables 2.7, 2.9 and Figure 2.17. The important virulence-related genes in this study are discussed in following paragraphs.

I. Leucine-rich repeat proteins (LRRPs), BspA-like family

The BspA-like surface protein has been initially characterised in *Bacteroides forsythus.* This surface protein contains a leucine-rich repeat motif (LRRs) functioning as a recognition motif for binding to fibronectin matrix [128]. The BspA-like surface proteins found in *E. histolytica* have unique LRR motifs similar to *Treponema pallidum* LrrA proteins , responsible for host cell adhesion and penetration into deep tissue [25,129,130]. Also, a 50 amino acid N-terminal domain was found to be conserved in 85 members of this protein family. Surprisingly, these LRRPs are localised on the surface of trophozoites but there is no classical membrane-targeting signal sequence present in these proteins. Thus, the conserved N-terminal domain mentioned above may have an important role in non-classical export and anchoring to the membrane [25,130]. After re-assembly of *E. histolytica* genome, Lorenzi *et al.*, 2010 reported that 114 genes were identified as members of this BspA-like protein family in the *E. histolytica* genome [25]. Also, 41 of these 114 genes, accounting for 35 % of this large protein family, were found a high association with TEs, possibly affecting to their expression [25].

In this present study, thirty-one, twenty-one and twenty-three members of BspA-like LRRP gene family were significantly upregulated in HM-1:IMSS, PVBM08B and IULA:1092:1 respectively compared to Rahman, with absolute $\log_2$FC of ≥ 2 as listed in Appendix Table 1.1. After intersecting markedly upregulated members of these three strains, thirteen genes were commonly identified in all these three strains as listed in Table 2.7. This could be inferred that this cluster of thirteen BspA-like surface proteins is likely to have an important role in pathogenesis of invasive amoebiasis. By nature of virulent infection, trophozoites potentially exploit these LRRPs to invade the colonic mucosa, filled with extracellular matrix, i.e. collagen, fibronectin and laminin. Fibronectin binding mediated by such EhLRRPs results in cytoskeleton rearrangement, motility as well as enzyme secretion via G-protein linked receptors and phosphokinase A-dependent signaling [131].

As previously described by Weedall *et al.*, 2012, a total of 512 genes with deep coverage in one or more strains of *E. histolytica* were listed, including 14 members of BspA-like protein family: EHI_002120, EHI_005660, EHI_008340, EHI_049160, EHI_094080, EHI_102380, EHI_113190, EHI_123820, EHI_137910, EHI_147680, EHI_163960, EHI_189090, EHI_191510 and EHI_192600, indicating their putative high copy number [70]. In this study, 3 of 13 upregulated *EhLRRP* genes: EHI_049160, EHI_123820 and EHI_191510 were found to be higher in copy number in these three virulent strains than Rahman (data not shown), indicating these three genes are putative high copy number genes. Therefore,

higher copy number and consequent higher expression levels of LRRP genes could be associated with an increased-virulence phenotype.

II. Galactose/N-acetylgalactosamine (Gal/GalNAc) lectin

As previously reported, virulent trophozoites exhibit certain adhesive molecules, such as Gal/GalNAc lectin, EhSTIRPs and KERP1, on their surfaces to bind and cause the cytopathic effect to the host epithelial cells, resulting in the mucosal invasion [8,11,22,38,50,132]. *E. histolytica* trophozoites undergo contact-dependent cytotoxicity to the host cells via the interaction of a lectin molecule localised on the parasite surface with a Gal/GalNAc-terminal oligosaccharide of the host cell. Ability of the parasite to adhere and kill the host colonic cells, neutrophils, T lymphocytes and macrophages could be impaired by the presence of 50 mM Gal or GalNAc in the culture media [66,133-135]. Also, trophozoites cannot adhere and trigger the cytopathic effect to the Chinese hamster ovary (CHO) cell mutants deficient in N-linked and O-linked glycosylation processes of Gal/GalNAc residues [132].

For the Gal/GalNAc lectin, this complex contains three components: a 170 kDa transmembrane heavy subunit (Hgl), a 31/35 kDa glycosylphosphatidylinositol (GPI)-anchored light subunit (Lgl) and a 150 kDa intermediate subunit (Igl), encoded by different gene families [38,135]. The heavy subunit which is encoded by 5 *Hgl* gene family members with 89-95% amino acid identity contains the carbohydrate recognition domain (CRD) responsible for binding the host's galactose residue [38,136]. The light subunit gene family contains at least 7 *Lgl* genes, showing less conserved 79-85% amino acid identity [38]. Both 31 and 35 kDa light subunits are the dominant isoforms and linked to the CRD of the heavy subunit by disulfide bonds [137]. The role of the light subunit is associated with amoebic virulence by participating in clustering of lectin complexes which is the first key step prior to host cell binding [80,138]. Finally, the intermediate subunit is encoded by two copies of *Igl* genes and non-covalently associated with the Hgl-Lgl lectin heterodimer [78].

As mentioned before in Chapter 1, cDNA differential display revealed lower expression of Lgl1 isoform transcripts in Rahman relative to HM-1:IMSS, indicating that downregulation of *Lgl* expression in Rahman is associated with its decreased virulence [79]. Interestingly, overexpression of the Lgl1 isoform in Rahman could not transform Rahman into the virulent state, suggesting certain regulations in Rahman [139]. Also, dominant negative N-truncated Lgl1 expression or downregulation of Lgl1 by monoxenic cultivation or antisense inhibition in HM-1:IMSS was associated with defective erythrophagocytosis [79,80,140]. However, this present study reveals no marked upregulation of any lectin

subunit gene in all virulent strains but conversely shows more than 4-fold downregulation of the three Gal/GalNAc lectin heavy subunit genes in PVBM08B compared to Rahman: EHI_042370, EHI_077500 and EHI_133900. Moreover, HM-1:IMSS and IULA:1092:1 show more than 4-fold downregulation of the two Gal/GalNAc light subunit genes: EHI_065330 and EHI_159870.

As previously reported by Davis *et al.*, 2007, *Lgl3* expression was higher 22-fold in Rahman compared to HM-1:IMSS, raising the hypothesis for its possible dominant negative mutant-like expression and corresponding reduced virulence [76]. For my RNA-Seq result, it is reasonable to state that downregulation of such two *Lgl* genes (EHI_065330 and EHI_159870) in HM-1:IMSS and IULA:1092:1 may be due to allelic preference between strains. Alternatively, both Hgl and Lgl isoforms upregulated in Rahman may be dominant negative forms, less functional than other isoforms expressed in virulent strains. Also, Katz *et al*., 2002 demonstrated that that overexpression of native *Lgl1* gene in HM-1:IMSS trophozoites transfected by a constructed plasmid has no influence on their virulence [80]. It could be explained that Gal/GalNAc lectin requires the combination of both the heavy and light subunits to form a heterodimer, so overexpression of a *Lgl1* gene caused misbalance in numbers between these two subunit molecules [80]. Based on the basis of its heterodimeric structure, it is possible to explain that the upregulation of particular lectin subunit genes in Rahman compared to each virulent strain in my RNA-Seq study might not represent the real difference in quantity of such heterodimeric lectin molecules.

III. Serine-threonine-isoleucine rich proteins (EhSTIRPs)

Besides the Gal/GalNAc lectin, another important adhesive molecule on the virulent trophozoite surface is a serine-threonine-isoleucine rich protein (EhSTIRP). These surface proteins are encoded by a multigene gene family containing four members: EHI_004340, EHI_012330, EHI_025700 and EHI_073630. Strikingly, most of these protein family members exhibit the unusual feature of expression profiles between life cycle stages and between strains of parasite. EhSTIRPs were reported to be highly expressed in all virulent trophozoites, i.e. HM-1:IMSS, 200:NIH and the invasive trophozoites isolated from infected colon, and to be not or very lowly expressed in nonvirulent conditions including *E. dispar* and *E. histolytica* Rahman and cystic stage of virulent *E. histolytica* strains [11,74,77,141].

Based on the HM-1:IMSS genomic reference and annotation in the AmoebaDB database version 4.2, three of four members of *EhSTIRP* gene family: EHI_004340, EHI_012330 and EHI_025700 are similar in a very large size of approximately 8 kb in the parasite genome [26]. Interestingly, the other *EhSTIRP* gene EHI_073630 is indeed the

largest annotated gene in the genome with a length of 15.2 kb. All these family members contain a single transmembrane domain and a short portion of 34 amino acid cytoplasmic tail and show very high conservation at their 3' end with greater than 99% nucleotide identity and less conservation at the 5' end with 88-94% identity between isotypes [11].

Relevant to their virulent functions, the cytotoxic ability of EhSTIRPs has been proven by comparing the release of lactate dehydrogenase from damaged Caco-2 colonic monolayer cells treated with wild type HM-1:IMSS and *EhSTIRP* dsRNA-treated HM-1:IMSS trophozoites and found to be drastically reduced to ~55% in *EhSTIRP*-downregulated parasites at all time points [11]. In addition, it was found that *EhSTIRP* dsRNA-treated HM-1:IMSS trophozoites showed the decrease in host cell adhesion compared to the wild type parasite after incubating on ice and washing [11]. Altogether, EhSTIRPs have putative functions in virulent parasites for host cell adhesion and cause subsequent host cell damage.

Expectedly, two *EhSTIRP* gene members in this study: EHI_012330 and EHI_025700 are markedly upregulated with absolute $\log_2$FC ≥ 2 in all three virulent strains whereas another member EHI_004340 shows strong upregulation only in HM-1:IMSS and PVBM08B as reported in Appendix Table 1.1. For these three *EhSTIRP* members, HM-1:IMSS and PVBM08B have much higher transcript levels ($\log_2$FC = 5.10-8.43) than less virulent IULA:1092:1 ($\log_2$FC = 1.72-2.63), inferring that differential *EhSTIRP* gene expression strongly contributes to virulence variability across virulent strains..

Conversely, it is interesting that the largest *EhSTIRP* gene (EHI_073630) was significantly downregulated in all these three virulent strains relative to Rahman with $\log_2$FC = -1.69, -1.52, -0.82 for HM-1:IMSS, PVBM08B and IULA:1092:1, respectively. Based on the recent microarray, Thibeaux *et al.*, 2013 recently reported that *EhSTIRP* transcript EHI_073630 was only significantly upregulated in HM-1:IMSS with fold change = 2.5, FDR-adjusted *P*-value = $2.70e^{-22}$, compared to Rahman in contact with the colon mucus whereas other three members showed significant upregulation in HM-1:IMSS both in culture and upon contact with the human colon [82]. Also, the authors demonstrated the co-expression of other functional transcripts involved in DNA-RNA regulation, cell signaling, stress response, proteolysis, translation-protein maturation, subcellular trafficking, cytoskeleton and biomolecular metabolism, shown to be upregulated solely during host mucosal contact [82]. This could be inferred that the expression of this gene set including this *EhSTIRP* gene EHI_073630 in virulent parasites is not ubiquitous and solely upregulated under the invasive condition, e.g. during contact to the mucus. In other words, it may be stated in the principle of allocation that virulent trophozoites possibly adapt to allocate their limited

energy for their cell division and growth by reducing the production of such nonessential transcripts in the non-enriched axenic culture condition.

Moreover, Weedall *et al.*, 2012 demonstrated that this *EhSTIRP* gene EHI_073630 which locates on scaffold DS571171 exhibits the most polymorphisms in the genome and is distantly related to the other three *EhSTIRP* gene members [70]. Based on SNP data in the AmoebaDB database, EHI_073630 contains the highest SNPs across all strains: total SNPs = 100, nonsynonymous SNPs = 63, synonymous SNPs = 37 and nonsyn/syn ratio = 1.7 [26]. As demonstrated in Figure 2.12, most of the polymorphisms found in this gene show homozygous sequence pattern in each strain. Strikingly, it was found that HM-1:IMSS and its two derived clones (i.e. HM-1A and HM-1B) have distinctive sequence divergence from the other eight strains that resemble each other, consistent with allelic dimorphism found in *Plasmodium* genes encoding merozoite surface proteins [70,142,143]. However, the question raising whether the large sequence divergence in HM-1:IMSS is associated with its downregulation in axenic condition still needs to be further investigated.

For three expressed *EhSTIRP* gene members ubiquitously in virulent strains, it is intriguing that these three gene members are conversely very low in expression in Rahman, implying that gene silencing exists in the nonvirulent parasite. MacFarlane and Singh, 2007 found that most of *EhSTIRP* coding sequences as well as their promoters in Rahman are very similar (≥ 98%) to those found in HM-1:IMSS, suggesting that the possible epigenetic mechanisms such as DNA methylation and histone deacetylation might be responsible for *EhSTIRP* gene silencing in Rahman [11,144-146]. However, no change in *EhSTIRP* expression was observed after treating Rahman trophozoites with a DNA methyltransferase inhibitor, i.e. 5-azacytidine, and a histone deacetylase inhibitor, i.e. trichostatin A [11]. It is interesting that *EhSTIRP* expression could be downregulated in HM-1:IMSS trophozoites transfected with a plasmid with construct encoding dsRNA specific to the highly conserved 3' end but this dsRNA-based silencing reverted to the normal wide type after one year of subculture [11].

Recently, an endogenous RNA interference (RNAi) pathway has been identified for its role in gene silencing in several human parasites including *G. lamblia*, *T. vaginalis*, *T. gondii*, *T. brucei* and *E. histolytica* [83-86]. The RNAi pathway in *E. histolytica* is mediated by a population of 27 nt small RNAs and their partners, Argonaute proteins (EhAGOs). Zhang *et al.*, 2013 demonstrated the presence of abundant 27 nt sRNAs which antisense mapped to the *EhSTIRP* genes (EHI_025700 and EHI_012330) only in Rahman but were absent in HM-1:IMSS [92]. Furthermore, overexpression of a Myc-tagged *EhSTIRP1* construct (EHI_025700) could be achieved in transfected HM-1:IMSS trophozoites but not in

Rahman, suggesting that antisense 27 nt sRNAs likely regulate the expression of these adhesion molecules in the nonvirulent Rahman strain [92]. More details regarding the antisense sRNA-mediated gene silencing in a strain-specific manner will be discussed in Chapter 5.

Figure 2.12: Sequence polymorphism of *EhSTIRP* gene EHI_073630 located on scaffold DS571171. The top three rows represent the reference HM-1:IMSS and its derivative strains: HM-1A and HM-1B, respectively. Across the full length of the gene, HM-1:IMSS and its derivatives exhibit sequence divergence compared to the other strains shown in the lower eight rows: Rahman, 2592100, PVBM08B, PVBM08F, IULA:1092:1, HK-9, MS84-1373 and MS27-5030, respectively. Polymorphic positions are indicated by different colours as follows: black for homozygous positions; grey for heterozygous positions; light grey for positions different from the reference; white for base not available. This figure is reproduced with permission from Weedall *et al.*, 2012 [70].

IV. Cysteine proteinases

In invasive amoebiasis, virulent trophozoites release extracellular cysteine proteinases to degrade the mucus barrier, i.e. MUC2, as well as the collagen and laminin matrix of the colonic epithelium for penetration to the deeper mucosa [8,9]. Additionally, these released cysteine proteinases enable parasites to resist the host immune defences such as secretory IgA and complement-mediated lysis [10,147-148]. Compared to the lysates of noninvasive *E. dispar*, pathogenic *E. histolytica* strains could release 10- to 1,000-fold more proteinases, reflecting their key role in virulence [149]. Initially, Bruchhaus *et al*., 1996 identified six cysteine proteinase genes (*EhCP-A1* to *EhCP-A6*) from a genomic library prepared from the axenic culture and found that only three genes: *EhCP-A1*, *EhCP-A2* and *EhCP-A5* constituted approximately 90% of total cysteine proteinase transcripts [47]. *EhCP-A5* (EHI_168240), a key cysteine proteinase gene for MUC2 degradation, is found as a pseudogene in *E. dispar* [46].

So far, 33 *E. histolytica* genes encoding cysteine proteinases have been identified in the parasite genome, based on their functional annotations in the AmoebaDB database version 4.2 [26]. A number of the cysteine proteinase gene family members have been identified for their differential expression between nonvirulent and virulent strains and between *in vitro* culture and *in vivo* infection [47,74,76,77,82,150]. Interestingly, most of the EhCP gene family members were not expressed in axenic culture condition and the cysteine proteinase activity of the lysates was progressively increased after the inoculation of axenic and xenic trophozoites into hamster livers, strongly suggesting their specific role responsible for the invasive infection and/or completion of the life cycle [10,47,151].

Davis *et al.*, 2007 revealed that a number of EhCPs: EhCP-A4 (EHI_050570), EhCP-A6 (EHI_151440), EhCP-B1 (EHI_117650) were upregulated ~3-fold in HM-1:IMSS relative to Rahman. Additionally, it is worth noting that major cysteine proteinases (EhCP-A1, EhCP-A2 and EhCP-A5) were relatively abundant in both HM-1:IMSS and Rahman strains, and not significantly different in their expression levels between these two strains. Conversely, EhCP-A3 (EHI_159610), EhCP-A7 (EHI_039610) and EhCP-B9 (EHI_181230) were higher expressed in Rahman than HM-1:IMSS [76].

In a recent microarray study comparing gene expression profiles between HM-1:IMSS and Rahman, it was found that there was upregulation of EhCP-A7 in both HM-1:IMSS and Rahman in response to the human colon contact, compared to those in axenic culture. Also, EhCP-A3 (EHI_159160) and EhCP-A8 (EHI_151400) were ubiquitously expressed in Rahman and showed higher expression than HM-1:IMSS both in culture and

during colon contact whereas EhCP-A4 was upregulated in Rahman solely upon contact with the human colon [82].

Consistent to prior studies, my RNA-Seq data reveals that EhCP-A3 was higher expressed in Rahman ($log_2FC = 5.52$) than HM-1:IMSS. EhCP-B8 (EHI_097900) shows high upregulation in these three virulent parasites as listed in Table 2.7. Also, other two cysteine proteinases: EhCP-A7 was considerably expressed in HM-1:IMSS ($log_2FC = 4.91$) whilst EhCP-B6 (EHI_126170) was highly upregulated in IULA:1092:1 ($log_2FC = 3.42$). Taken together, these distinctive expression patterns among the virulent strains suggest that such cysteine proteinases may possess different non-redundant functions [76].

V. AIG1-like family proteins

AvrRpt2-induced gene-1 (AIG1) family proteins, firstly characterised in *Arabidopsis thaliana*, are small GTPases responsible for bacterial resistance in *A. thaliana* [152]. Interestingly, AIG1-like proteins are found in *E. histolytica* and encoded by a large gene family. The AIG1-like family in *E. histolytica* contains 29 members physically distributed in 3 clusters [25]. Of these 29 members, 18 genes are close to TEs, accounting for 62% of physical association with repetitive elements [25]. Since *E. histolytica* trophozoites colonise with the colon microbiome and feed on bacteria, AIG1-like proteins may be responsible for antibacterial activity.

Gilchrist *et al*., 2006 reported the increase of *AIG1* mRNA levels in HM-1:IMSS trophozoites isolated from a murine model of amoebic colitis using an Affymetrix array, indicating a possible important role in defense against intestinal bacteria [77]. Comparative DNA microarray studies by MacFarlane *et al.*, 2006 revealed that 415 genes including AIG1-like proteins and heat shock proteins have significantly lower expression levels in nonvirulent *Entamoeba dispar* SAW760 than in *E. histolytica* HM-1:IMSS [74]. Following this transcriptional difference, it was hypothesised that the association of TEs with these AIG1 genes could enhance the expression levels of these genes, and contribute to the increase of virulence [25]. It was previously reported that EhLINE and EhSINE retrotransposons are organised in clusters especially at synthenic break points and contributing to genomic evolution via rearrangement and amplification [153]. However, the question whether the amplification of this family was promoted by the close proximity of TEs needs to be elucidated [25].

Contrastedly, it was shown in the current data that there were only two AIG1 genes (EHI_176280 and EHI_180390) showing upregulation and two other AIG1 genes (EHI_176590 and EHI_176700) showing downregulation with absolute $log_2FC \geq 2$ in the

three virulent strains as listed in Table 2.7 and 2.9, respectively. These modulated transcripts possibly suggest allelic differences in such large AIG1 gene family. Also, the experiment was designed using axenically cultured laboratory strains which grow without bacteria, so it is possible for such virulent parasites to preferably downregulate such AIG1 transcripts with putative antibacterial function, not essential in the axenic condition [81].

VI. Peroxiredoxins

A major host response in the first stage of infection is the release of NO and ROS by host immune effector cells including neutrophils, monocytes, tissue macrophages and dendritic cells to kill the parasite [13,14]. Trophozoites can overcome this threat by using their surface associated molecules, e.g. peroxiredoxin, SOD and flavin reductase [15-19]. Both SOD and flavin reductase play a key role in the production of hydrogen peroxide (H_2O_2) in the presence of oxygen radicals released from the host immune cells. Peroxiredoxin then counteracts the toxicity of produced hydrogen peroxides by reducing them into water molecule (H_2O) [18,154]. In HM-1:IMSS, peroxiredoxin is a surface-associated molecule co-localised with Gal/GalNAc lectin molecule in the lectin-peroxiredoxin complex at the host adhesion site, and plays an important role in ROS degradation against host oxidant attack [155].

In this study, I found that three peroxiredoxin genes (EHI_145840, EHI_001420 and EHI_123390) were significantly upregulated and commonly found in all three virulent strains but only one single peroxiredoxin gene (EHI_145840) was more than 4-fold differentially expressed in all three virulent strains than Rahman as listed in Table 2.7. Amongst these three virulent strains, it is interesting that seven peroxiredoxin genes were individually upregulated in IULA:1092:1 whereas only single and two peroxiredoxin genes were found in PVBM08B and HM-1:IMSS, respectively as detailed in Appendix Table 1.1. However, it was found that other two peroxiredoxin genes were downregulated in HM-1:IMSS (EHI_114010, $\log_2$FC = -2.21) and PVBM08B (EHI_183180, $\log_2$FC = -3.57) respectively, relative to Rahman.

Based on the HTSeq-count data, it is possible to presume that the virulent parasites preferably upregulate a peroxiredoxin (EHI_145840) as a main isoform to counteract the host response since this isoform (EHI_145840) shows the highest level of raw expressed transcripts relative to other isoforms in all four strains enrolled in this study (data not reported). Additionally, it was observed that normalised HTSeq-count of EHI_145840 in IULA:1092:1 is 123.86, less than those in PVBM08B (1,798.32) and HM-1:IMSS (398.18). This finding is likely to explain that upregulation of other six peroxiredoxin genes in

IULA:1092:1 occurs to provide compensatory transcripts for the parasite survival under the host oxidative stress. In accordance with previous studies, it was reported that peroxiredoxin was more highly expressed in virulent HM-1:IMSS than Rahman in both transcriptomic and proteomic levels [16,73]. This current finding promisingly supports that these three laboratory-adapted virulent parasites have potential to counteract host reactive molecules effectively, resulting in more survival and virulence compared to Rahman.

VII. C2 domain-containing proteins

In *E. histolytica*, calcium ions (Ca^{2+}) have been proven for their globally regulatory functions in many biochemical processes including signaling [156], cell motility [157,158], actin dynamics and phagocytic cup formation [159,160], fibronectin adhesion [161], transcriptional regulation [158,162,163], host cell lysis [164,165] and developmental stage conversion [166,167]. Generally, intracellular Ca^{2+} concentration is controlled by certain calcium-binding proteins containing Ca^{2+} binding domains such as the C2 domain, EF hand motif, grainin [156,168]. The C2 domain possesses 120 amino acid residues responsible for phospholipid-binding activity in a Ca^{2+} dependent manner. Members of the C2 domain superfamily have a variety of cellular functions, e.g. signal transduction, vesicular trafficking, second messenger production and transcriptional regulation. One of the C2 domain-containing proteins which was firstly characterised in *E. histolytica* is a 22 kDa EhC2A (EHI_069320), found in amoebic phagosomes [169].

Moreno *et al*., 2010 demonstrated that EhC2A interacts and translocates upstream regulatory element 3-binding protein transcription factor (URE3-BP) to plasma membrane in response to intracellular Ca^{2+} flux, possibly due to host cell phagocytosis [168]. Interestingly, URE3-BP controls the transcriptional levels of certain virulence factors in *E. histolytica* including the heavy subunit of Gal/GalNAc lectin (*Hgl5*) and ferredoxin genes [158]. Therefore, the recruitment of transcription factor URE3-BP to the plasma membrane results in modulation of URE3-BP regulated transcripts [168]. Additionally, EhC2B (EHI_059860) was found to contain a similar molecular weight with a highly conserved C2 domain, 75% amino acid identical to EhC2A. However, this EhC2B molecule was not coimmunoprecipitated with URE3-BP. It implies that there might be functional divergence between these two structurally similar proteins. Since several C2 domain-containing proteins have been reported for their function in targeting other proteins including transcription factors to cell membranes, EhC2B may have a potential role in acting as a molecular scaffold to anchor associated proteins with the membrane [168]. As mentioned above, EhC2A-mediated transcriptional regulation in response to the increased intracellular Ca^{2+} flux might affect parasite virulence.

In this study, it was found that both EhC2A and EhC2B were greater than 4-fold highly expressed in HM-1:IMSS with log_2FC of 4.82 and 6.94, respectively whereas EhC2A and the other C2 domain protein EhC2D (EHI_118130) were upregulated with log_2FC of 4.27 and 2.10 in PVBM08B, relative to Rahman. Conversely, these C2 domain proteins showed less expression in IULA:1092:1 with log_2FC of 0.47 and 1.50 for EhC2A and EhC2B, respectively.

For other calcium-binding proteins, it was found that the grainin-1 paralogue (EHI_120360) showed marked upregulation in Rahman with log_2FC of 2.01 in relation to PVBM08B as listed in Appendix Table 2.6. Consistent with proteomic analysis of Davis *et al.*, 2006, grainin-1 and grainin-2 show upregulation in Rahman, compared to HM-1:IMSS, and marked reduction in grainin expression was also found in HM-1:IMSS trophozoites after infecting human intestinal xenografts, suggesting the association between upregulation of grainin and decreased virulence [16]. Also, as shown in Table 2.10, there are common transcriptomic modulations of six EF-hand calcium-binding proteins in virulent strains: EHI_079290, EHI_096640 and EHI_148810 for upregulation; EHI_016120, EHI_151890 and EHI_197510 for downregulation, implying that Ca^{2+} -dependent regulatory mechanisms in virulent trophozoites are selectively controlled by a specific set of such calcium-binding proteins.

Due to the involvement of Ca^{2+} in a vast variety of cellular processes, this RNA-Seq analysis suggests that differences in isoforms and transcript levels of these calcium-binding proteins among *E. histolytica* strains are likely to reflect their varied cellular regulations, resulting in differential virulence.

VIII. Transcription factors

Transcriptomic differences between nonvirulent and virulent *E. histolytica* strains have been previously reported both in axenic culture and during host invasion, indicating that differences in virulence are likely to be a consequence of transcriptional variability among parasite strains [74,76,77,82]. Transcriptional regulation of particular genes in eukaryotic organisms including *E. histolytica* is mediated by specific transcription factors (TFs) [162,170-174]. As identified in the complete genomic data, there are fourteen superfamilies of specific TFs in *E. histolytica,* i.e. MYB, bZIP, Cys_2His_2 Zinc Finger, CBF/NFYA, HMG1, AT-hook, Cxc, MADS, GATA, HSF, Homeodomain, WRKY, CENPB and STAT [174].

As a result of their reduction in genomic size during the evolution of parasitism, *Entamoeba* and Apicomplexan parasites have reduced their proteome sizes and most of TFs, compared to their free-living protist lineages [174]. However, particular superfamilies of

specific TFs, such as MYB in *Entamoeba* and AP2 in Apicomplexa, have been evolutionarily expanded in a lineage-specific manner in such protozoan parasites to be the majority of their transcriptional regulators. In *E. histolytica* and *T. vaginalis*, the expanded MYB superfamily was predicted to be a major specific TF cluster in their transcriptome [24,174,175]. Global diversity of specific TFs among protist parasites due to their gene loss and lineage-specific expansions potentially implies the crucial roles of specific TFs in regulating the parasite transcriptome related to their particular lifestyle.

MYB DNA-binding domains (MYB DBD) are approximately 52 amino acids long and highly conserved amongst eukaryotic superkingdom including fungi, plants and vertebrates [176]. In *E. histolytica*, 32 different open reading frames encoding proteins with a putative MYB DBD were identified with varying sizes ranging from 15 to 83 kDa [177]. Also, such MYB DBD-containing proteins can be assigned into three different protein families (Family I, II and III) based on the number of repeats found in their MYB DBD structure [177]. A total of 15 MYB DBD-containing proteins which comprise two repeats (R2 and R3) in their domain are classified into EhMybR2R3 Family I. Family II comprises five members of single repeat MYB DBDs with telomeric binding function whilst nine members of single repeat proteins with a SHAQKYF motif are classified into EhMybSHAQKYF Family III. The other three proteins with a single repeat (EHI_000550, EHI_128200 and EHI_142140) are identified as MYB-related proteins.

In *T. vaginalis*, MYB DBD-containing proteins play an important role in transcriptional regulation of the adhesion protein *ap65-1* gene responsible for the host epithelial cell adhesion [175,178]. In *E. histolytica,* previous microarray analyses revealed the upregulation of certain *EhMybR2R3* genes in HM-1:IMSS strain, e.g. *EhMyb10* (EHI_129790) during mice colon infection; *EhMyb3* (EHI_012420 and paralogous EHI_063550) in response to heat shock stress [77,179].

In this present study, six members of *EhMybR2R3* gene family (EHI_009930, EHI_012420, EHI_063550, EHI_098070, EHI_166410 and EHI_168310) and single *EhMybSHAQKYF* gene (EHI_135150) were found to be significantly downregulated in all three virulent strains relative to nonvirulent Rahman, irrespective of their $\log_2$FC, as listed in Tables 2.8 and 2.10. Conversely, these three virulent strains show the upregulation of single *EhMybSHAQKYF* gene (EHI_136420), EhCDC5-like Myb related gene (EHI_000550) and seven gene members of Cys_2His_2 Zinc Finger protein family (EHI_017720, EHI_055640, EHI_091050, EHI_096780, EHI_105080, EHI_122760 and EHI_176800). This finding suggests that the virulent trophozoites potentially regulate their gene expression with a unique set of specific TFs different from the nonvirulent trophozoite.

The *in silico* analysis previously reported by Meneses *et al*., 2010 demonstrated a list of 246 putative *E. histolytica* genes which were potentially regulated by EhMybR2R3 transcription factor proteins due to the presence of consensus Myb recognition element (MRE) sequence in their gene promoters [177]. Interestingly, the majority of these putative genes with the MRE sequence play a role in signaling and vesicular transport (n = 53), DNA and RNA regulation (n = 37). As such, sixteen putative kinase genes were found to contain the MRE sequence in their promoter regions [177]. This implies that downregulation of *EhMybR2R3* genes in the trophozoite transcriptome would largely influence such biological processes and potentially result in an aberrant behavior of virulent strains.

In this work, RNA-Seq data and InterProScan protein domain analysis revealed the pronounced downregulation of signaling genes, especially for protein kinases in all three virulent strains as listed in Table 2.10 and Figure 2.23. Thus, it is possible that the reduction of gene transcripts involved in signaling pathways would result from downregulation of R2R3 MYB DBD-containing protein gene family. Moreover, the intriguing hypothesis is that a distinctive upregulation of six Zn finger gene family members potentially regulates the expression of virulence-associated genes in these three virulent strains. Essentially, these findings strongly suggest that diversity of specific TF superfamilies in the parasite genome enables the parasites to regulate a distinctive set of genes responsible for their particular behavior.

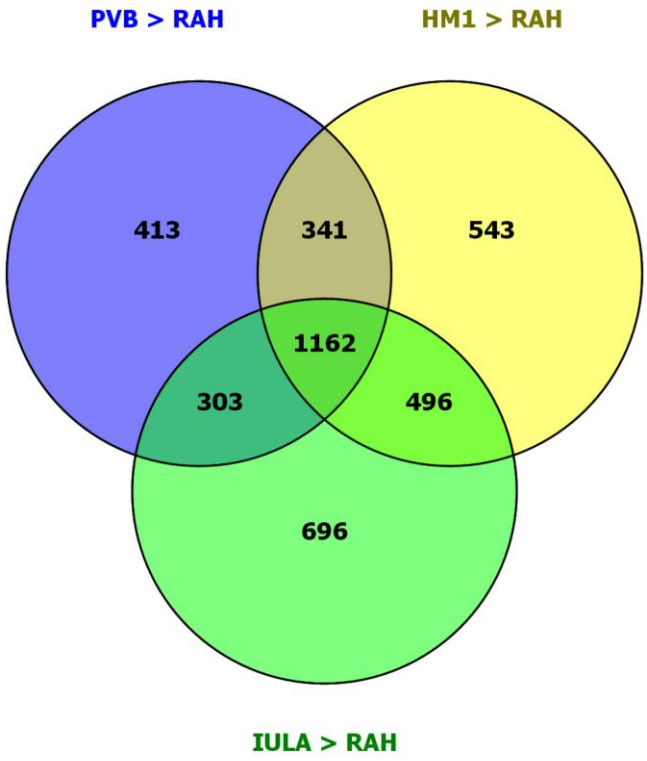


Figure 2.13: The number of genes known to be significantly upregulated (FDR-adjusted *P*-value < 0.05) in the three virulent strains (i.e. PVBM08B, HM-1:IMSS and IULA:1092:1) relative to nonvirulent Rahman. The intersection of gene members in each coloured circle is based on AmoebaDB_IDs. A total of 1,162 upregulated genes regardless of their $\log_2$FC are commonly found in all these three virulent strains.

Table 2.6: The 38 most frequent functionally annotated transcripts significantly upregulated (FDR-adjusted *P*-value < 0.05, regardless of log_2FC) in all three virulent *E. histolytica* strains.

group	Functional gene annotation	Number of genes
1.	leucine-rich repeat protein, BspA family	16
2.	Rab family GTPase	12
3.	protein kinase domain-containing protein	10
4.	heat shock protein 70, putative	7
5.	WD domain-containing protein	7
6.	zinc finger protein, putative	7
7.	actin, putative	6
8.	proteasome regulatory subunit, putative	6
9.	26s proteinase regulatory subunit, putative	5
10.	HEAT repeat domain-containing protein	5
11.	Rho guanine nucleotide exchange factor, putative	5
12.	RNA recognition motif domain-containing protein	5
13.	tyrosine kinase, putative	5
14.	long-chain-fatty-acid--CoA ligase, putative	4
15.	phospholipid-transprting P-type ATPase, putative	4
16.	protein kinase, putative	4
17.	Rho GTPase-activating protein, putative	4
18.	5'-3' exonuclease domain-containing protein	3
19.	actin-binding protein, cofilin/tropomyosin family	3
20.	ankyrin repeat protein, putative	3
21.	ATP-binding cassette protein, putative	3
22.	C2 domain-containing protein	3
23.	CXXC-rich protein	3
24.	cysteine proteinase, putative	3
25.	DEAD/DEAH box helicase, putative	3
26.	EF-hand calcium-binding domain-containing protein	3
27.	peroxiredoxin	3
28.	protein phosphatase, putative	3
29.	RhoGAP domain-containing protein	3
30.	ribosomal protein L17, putative	3
31.	ribosomal protein S24, putative	3
32.	RNA-binding protein, putative	3
33.	serine-threonine-isoleucine rich protein, putative	3
34.	transporter, major facilitator family	3
35.	UBA/TS-N domain-containing protein	3
36.	ubiquitin-conjugating enzyme family protein	3
37.	ubiquitin carboxyl-terminal hydrolase domain-containing protein	3
38.	zinc finger domain-containing protein	3

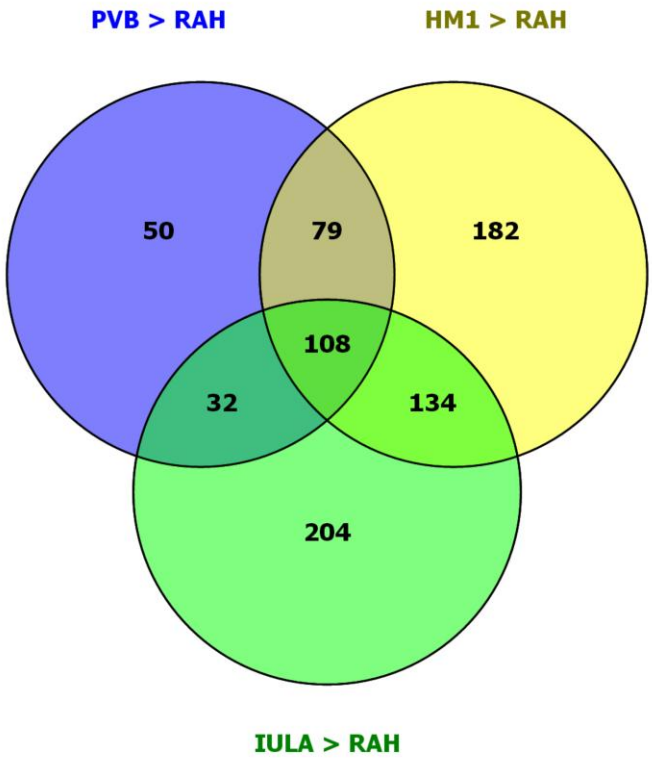


Figure 2.14: The number of significantly upregulated genes in the three virulent strains where $log_2FC \geq 2$. The intersection of gene members in each coloured circle is based on AmoebaDB_IDs. Only 108 of 1,162 genes upregulated in all three virulent strains as shown in Figure 2.13 have higher expression levels with $log_2FC \geq 2$ than Rahman.

Table 2.7: Summary of 108 upregulated DE transcripts with $log_2FC \geq 2$, commonly found in three virulent strains, assigned to 11 functional categories with their functional gene annotations and AmoebaDB_IDs.

Gene Category	Functional gene annotation	Number of genes	AmoebaDB_ID
Surface-associated	surface antigen ariel1, putative	1	EHI_123850
Host cell killing and mucosal invasion	cysteine proteinase, putative	1	EHI_097900
	serine-threonine-isoleucine rich protein	2	EHI_012330, EHI_025700
	leucine-rich repeat protein, BspA family	13	EHI_015120, EHI_018840, EHI_034610, EHI_041470, EHI_049160, EHI_070330, EHI_095060, EHI_112290, EHI_123820, EHI_176480, EHI_184260, EHI_191510, EHI_199270
Oxidative stress response	peroxiredoxin	1	EHI_145840
Bacterial killing	AIG1 family protein	2	EHI_176280, EHI_180390
Nucleic acid interaction	zinc finger protein, putative	2	EHI_091050, EHI_105080
	replication protein, pseudogene, putative	1	EHI_190200
	kinetochore protein Spc25 domain-containing protein	1	EHI_181520
	Myb family DNA-binding protein, SHAQKYF family	1	EHI_136420
	regulator of nonsense transcripts, putative	2	EHI_043440, EHI_193520
Ribosomal structure	60S ribosome subunit biogenesis protein NIP7, putative	1	EHI_031350
Protein foling	heat shock protein, putative	1	EHI_034710
	chaperone clpB, putative	1	EHI_155060
Signaling	protein kinase domain-containing protein	2	EHI_059040, EHI_144590
	tyrosine kinase, putative	3	EHI_117680, EHI_123840, EHI_148550
	Rap/Ran GTPase-activating protein, putative	1	EHI_108750

Table 2.7: Summary of 108 upregulated DE transcripts with $log_2FC \geq 2$, commonly found in three virulent strains, assigned to 11 functional categories. (Continued)

Gene Category	Functional gene annotation	Number of genes	AmoebaDB_ID
Signaling	dedicator of cytokinesis domain-containing protein	1	EHI_185270
Protein degradation	26S proteinase regulatory subunit, putative	1	EHI_053020
Miscellaneous	serine acetyltransferase 1	1	EHI_021570
	Fe-S cluster assembly protein NifU, putative	1	EHI_049620
	CXXC-rich protein	2	EHI_050970, EHI_082260
	PP-loop family protein	1	EHI_108760
	glutamic acid-rich protein, putative	1	EHI_053200
	tRNA-Leu (anticodon: CAA)	1	EHI_095430
	molybdenum cofactor synthesis protein3, putative	1	EHI_118040
	iron-sulfur flavoprotein, putative	1	EHI_138480
	Skp1 family protein	1	EHI_174180
	cdc48-like protein, putative	1	EHI_176970
	dextranase precursor, putative	1	EHI_182460
	dentin sialophosphoprotein precursor, putative	1	EHI_188600
	predicted protein	1	EHI_201420
Hypothetical	N/A	55	EHI_004070, EHI_004410, EHI_005657, EHI_010160, EHI_015220, EHI_015980, EHI_029500, EHI_034840, EHI_037440, EHI_047620, EHI_049820, EHI_051440, EHI_054670, EHI_054780, EHI_056110, EHI_057950, EHI_059330, EHI_067600, EHI_070130, EHI_071210, EHI_075430, EHI_080860, EHI_080880, EHI_083380, EHI_087110, EHI_087740, EHI_091740, EHI_113200, EHI_113950, EHI_118230, EHI_119750, EHI_121060, EHI_123120, EHI_128800, EHI_133780, EHI_134710, EHI_136480, EHI_145610, EHI_146130, EHI_151340, EHI_152360, EHI_153050, EHI_154160, EHI_160970, EHI_163360, EHI_169670, EHI_172000, EHI_174580, EHI_180410, EHI_180940, EHI_184500, EHI_187800, EHI_188860, EHI_198220, EHI_200950
	Total	**108**	

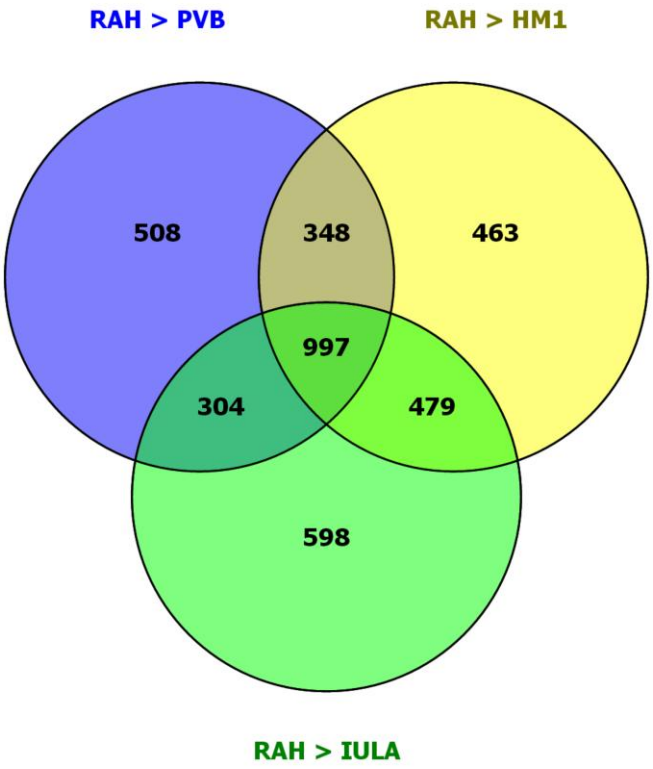


Figure 2.15: The number of genes known to be significantly downregulated (FDR-adjusted *P*-value < 0.05) in the three virulent strains relative to nonvirulent Rahman. The intersection of gene members in each coloured circle is based on AmoebaDB_IDs. A total of 997 genes regardless of their $\log_2$FC are commonly downregulated in all these three virulent strains.

Table 2.8: The 30 most frequent functionally annotated transcripts significantly downregulated (FDR-adjusted *P*-value < 0.05, regardless of log_2FC) in all three virulent *E. histolytica* strains.

group	Function gene annotation	Number of genes
1.	protein kinase domain-containing protein	18
2.	Rab family GTPase	15
3.	protein kinase, putative	14
4.	RhoGAP domain-containing protein	10
5.	Ras guanine nucleotide exchange factor, putative	9
6.	tyrosine kinase, putative	9
7.	WD domain-containing protein	8
8.	leucine-rich repeat protein, BspA family	7
9.	Rab GTPase-activating protein, putative	7
10.	Myb-like DNA-binding domain-containing protein	6
11.	protein phosphatase domain-containing protein	5
12.	Rho guanine nucleotide exchange factor, putative	5
13.	RNA recognition motif domain-containing protein	5
14.	zinc finger domain-containing protein	5
15.	acetyltransferase, GNAT family	4
16.	AIG1 family protein, putative	4
17.	DnaJ family protein	4
18.	protein tyrosine kinase domain-containing protein	4
19.	Ras family GTPase	4
20.	CDP-alcohol phosphatidyltransferase family protein	3
21.	CXXC-rich protein	3
22.	EF-hand calcium-binding domain-containing protein	3
23.	importin alpha, putative	3
24.	leucine-rich repeat-containing protein	3
25.	LIM zinc finger domain-containing protein	3
26.	ser/thr protein phosphatase family protein	3
27.	TBC domain-containing protein	3
28.	thioredoxin, putative	3
29.	ubiquitin carboxyl-terminal hydrolase domain-containing protein	3
30.	WD repeat protein	3

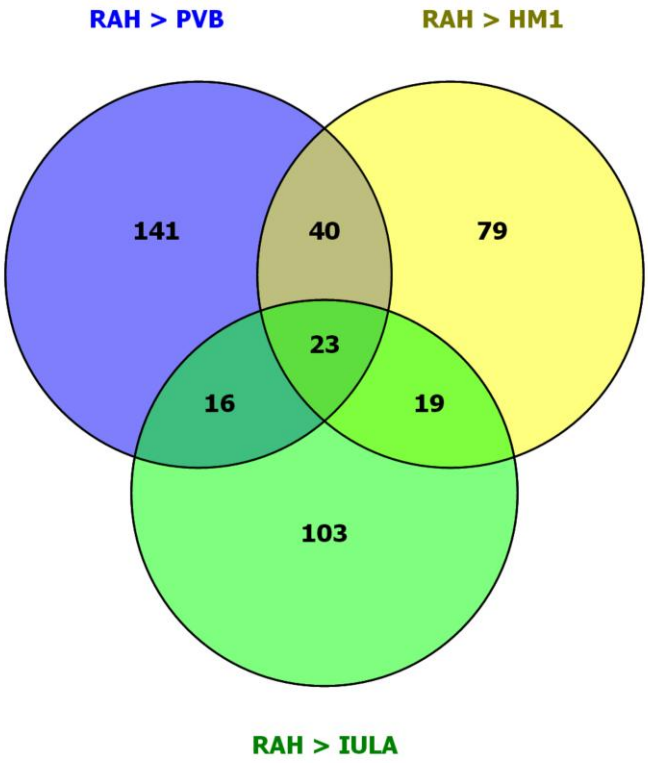


Figure 2.16: The number of significantly downregulated genes in the three virulent strains where $log_2FC \leq -2$. The intersection of gene members in each coloured circle is based on AmoebaDB_IDs. Only 23 of 997 genes downregulated in all three virulent strains as shown in Figure 2.15 show lower expression levels ($log_2FC \leq -2$) when compared to Rahman.

Table 2.9: Summary of 23 downregulated DE transcripts with $log_2FC \leq -2$ commonly found in three virulent strains, assigned to 8 functional categories with their functional gene annotations and AmoebaDB_IDs.

Gene Category	Function gene annotation	Number of genes	AmoebaDB_ID
Surface-associated	surface antigen ariel1, putative	1	EHI_172850
Bacterial killing	AIG1 family protein	2	EHI_176590, EHI_176700
Nucleic acid interaction	Myb-like DNA-binding domain-containing protein	1	EHI_063550
Ribosomal structure	60S ribosomal protein L38, putative	1	EHI_023840
Signaling	protein kinase domain-containing protein	1	EHI_023860
	WD domain-containing protein	1	EHI_023870
Protein degradation	ubiquitin-conjugating enzyme family protein	1	EHI_023880
Miscellaneous	metallo-beta-lactamase superfamily protein	1	EHI_115720
	nuclear movement protein, putative	1	EHI_023890
	rhodanase-like domain-containing protein	1	EHI_067950
Hypothetical	N/A	12	EHI_006180, EHI_019860, EHI_023850, EHI_047110, EHI_064440, EHI_069940, EHI_072740, EHI_096610, EHI_135600, EHI_192530, EHI_023900, EHI_095100
	Total	**23**	

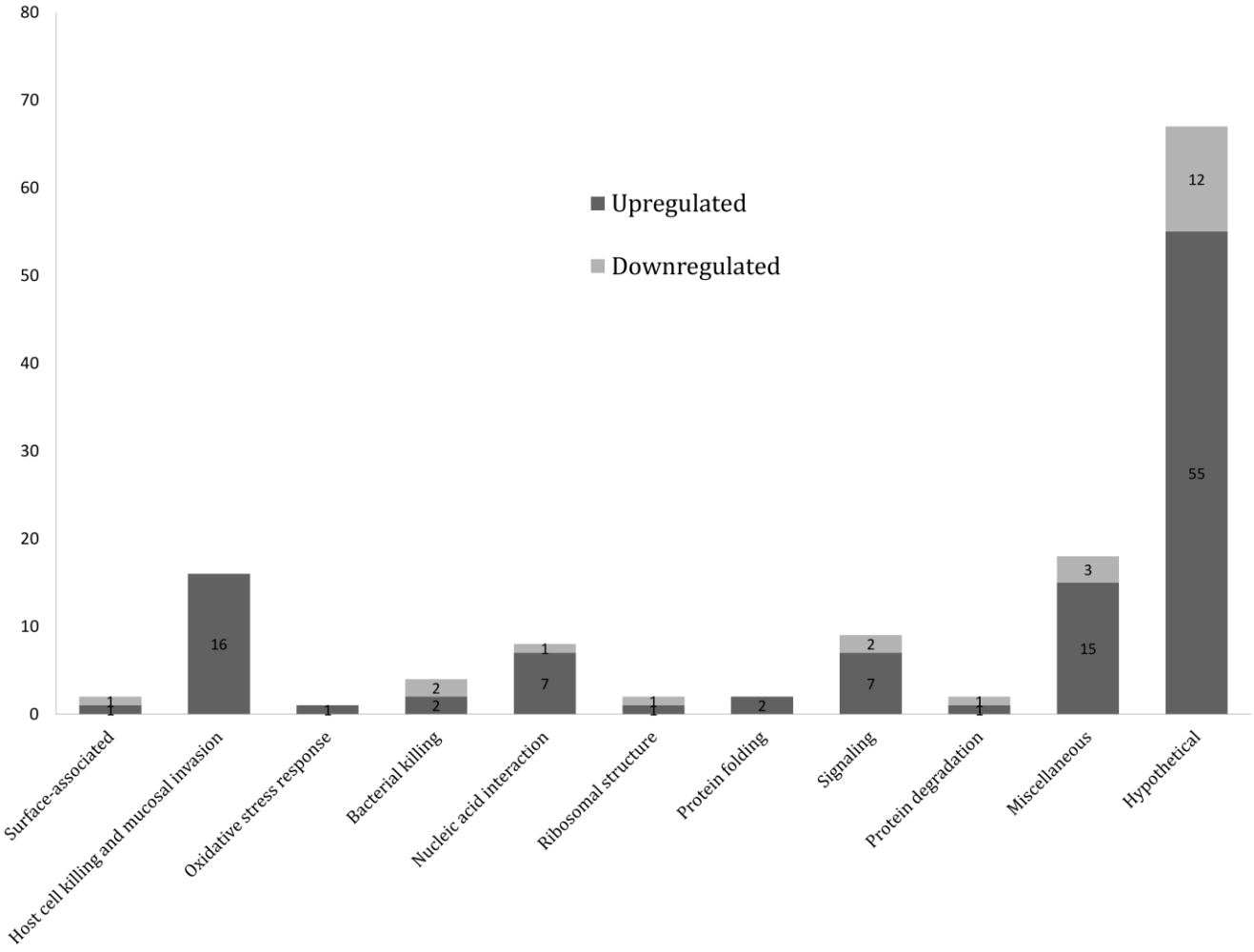


Figure 2.17: The number of modulated transcripts in all three virulent strains with $log_2FC \geq 2$ for upregulation and $log_2FC \leq -2$ for downregulation, based on their functional categories in Tables 2.7 and 2.9. Strikingly, categories of host cell killing and mucosal invasion, nucleic acid interaction as well as signaling are markedly upregulated in such virulent *E. histolytica* strains.

Table 2.10: Functional genes with transcriptomic modulations in all three virulent strains (n = 417), regardless of their log_2FC. These modulated transcripts can be assigned into 75 functional annotations as listed below.

group	Functional gene annotation	Number of genes		
		Total	upregulated	downregulated
1.	protein kinase domain-containing protein	28	10	18
2.	Rab family GTPase	27	12	15
3.	leucine-rich repeat protein, BspA family	23	16	7
4.	protein kinase, putative	18	4	14
5.	WD domain-containing protein	15	7	8
6.	tyrosine kinase, putative	14	5	9
7.	RhoGAP domain-containing protein	13	3	10
8.	Rho guanine nucleotide exchange factor, putative	10	5	5
9.	RNA recognition motif domain-containing protein	10	5	5
10.	Ras guanine nucleotide exchange factor, putative	10	1	9
11.	heat shock protein 70, putative	9	7	2
12.	Rab GTPase-activating protein, putative	9	2	7
13.	zinc finger protein, putative	8	7	1
14.	zinc finger domain-containing protein	8	3	5
15.	AIG1 family protein, putative	7	3	4
16.	myb-like DNA-binding domain-containing protein	7	1	6
17.	HEAT repeat domain-containing protein	6	5	1
18.	long-chain-fatty-acid--CoA ligase, putative	6	4	2
19.	CXXC-rich protein	6	3	3
20.	EF-hand calcium-binding domain-containing protein	6	3	3
21.	ubiquitin carboxyl-terminal hydrolase domain-containing protein	6	3	3
22.	protein phosphatase domain-containing protein	6	1	5
23.	phospholipid-transporting P-type ATPase, putative	5	4	1
24.	ankyrin repeat protein, putative	5	3	2
25.	DEAD/DEAH box helicase, putative	5	3	2
26.	transporter, major facilitator family	5	3	2
27.	ubiquitin-conjugating enzyme family protein	5	3	2
28.	leucine-rich repeat-containing protein	5	2	3
29.	LIM zinc finger domain-containing protein	5	2	3
30.	Rap/Ran GTPase-activating protein, putative	5	2	3

Table 2.10: Functional genes with transcriptomic modulations in three virulent strains (n = 417), regardless of their log_2FC. (Continued)

group	Functional gene annotation	Number of genes		
		Total	upregulated	downregulated
31.	protein tyrosine kinase domain-containing protein	5	1	4
32.	C2 domain-containing protein	4	3	1
33.	RNA-binding protein, putative	4	3	1
34.	serine-threonine-isoleucine rich protein, putative	4	3	1
35.	leucine-rich repeat / protein phosphatase 2C domain-containing protein	4	2	2
36.	surface antigen ariel1, putative	4	2	2
37.	TBC domain-containing protein	4	1	3
38.	thioredoxin, putative	4	1	3
39.	WD repeat protein	4	1	3
40.	acetyltransferase, putative	3	2	1
41.	cysteine proteinase, putative	4	3	1
42.	helicase, putative	3	2	1
43.	high mobility group (HMG) box domain-containing protein	3	2	1
44.	phospholipase, patatin family protein	3	2	1
45.	protein tyrosine phosphatase, putative	3	2	1
46.	AAA family ATPase, putative	3	1	2
47.	casein kinase II regulatory subunit family protein	3	1	2
48.	dual specificity protein phosphatase, putative	3	1	2
49.	fatty acid elongase, putative	3	1	2
50.	HAD hydrolase, family IA, variant 3	3	1	2
51.	IBR domain-containing protein	3	1	2
52.	myotubularin, putative	3	1	2
53.	nucleosome assembly protein, putative	3	1	2
54.	Rho family GTPase	3	1	2
55.	serine/threonine-protein kinase, putative	3	1	2
56.	ARF GTPase-activating protein, putative	2	1	1
57.	citrate transporter, putative	2	1	1
58.	dihydrouridine synthase (Dus) family protein	2	1	1
59.	glucosamine 6-phosphate N-acetyltransferase, putative	2	1	1
60.	GTP-binding protein, putative	2	1	1

Table 2.10: Functional genes with transcriptomic modulations in three virulent strains (n = 417), regardless of their log_2FC. (Continued)

group	Functional gene annotation	Number of genes		
		Total	upregulated	downregulated
61.	haloacid dehalogenase-like hydrolase domain-containing protein	2	1	1
62.	hydrolase, alpha/beta fold family domain-containing protein	2	1	1
63.	inositol polyphosphate 5-phosphatase, putative	2	1	1
64.	leucine-rich repeat and phosphatase domain-containing protein	2	1	1
65.	longevity-assurance family protein	2	1	1
66.	Myb family DNA-binding protein, SHAQKYF family	2	1	1
67.	peptidyl-prolyl cis-trans isomerase, putative	2	1	1
68.	PH domain-containing protein kinase, putative	2	1	1
69.	PP-loop family protein	2	1	1
70.	pumilio family RNA-binding protein	2	1	1
71.	receptor protein kinase, putative	2	1	1
72.	RNA polymerase III subunit, putative	2	1	1
73.	Sec1 family protein	2	1	1
74.	Sec7 domain protein	2	1	1
75.	transporter, auxin efflux carrier (AEC) family	2	1	1

2.3.7 Cluster analysis of all differentially expressed genes unravels the spectrum of co-upregulation pattern of transcript populations in the virulent strains, suggesting their potential role in strain-specific virulence

Besides Venn diagrams, I also tried to explore the pattern of transcriptional differences across the strains by hierarchical clustering of 7,024 DE genes based on their relative expression pattern (log_2FC) across 6 pairs of contrast. In Figure 2.18, all 7,024 significantly DE genes could be categorised into 9 clusters. The first three columns represent three contrast pairs of Rahman vs PVBM08B, Rahman vs HM-1:IMSS and Rahman vs IULA:1092:1, respectively. The 4th and 5th columns represent two pairs of PVBM08B vs HM-1:IMSS and PVBM08B vs IULA:1092:1, respectively. Lastly, the 6th column represents a pair of HM-1:IMSS vs IULA:1092:1. Strikingly, 98 DE genes were grouped together in 6th cluster, showing remarkable differences among columns of the heatmap, compared to the other clusters.

In this 6th cluster, the majority of genes in the first two columns are depicted with light blue to deep blue colour (average log_2FC = -4.63 and -4.42 for 1st and 2nd columns, respectively), meaning that this group of DE genes has downregulated expression in Rahman, compared to PVBM08B and HM-1:IMSS. Conversely, the last two columns of 6th cluster are highlighted with orange (average log_2FC = 4.13 and 3.92), indicating higher expression in PVBM08B and HM-1:IMSS than IULA:1092:1. For the two middle columns, their average log_2FC values are -0.50 and 0.22, referring to similar expression between Rahman and IULA:1092:1 and between PVBM08B and HM-1:IMSS, respectively.

It is likely to imply that this unique cluster represents a set of genes showing differential expression across the strains with high transcript levels in the two most virulent strains (i.e. PVBM08B and HM-1:IMSS) and low transcript levels in less virulent IULA:1092:1 and nonvirulent Rahman. To further scrutinise the biological relevance of this 6th cluster, 2nd cluster analysis was done and shown in Figure 2.19. All 98 DE genes retrieved from the 6th cluster in Figure 2.18 could be cateogorised into five subclusters, as detailed in Table 2.11.

Based on their relative expression levels, different colour spectra in the first two columns of the 1st subcluster clearly indicate that its gene members show higher transcript levels in HM-1:IMSS (deep blue, average log_2FC = 6.08) than PVBM08B (light blue, average log_2FC = 4.58), relative to Rahman. Differently, the 2nd subcluster shows greater in average transcript levels and number of genes than the 1st subcluster and also displays similar expression levels between these two virulent strains relative to Rahman with average log_2FC = 6.88 for PVBM08B and 6.66 for HM-1:IMSS. In contrast to the 1st subcluster, the 4th

subcluster shows a set of genes with higher upregulation in PVBM08B than HM-1:IMSS with average log_2FC = 5.49 and 1.96, respectively. Also, the same transcriptional differences between PVBM08B and HM-1:IMSS could be observed in the last two columns in comparison to IULA:1092:1.

Essentially, it is interesting that there are different spectra of expression levels of such 98 DE genes among these four *E. histolytica* strains and cluster analysis can categorise such genes with similar co-upregulation pattern in the virulent parasites into 5 subclusters. From the 1st subcluster, the majority of co-upregulated functional members are BspA-like LRRPs (EHI_018840, EHI_034610, EHI_102380 and EHI_105370). Regarding their FDR-corrected *P*-values in the previous DGE results (data not shown), three LRRP members (EHI_018840, EHI_102380 and EHI_105370) show statistically significant upregulation in HM-1:IMSS compared to PVBM08B, indicating that HM-1:IMSS has a greater potential to invade the host tissue than PVBM08B. On the contrary, the 4th subcluster reveals the greater expression of peroxiredoxin (EHI_145840), DNA polymerase (EHI_018010), multidrug resistance-associated protein (EHI_084730) in PVBM08B than HM-1:IMSS. Similarly, significant testings reveal corrected *P*-values less than 0.05 for peroxiredoxin (EHI_145840) and multidrug resistance-associated protein (EHI_084730), reflecting the possible higher capability to survive and multiply under conditions of host stress, i.e. ROS attack and antibiotic inhibition.

Strikingly, as shown in the 2nd subcluster, all three *EhSTIRP* gene members (EHI_004340, EHI_012330 and EHI_025700) which were found to be upregulated in the axenic condition show a similar pattern in upregulation to other virulence-associated genes including C2 domain-containing protein (EHI_059860), BspA-like LRRP (EHI_127710), WD domain-containing protein (EHI_092070) and 70 kDa heat shock proteins (EHI_021780 and EHI_133950). Intriguingly, a similar upregulation pattern are found in such virulence-associated genes related to pathogenic processes e.g. host cell killing, mucosal invasion and stress response. In addition, members of G-protein signaling system, i.e. Ras family GTPase (EHI_058520), RhoGAP domain-containing protein (EHI_199570) and Rab GDP dissociation inhibitor alpha (EHI_164890), are shown to be upregulated with such above virulence-associated genes. Therefore, it could be speculated that their expression is potentially stimulated through the activation of such G-protein signaling.

For IULA:1092:1, all these 98 DE genes in the 5th column (logFC_PVBvsIULA) show positive log_2FC values, indicating higher expression in PVBM08B than IULA:1092:1 and 83 of such 98 genes show significantly differential expression in the previous DGE test (data not shown). Moreover, 96 of 98 DE genes in the last 6th column (logFC_HM1vsIULA) except

multidrug resistance-associated protein (EHI_084730) and one hypothetical gene (EHI_033450) display positive log_2FC values, indicating higher expression in HM-1:IMSS than IULA:1092:1 and 82 of these 98 genes show the statistical significance in DGE results. One plausible interpretation for this finding is that IULA:1092:1 is less virulent than PVBM08B and HM-1:IMSS. This is consistent with the PCA plot in Figure 2.8, showing that IULA:1092:1 transcriptome libraries were plotted separately from more virulent PVBM08B and HM-1:IMSS transcriptome libraries. Taken together, these explorative analyses can comprehensively cluster a core set of DE genes which exhibit distinctive expression profiles across six pairs of contrasting strains, suggesting their potential role in strain-specific virulence and provide compelling biological interpretations as explained above.

In other words, a core set of 98 DE genes identified from the cluster analysis unveils a high degree of transcriptional variation among virulent strains, suggesting such 98 DE genes are likely to be major virulence-determining factors. Therefore, it might be substantially advantageous for development of a novel therapeutic drug to effectively treat patients with invasive amoebiasis.

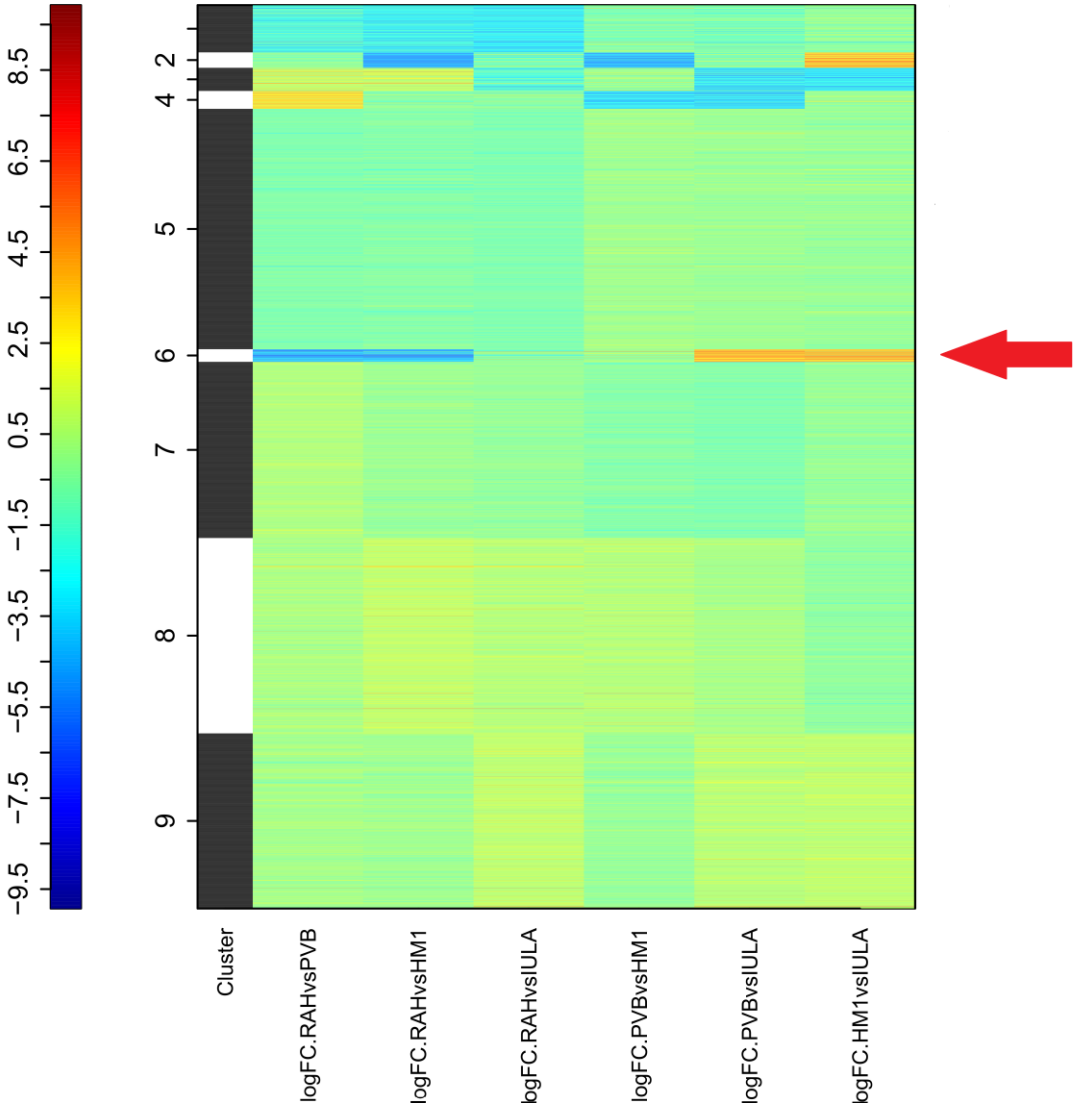


Figure 2.18: **Agglomerative hierarchical clustering of all DE genes based on their relative expression levels across all six contrast pairs.** Colour spectrum bar on the left represents the relative expression levels (log_2FC). The leftmost column indicates clusters defined by hierarchical clustering. All of 7,024 DE genes retrieved from six contrast pairs can be grouped into nine clusters, based on their pattern of expression levels across all strains. Interestingly, 6th cluster demonstrates a group of DE genes with distinctive pattern of log_2FC across six contrast pairs. Members of DE genes in 6th cluster have high levels of expression in two virulent strains, i.e. HM-1:IMSS and PVBM08B, but low expression levels in virulent IULA:1092:1 strain and nonvirulent Rahman strain. This 6th cluster are further categorised into five subclusters shown in Figure 2.19.

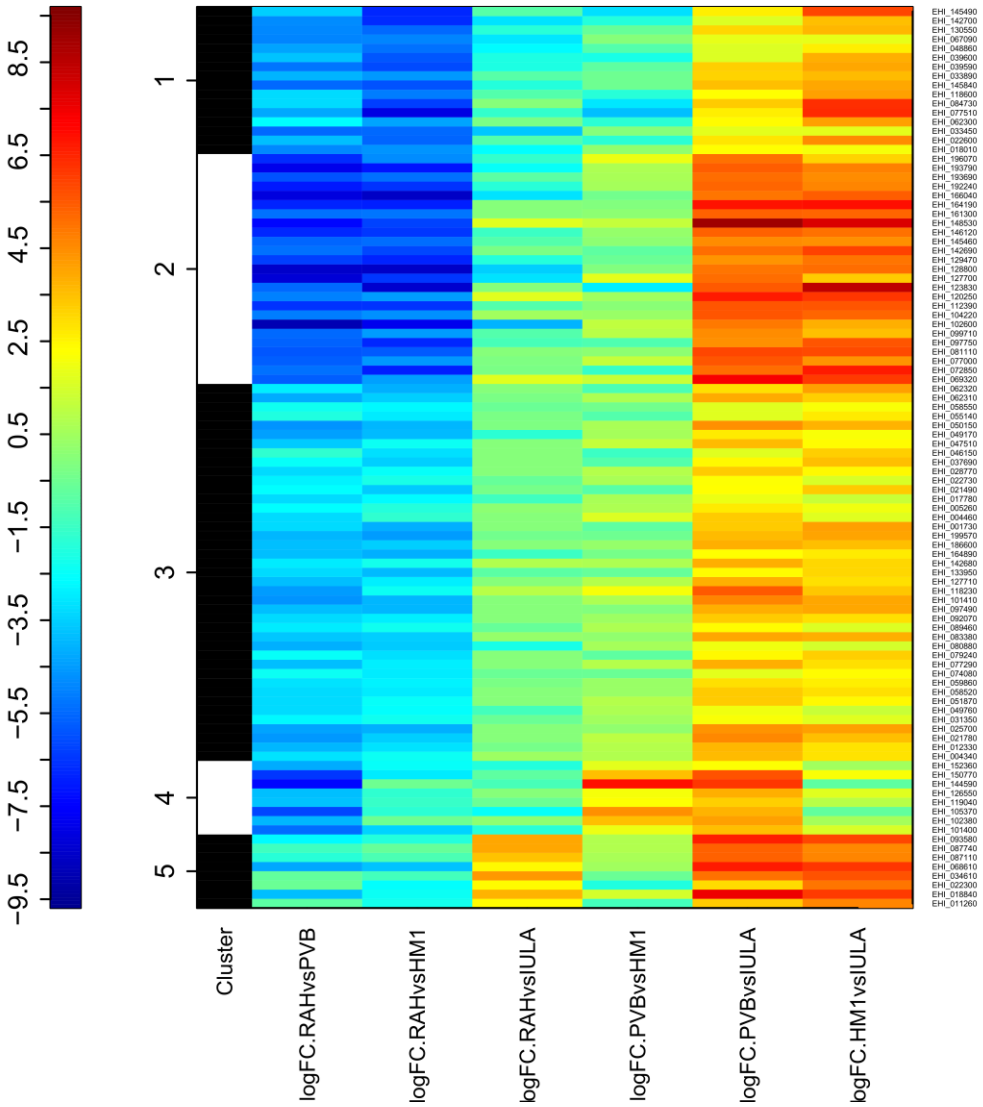


Figure 2.19: **Agglomerative hierarchical clustering of 98 DE genes retrieved from the 6th cluster in previous analysis (Figure 2.18).** All of these 98 genes can be further categorised into five subclusters, based on their pattern of expression levels. Obviously, different co-expression patterns of DE genes among the parasite strains are demonstrated. The details of each subcluster are summarised in Table 2.11. Interestingly, DE genes in 1st, 2nd and 3rd clusters show similar expression levels in the middle two (3rd: Rahman vs IULA:1092:1 and 4th: PVBM08B vs HM-1:IMSS) columns.

Table 2.11: Summary of 2nd cluster analysis results of 98 DE genes retrieved from the 6th cluster of the heatmap in Figure 2.18, including functional gene annotation, number of genes and AmoebaDB_IDs.

Sub cluster	Functional gene annotation	Number of genes	AmobaDB_ID
1	leucine-rich repeat protein, BspA family	4	EHI_018840, EHI_034610, EHI_102380, EHI_105370
	AIG1 family protein, putative	2	EHI_119040, EHI_126550
	heat shock protein 70, putative	1	EHI_150770
	60S ribosomal protein L6, putative	1	EHI_093580
	protein kinase domain-containing protein	1	EHI_144590
	hypothetical protein	7	EHI_011260, EHI_022300, EHI_068610, EHI_087110, EHI_087740, EHI_101400, EHI_152360
2	serine-threonine-isoleucine rich protein, putative	3	EHI_004340, EHI_012330, EHI_025700
	heat shock protein 70, putative	2	EHI_021780, EHI_133950
	C2 domain-containing protein	1	EHI_059860
	leucine-rich repeat protein, BspA family	1	EHI_127710
	WD domain-containing protein	1	EHI_092070
	Ras family GTPase	1	EHI_058520
	RhoGAP domain-containing protein	1	EHI_199570
	P-glycoprotein-2, putative	1	EHI_186600
	60S ribosome subunit biogenesis protein NIP7, putative	1	EHI_031350
	peptidyl-prolyl cis-trans isomerase, FKBP-type , putative	1	EHI_051870
	Rab GDP dissociation inhibitor alpha, putative	1	EHI_164890
	hypothetical protein	11	EHI_049760, EHI_074080, EHI_077290, EHI_079240, EHI_080880, EHI_083380, EHI_089460, EHI_097490, EHI_101410, EHI_118230, EHI_142680
3	AIG1 family protein	3	EHI_072850, EHI_102600, EHI_129470
	leucine-rich repeat protein, BspA family	2	EHI_148530, EHI_161300
	C2 domain-containing protein	1	EHI_069320
	surface antigen ariel1, putative	1	EHI_005260
	heat shock protein 70, mitochondrial, putative	1	EHI_127700
	coiled-coil domain-containing protein 25, putative	1	EHI_021490

Table 2.11: Summary of 2nd cluster analysis results of 98 DE genes retrieved from the 6th cluster of the heatmap in Figure 2.18. (Continued)

Sub cluster	Functional gene annotation	Number of genes	Amobadb_ID
3	signal recognition particle 54 kDa protein, putative	1	EHI_022730
	splicing factor 3B subunit 1, putative	1	EHI_049170
	HEAT repeat domain-containing protein	1	EHI_050150
	ethanolamine phosphotransferase, putative	1	EHI_055140
	pre-mRNA cleavage factor I 25 kDa subunit, putative	1	EHI_077000
	DNA mismatch repair protein Msh2, putative	1	EHI_123830
	DNA polymerase, putative	1	EHI_164190
	Ras family GTPase, pseudogene	1	EHI_058550
	hypothetical protein	24	EHI_001730, EHI_004460, EHI_017780, EHI_028770, EHI_037690, EHI_046150, EHI_047510, EHI_062310, EHI_062320, EHI_081110, EHI_097750, EHI_099710, EHI_104220, EHI_112390, EHI_120250, EHI_128800, EHI_142690, EHI_145460, EHI_146120, EHI_166040, EHI_192240, EHI_193690, EHI_193790, EHI_196070
4	peroxiredoxin	1	EHI_145840
	DNA polymerase, putative	1	EHI_018010
	NADPH-dependent FMN reductase domain-containing protein	1	EHI_022600
	multidrug resistance-associated protein, putative	1	EHI_084730
	calcineurin catalytic subunit A, putative	1	EHI_118600
	hypothetical protein	3	EHI_033450, EHI_062300, EHI_077510,
5	U3 small nucleolar ribonucleo protein MPP10, putative	1	EHI_048860
	endonuclease V, putative	1	EHI_142700
	hypothetical protein	6	EHI_033890, EHI_039590, EHI_039600, EHI_067090 EHI_130550, EHI_145490
	Total (subclusters 1st, 2nd, 3rd, 4th and 5th)	**98**	

2.3.8 Sequence divergence in genes implicated in host-parasite interaction is significantly correlated with transcriptional variability across *E. histolytica* strains

Based on the Red Queen hypothesis raised by Van Valen, 1973, coevolved species such as host and parasite can drive the molecular evolution of each other [180]. Both host and parasite need to continuously adapt to gain reproductive fitness and survive under selective pressures from the changing environment and interacting species [180,181]. Molecular evolution of genes involved in host resistance and parasite infectivity should be driven faster than others [182,183]. As mentioned above, it is thus hypothesised that an exclusive set of *E. histolytica* genes directly involved in the host-parasite interaction should have evolved at a faster rate than other genes, due to the 'molecular arm races' for host-parasite coevolution [180-183].

It was recently reported by Weedall *et al.*, 2012 that sequence variation among genomes of *E. histolytica* strains was quite low (0.312-0.857 SNPs per kb), different from *Plasmodium falciparum* which shows higher sequence diversity with 1.31 SNPs per kb [70,184]. However, it is intriguing that a unique set of genes displays high sequence polymorphisms across the sequenced strains. Across all 8,333 *E. histolytica* genes, a total of 3,022 genes exhibit intraspecific SNPs and the majority of these genes (1,644 genes, 54.4%) encode for hypothetical proteins [70]. Among these 3,022 genes, 53 genes with ≥ 5 nonsynonymous homozygous SNPs across sequenced strains were identified as highly polymorphic genes. It is worth noting that these nonsynomyous SNPs are more common in genes associated with the host-parasite interaction such as EhSTIRPs, the intermediate chains of Gal/GalNAc lectin Lgl1 and Lgl2, BspA-like LRRPs and AIG1-like family proteins. Also, a large number of SNPs could be detected in regulatory genes, i.e. protein kinase domain-containing proteins, tyrosine kinases implicated in protein phosphorylation and signaling pathways as well as 70 kDa heat shock proteins responsible for stress response [70].

Therefore, these findings of sequence polymorphisms reported by Weedall *et al.*, 2012 are consistent with the Red Queen hypothesis of antagonistic coevolution between virulent *E. histolytica* parasites and their human host. Interestingly, these genes are also directly implicated for parasite survival and amoebic virulence and likely to exhibit differential expression across strains. It would follow that genes that are under positive selection (i.e. selection to change) would also be under selective pressure for changes in transcript levels, as this is another route to phenotypic variability. It would also follow that

the changes in primary DNA sequence would also lead to changes in gene expression as they could influence the binding of the transcriptomic machinery and of transcription factors.

To test this hypothesis, 98 DE genes which are obtained from the previous cluster analysis (the 6th cluster, Figure 2.18) and show the remarkable differences in expression across the strains were studied to assess whether overall genotypic differences between strains are linked to transcriptional variation.

The details of SNPs in these 98 DE genes across all *E. histolytica* strains are available in the AmoebaDB database version 4.2, as summarised in Appendix Table 3. It is striking that average SNP sites per kilobase of these 98 DE genes across all strains are 9.31 SNPs/kb, more than 10-fold higher than average value across the genome, previously mentioned. As shown in Figure 2.20, the numbers of total SNP sites across all strains were plotted against the maximal transcriptional differences across all strains, represented by $\log_2$-transformed values of the ratio between maximal FPKM and minimal FPKM observed for a particular gene.

The scatterplot shows a strong significant correlation (r = 0.3097, P-value = 0.0019) between polymorphisms and transcriptional variablility across these 98 genes, indicating that a particular gene with a faster rate of evolution tends to have a more variable transcription when compared across all strains. In addition, it is likely that transcriptional regulation is less tight in a gene with more variable sequence. In other words, evolutionary change of sequence potentially leads to alteration in transcriptional regulation and subsequent differential abundance of such polymorphic gene across the parasite strains. Essentially, host selective pressures are the key drivers of sequence polymorphisms and variable in each region of the *E. histolytica* genome [185]. Therefore, it could be stated that different mRNA levels and flexibility in transcriptional regulation depend on the polymorphic levels of genes.

To confirm the hypothesis in relevance to differential virulence, the correlation between sequence variation and transcriptional variability was explored only in a pair of nonvirulent Rahman and virulent HM-1:IMSS, as shown in Figure 2.21. Compared to the 1st scatterplot in Figure 2.20, slightly less positive correlation with statistical significance (r = 0.2018, P-value = 0.0464) indicates that genetic variation likely causes transcriptional variation, contributing to differential virulence between such nonvirulent and virulent strains.

In addition, the key consideration is the influence of sequence polymorphisms on the mapping of raw read sequences to the genomic reference. The number of polymorphic sites in each strain is counted by comparing the sequence of a particular gene with the HM-1:IMSS genomic reference. Higher number of SNP sites would represent higher sequence divergence when compared to the HM-1:IMSS reference. Taking into consideration, I have used the HM-1:IMSS genomic sequence as a reference for mapping and annotation in the bioinformatic analysis. Additionally, TopHat's mapping algorithm, by default, allows one or two base differences between aligned read and the reference sequence [111,114,118]. As such, any read with many mismatches would be disregarded, resulting in underestimation of aligned reads. As plotted in Figures 2.20 and 2.21, it is suggested that transcriptional difference of genes with high sequence polymorphisms may be partly affected by artefacts resulting from mapping to the HM-1:IMSS genomic reference.

However, transcriptional variation can be indeed a consequence of not only sequence polymorphisms but also gene gain or gene loss and gene copy number variation among the strains [70]. Also, it was recently reported that antisense sRNAs can regulate transcriptional levels of *E. histolytica* parasites by the endogenous RNAi pathway in a strain-specific manner [85,86,92]. Essentially, it needs to be further investigated in functional studies to determine how the molecular evolution of sequence divergence has an influence on the transcriptional control and virulence variability in this parasite.

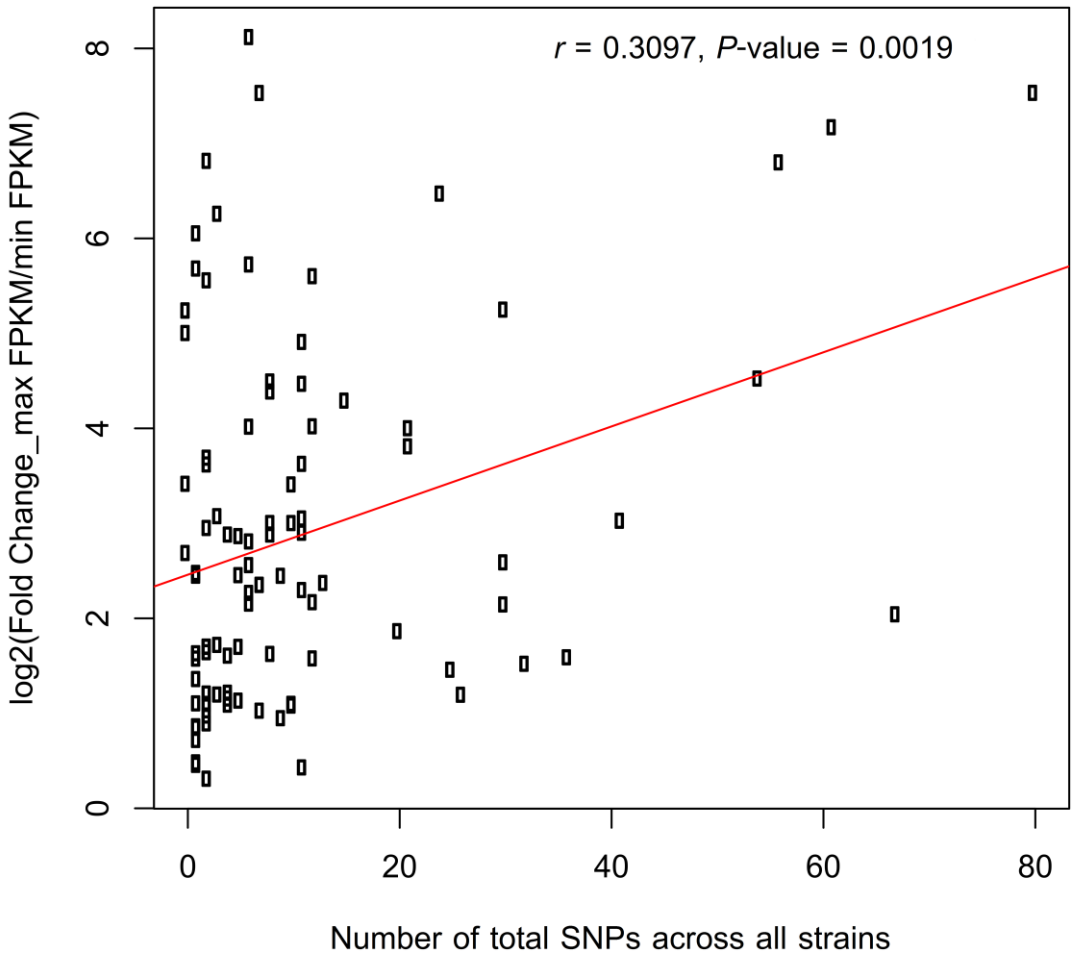


Figure 2.20: Significant positive correlation (*r* = 0.3097, *P*-value = 0.0019) between levels of single nucleotide polymorphisms and transcriptional variability of 98 DE genes among the four *E. histolytica* strains. Transcriptional variability of a particular gene is represented in terms of $\log_2$-transformed value of fold change computed by the ratio of maximum FPKM and minimum FPKM seen in these four strains. The comparatively high degree of sequence divergence is associated with a vast range of transcript levels across all strains, most likely reflecting a varied regulation of expression in such DE genes.

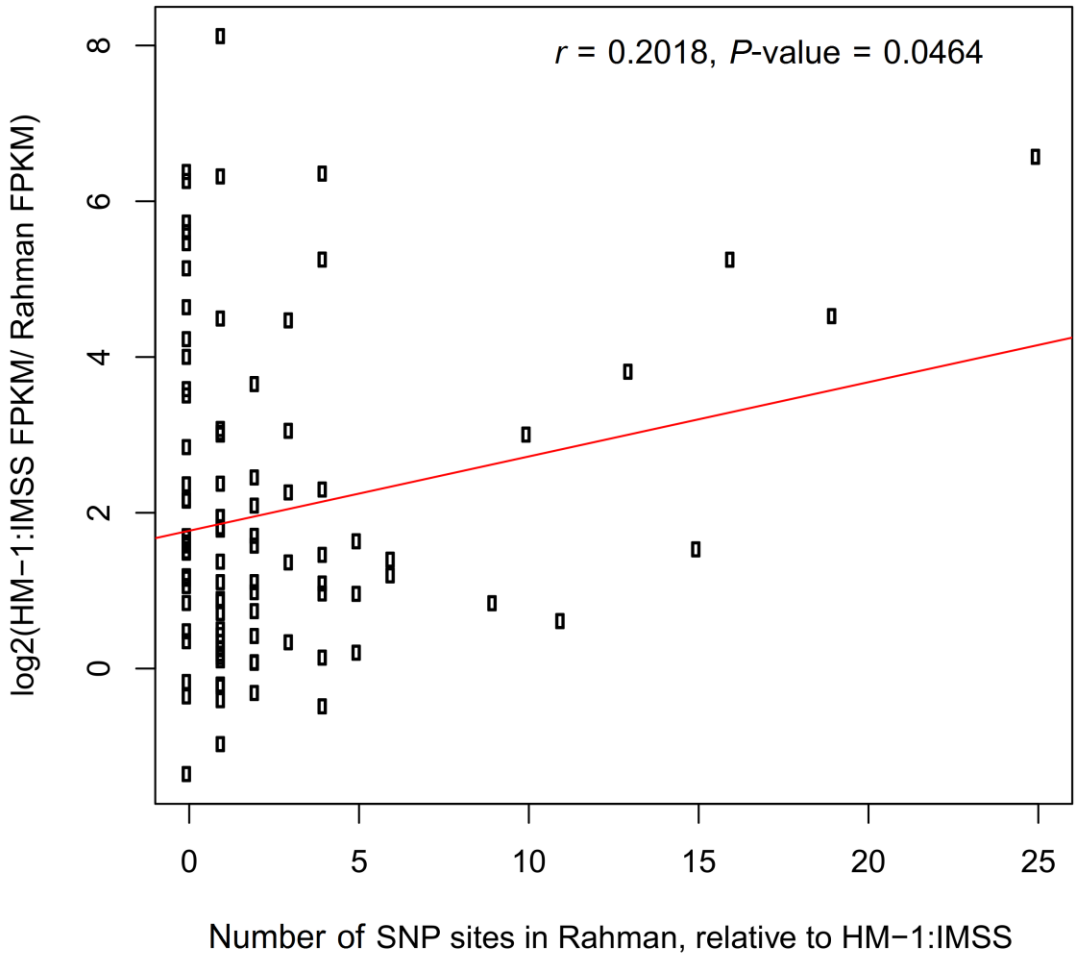


Figure 2.21: Significant positive correlation (*r* = 0.2018, *P*-value = 0.0464) between levels of single nucleotide polymorphisms and transcriptional variability of 98 DE genes in Rahman relative to HM-1:IMSS.

2.3.9 Functional characterisation and annotation of protein domain signatures reveals biological cellular functions potentially involved in virulence

Currently, protein domains and protein sequence motifs can be used as signatures of protein families for functional annotation purposes. Several bioinformatic tools have been developed to predict functional domains of a particular protein for its functional assignments, especially in enzymes which predicted domain(s) and motif(s) would define their putative functions. Also, this method is very useful for functional annotation in case of proteins with unknown function or hypothetical proteins (HPs) [186]. For *E. histolytica*, whole genomic analyses found 8,333 genes in total which include 4,478 genes encoding HPs, accounting for 53.74 % of total genes [25,46]. Therefore, the functional assignment for protein domain signatures in group of proteins of interest including HPs will provide a better framework of biological functions involved in pathogenesis and virulence in *E. histolytica* as well as prioritise protein candidates for therapeutic purposes.

Functional characterisation and annotation could be accomplished using several public protein domain databases, e.g. Pfam [187], PRINTS [188], PROSITE [189], SMART [190], PANTHER [191], TIGRFAMs [192], SUPERFAMILY [193] and PIRSF [194]. Herein, IntroProScan was applied to search functionally annotated domains and motifs using signature recognition methods of the InterPro Consortium for several databases such as Pfam, PRINTS, PROSITES, PANTHER and SMART [123]. In this present study, FASTA formatted protein sequences of 1,162 upregulated and 997 downregulated DE genes in the three virulent strains were verified against the Pfam database to identify putative functional protein domains and motifs in such two sets of DE genes. The functionally annotated protein domains and motifs are listed for the top 30 prevalent annotations as illustrated in Figures 2.22 and 2.23 for upregulated and downregulated groups, respectively.

2.3.10 Protein phosphorylation and Ras-regulated G-Protein signaling are the key regulatory processes in *E. histolytica*

Complete genome sequencing as well as genome-wide transcriptomic studies have found that *E. histolytica* possesses a large number of cell signaling molecules involved in a diverse variety of cellular processes [36,76,77]. Prominently, protein kinases (PKs) constitute a large fraction of the human protist genomes such as *Leishmania major* and *Trypanosoma spp*. [195,196]. Anamika *et al.*, 2008 identified a large protein kinase repertoire (kinome) consisted of 307 PKs in *E. histolytica* [197]. Basicly, PKs play an important role in almost all major signal transduction pathways of eukaryotic cells. Protein phosphorylation catalysed by the kinase activity functions as a regulatory switch for many cellular activities, including transcription, metabolism, cytoskeletal rearrangement, cell division and cell movement via mechanisms of signal transduction [198].

In *E. histolytica*, the interaction between trophozoites and host extracellular matrix results in induced signaling which triggers invasion [199]. A large family of over 90 transmembrane kinases (TMKs) have been identified in *E. histolytica* and thought to have nonredundant functions involved in growth and phagocytosis [200]. Differently, only nine putative TMKs were predicted in the free-living sister species, *Dictyostelium discoideum*, so the presence of large TMK gene family in *E. histolytica* promisingly suggests the necessitation of the parasite in sensing and responding to a wider variety of extracellular stimuli within the host environment, compared to the free-living condition [201]. Also, kinome analysis of *E. histolytica* revealed that the parasite possesses a complex network of protein phosphorylation implicated with several unusual PKs [197].

The Ras superfamily GTPases have been highly studied in *E. histolytica* [202]. The genome of *E. histolytica* harbors more than 170 annotated members of Ras superfamily GTPases in the AmoebaDB database, indicating an important role in G-protein signaling system [26,202]. The Ras superfamily can be categorised into five families: Ras, Rho, Ran, Rab and Arf GTPases [203]. The Ras family typically controls cell proliferation and survival. The Rho family is responsible for cell morphology, actin filament organisation, cell cycle and gene expression. The Ran family involves in nuclear-cytoplasmic transport. Lastly, the Rab and Arf families regulate vesicular transport [202,204]. However, experimental studies of Ras signaling as well as Ras regulators, i.e. GEFs and GAPs, are still understudied. [202].

As shown in Figure 2.22, the most prevalent domain found in the cluster of 1,162 upregulated genes in the three virulent strains is a PK domain (PF00069). Also, the Ras family (PF00071) belonging to a Ras superfamily GTPase family as mentioned above is

ranked as the top third order (n = 15). Likewise, for the cluster of 997 downregulated genes, PK domain and Ras family are remarkably more prevalent than others, as shown in Figure 2.23. As shown in Figures 2.22 and 2.23, downregulated PK (n = 45) and Ras family domains (n = 31) are twice in number of proteins consisting of these two domains when comparing to those in the upregulated cluster (n = 21 for PK and 15 for Ras family).

Similarly, the previous DGE tests reveal transcriptomic modulations of genes mostly involved in protein phosphorylation and G-protein signaling. As listed in Table 2.10, most of the functional gene annotations involved in protein phosphorylation and G-protein signaling show higher number of downregulated transcripts than those of upregulated transcripts. In agreement with the DGE tests, the InterProScan results show the different distribution of these protein domains between upregulated and downregulated clusters. It is striking that the PK and Ras family domains were much more obviously downregulated compared to other domains in the downregulated cluster as shown in Figure 2.23.

From this observation, different members of the same conserved family involved in phosphorylation and signaling were upregulated or downregulated at the same time. It is worth noting that the genome of simple protozoan parasite *E. histolytica* contains indeed a large portion of conserved gene families linked to signaling pathways [197,200]. Therefore, the interesting question how parasites can accurately regulate the expression of a subset of such large multigene families still needs to be further elucidated, possibly for epigenetic mechanisms.

Besides the Ras family, the Rho family GTPases as well as their two Rho regulators: guanine nucleotide exchange factor (RhoGEF) and GTPase-activator protein (RhoGAP) primarily regulate the dynamics of actin cytoskeleton and actin filament-based processes in *E. histolytica*, including movement, phagocytosis, tissue invasion as well as surface receptor capping for host immune evasion [199,205-207]. Therefore, regulation of actin dynamics plays an important role in pathogenesis-related processes as well as trophozoite survival. Indeed, 22 Rho family GTPases (EhRhos) were identified in the *E. histolytica* genome [208]. Surprisingly, the InterProScan results do not show any upregulation or downregulation of Rho domains of these 22 Rho family GTPase members in this study. However, both RhoGEF and RhoGAP domains were found to be upregulated and downregulated as shown in Figures 2.22 and 2.23, respectively. In the three virulent strains, 7 RhoGEFs and 7 RhoGAPs were upregulated but conversely, 12 RhoGEFs and 11 RhoGAPs were downregulated. Essentially, it is likely to imply that virulent parasites have a specific molecular switch system that can activate or inhibit expression of gene members in signaling pathways including PKs, Ras family GTPases, RhoGEFs and RhoGAPs as mentioned above.

Similar to previous transcriptomic studies, protein kinases, RhoGAPs and protein phosphatases were found to be modulated in both HM-1:IMSS and Rahman in the axenic condition, suggesting the possible allele-specific expression between strains [76]. It was also reported for the upregulation of PKs, TMKs, Ras and Rho family GTPases in HM-1:IMSS trophozoites inoculated to the mice colon, strongly indicating their important role in adaptation to the host environment [77].

As explained above, a number of these two domains constitute a large fraction of both upregulated and downregulated proteomes, reflecting the great impact of signaling in regulating diverse cellular processes of this parasite. These two major regulatory pathways, i.e. PK-dependent and G-protein signaling pathways, enable trophozoites to interact with a vast variety of extracellular signal cues, essential for their survival and host-parasite interaction. Therefore, this current study reveals that protein phosphorylation and Ras-regulated G-Protein signaling are the key essential steps for regulating a wide variety of cellular processes and the transcriptional modulations of such major signaling pathways potentially result in differences of trophozoite pathogenicity and virulence among *E. histolytica* strains.

2.3.11 Co-upregulation of actin cytoskeleton and actin-modulating domains indicates the increase of actin-filament based processes in virulent parasites

As previously reported, actomyosin cytoskeleton centrally contributes to *E. histolytica* pathogenesis due to its diverse functions directly involved with reorganisation of cellular component, cell movement and morphological changes, cell division, phagocytosis, host cell adhesion as well as interaction with host extracellular matrix [206,209]. Trophozoites with highly active motility would have an advantage for moving from an ulcerative lesion site to the bloodstream and subsequent hematogenous spreading to extraintestinal organs [209]. Also, the actin cytoskeleton plays a crucial role in maintaining structural integrity of the parasite adhesion molecules, i.e. Gal/GalNAc lectins at the host cell adherence site [210]. It was evidenced in genetically engineered *E. histolytica* strain LMM that both *in vitro* and *in vivo* parasite motility and host cell cytotoxicity were drastically reduced by disruption of cytoskeletal myosin II activity, indicating that virulence is regulated by the amoebic cytoskeleton [211,212].

Based on my protein domain data, not only actin (PF00022) but also other actin-modulating domains were co-upregulated, emphasising the important role of the actin cytoskeleton in virulent trophozoites. Other actin-binding domains were also found to be upregulated in the InterProScan result, such as calponin homology (CH) domain (PF00307),

zinc-binding domain present in **L**in-11, **I**sl-1 & **M**ec-3 (PF00412), Wiskott-Aldrich syndrome homology region 2 (WH2) domain (PF02205), gelsolin repeat (PF00626) and cofilin/tropomyosin-type actin-binding protein (PF00241) as demonstrated in Figure 2.22.

First, many cytoskeletal proteins contain two copies of the CH domain in a tandem arrangement [213]. Also, a single CH domain could be found in regulatory proteins of the signal transduction pathways [214,215]. The microarray study of Davis *et al*, 2007 showed the significant upregulation of the CH domain-containing protein (XM_652357.1) with $\log_2$FC = 3.0, *P*-value = $2.35e^{-4}$ in HM-1:IMSS compared to Rahman [76]. Second, LIM domain is a cysteine and histidine rich domain containing two zinc fingers. This domain plays a role in cytoskeletal reorganisation and protein-protein/protein-DNA interactions [215,216]. Differential in-gel 2D electrophoresis of the proteomes of HM-1:IMSS and Rahman, performed by Davis *et al.,* 2006 showed the upregulation of six proteins in HM-1:IMSS, including a LIM domain-containing protein [16]. Third, WH2 motif and cofilin domain can be found in Wiskott-Aldrich syndrome protein (WASP) and suppressor of cAMP receptor (SCAR). These WASP/SCAR family proteins function as nucleation-promoting factors in concert with the Arp2/3 complex [215,217-219].

This protein domain data show the consistence with my RNA-Seq result presenting the upregulation of six actin genes and three genes encoding actin-binding proteins (cofilin/tropomyosin family) in all three virulent strains as listed in Table 2.6. However, the marked upregulation of actin transcripts with $\log_2$FC ≥ 2 could be found in only HM-1:IMSS. In addition, HM-1:IMSS displays more than 4-fold higher expression of Arp2/3 complex 21 kDa subunit (EHI_174910) than the other strains, implying its higher capability to trigger the actin nucleation and subsequent actin filament-based processes, compared to other virulent strains.

2.3.12 Increase of proteolysis-related transcripts suggests the high protein turnover rate and active metabolism in virulent parasite strains

The proteasomal degradation pathway is important for several cellular processes in all cells and tissues of eukaryotic organisms, including control of gene expression, cell cycle, development, as well as rapid protein turnover [220,221]. Previously reported by Dustin *et al*., 2013, *E. histolytica* trophozoites have a remarkable ubiquitin-dependent protein degradation system [221]. Proteasome inhibitors can retard trophozoite growth in *E. histolytica* as well as encystation process in *E. invadens* [222]. Also, transcriptomic changes in this protein turnover pathway were reported to be associated with variation of virulence among strains [77,82]. Most recently, Thibeaux *et al.*, 2013 demonstrated that there were

significantly marked upregulation of ubiquitin (EHI_083410 and EHI_178340) and probable proteasome subunit beta type 2 (EHI_078710) in HM-1:IMSS in response to contact with the human colonic explant [82]. Herein, proteins with proteasome subunit domain (PF00227) and proteasome subunit A N-terminal signature (PF10584) were found to be exclusively upregulated (n = 11 and 7, respectively) in the three virulent strains as demonstrated in Figure 2.22. Therefore, it seems that upregulation of proteolysis-related genes in the transcriptomes of virulent strains would indicate the high protein turnover rate as well as the active metabolic state in virulent parasites, potentially contributing to their virulence.

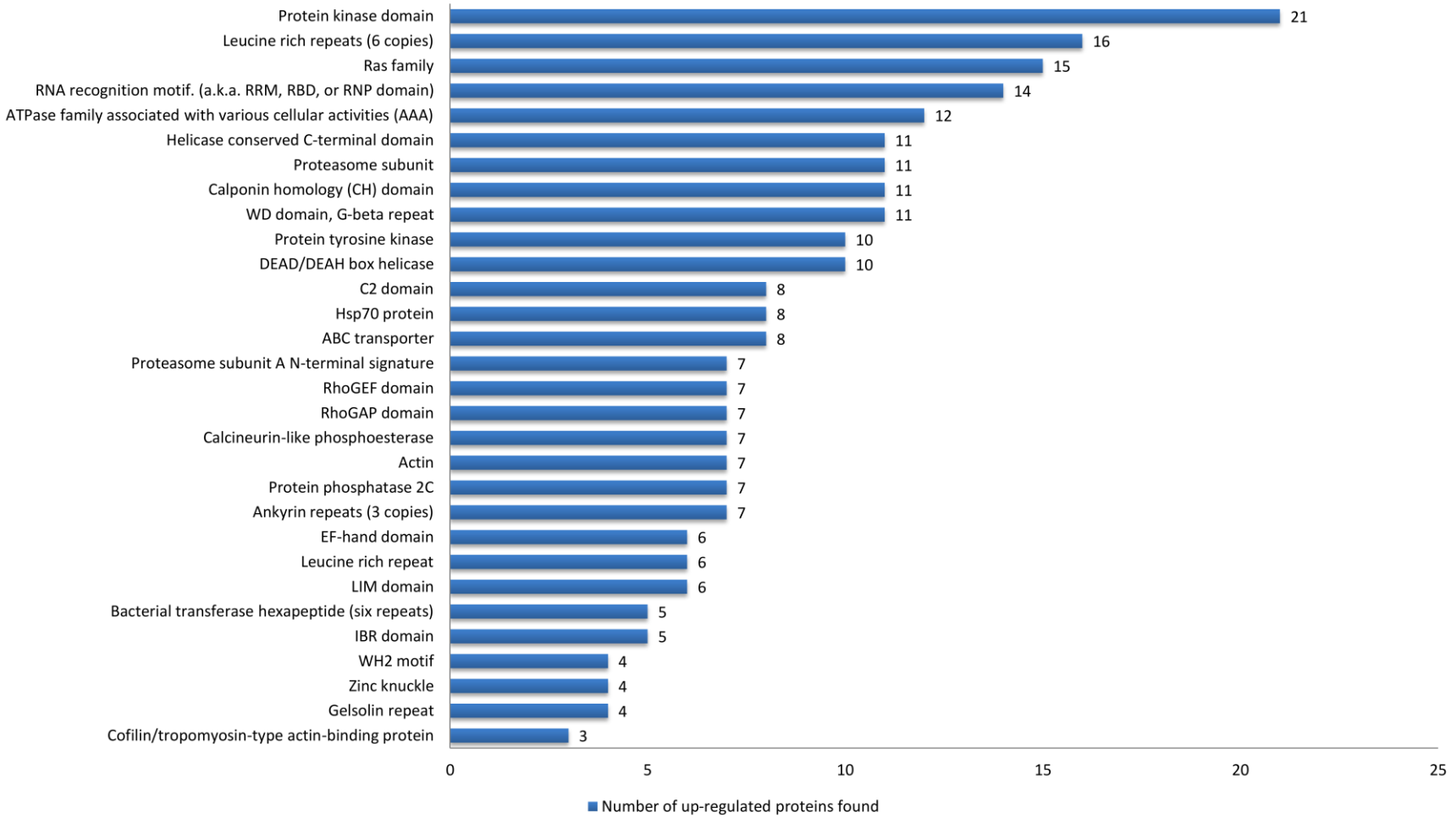


Figure 2.22: The 30 most prevalent functionally annotated protein domains/motifs found in 1,162 upregulated DE proteins in the three virulent strains (i.e. PVBM08B, HM-1:IMSS and IULA:1092:1). The abbreviations represent as follows: Ras = Ras subfamily of RAS small GTPases; AAA = ATPases associated with a variety of cellular activities; CH = Calponin homology; WD = Beta-transducin repeat; C2 = Protein kinase C conserved region 2 (CalB); Hsp70 = 70 kilodalton heat shock protein; ABC = ATP binding cassette; RhoGEF = Guanine nucleotide exchange factor for Rho/Rac/Cdc42-like GTPases; RhoGAP = GTPase-activator protein for Rho/Rac/Cdc42-like GTPases; LIM = Zinc-binding domain present in Lin-11, Isl-1 & Mec-3; IBR = In Between Ring fingers; WH2 = Wiskott-Aldrich syndrome homology region 2.

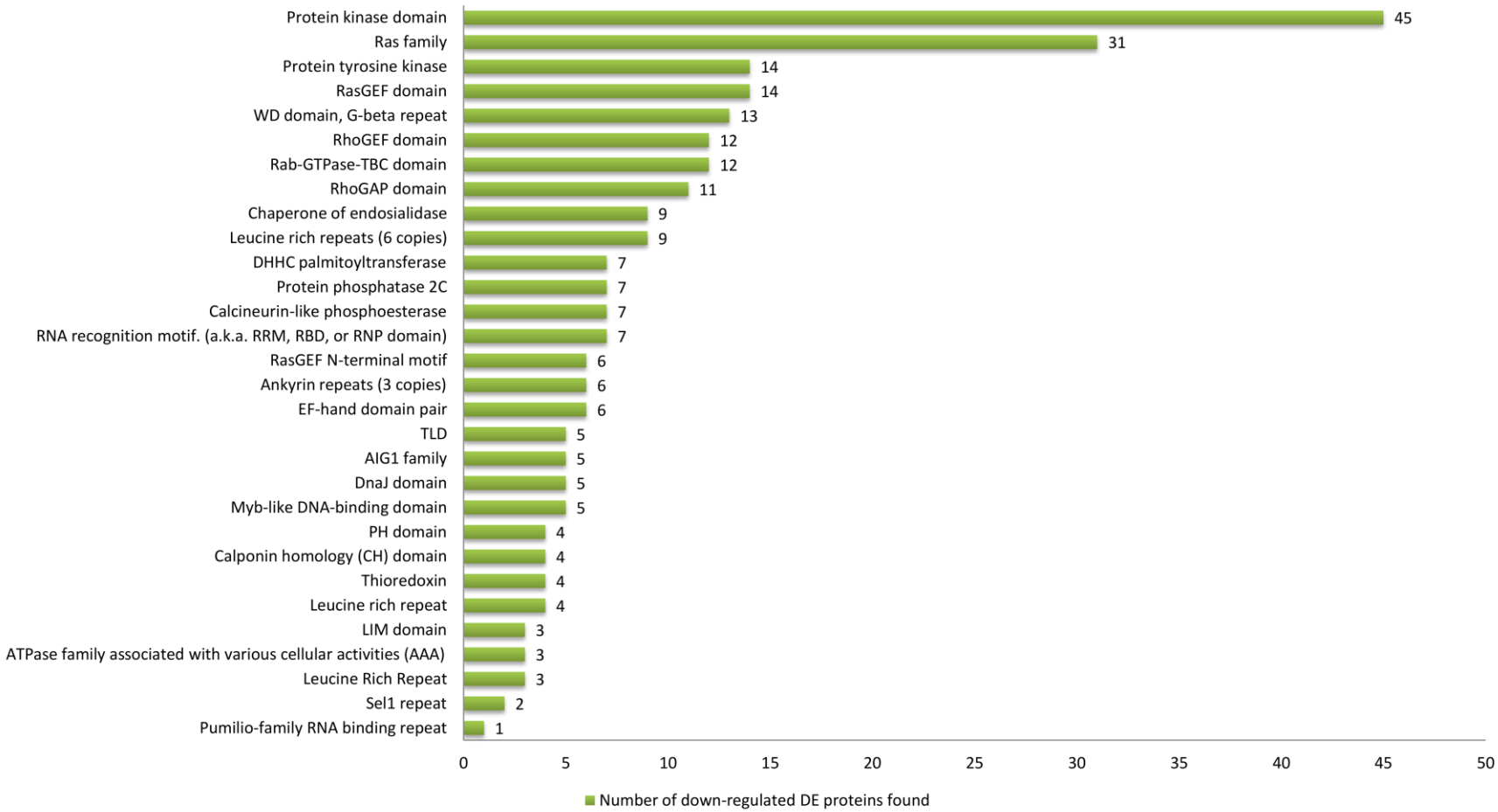


Figure 2.23: The 30 most prevalent functionally annotated protein domains/motifs found in 997 downregulated DE proteins in the three virulent strains. The abbreviations represent as follows: RasGEF = Guanine nucleotide exchange factor for Ras-like small GTPases; Rab = Rab subfamily of small GTPases; TBC = Domain in Tre-2, BUB2p, and Cdc16p; TLD = TBC/LysM-associated domain; AIG1 = AvrRpt2-induced gene-1; DnaJ = 40 kilodalton heat shock protein; PH = Pleckstrin homology; Sel1 = Sel1-like repeats.

2.3.13 GO enrichment analysis

After getting the list of 1,162 upregulated and 997 downregulated DE genes commonly found in three virulent strains of *E. histolytica*, i.e. PVBM08B, HM-1:IMSS and IULA:1092:1 as shown in Figures 2.13 and 2.15, GO enrichment analysis (available at http://AmoebaDB.org) was applied for these two sets of DE genes. Given a list of DE genes which are upregulated or downregulated as above, the ontology enrichment analysis will explore which terms have overrepresentation or underrepresentation compared to the rest by calculating a ratio of fold enrichment (the percentage of genes annotated to a GO term of interest in the sample divided by the percentage of genes with the identical term in the background) [223-225]. These over-represented or under-represented GO terms would reflect the predominant ontologies and lead us to better understanding in biology of this parasite's virulence.

I. GO terms identified in upregulated gene cluster

For the 1,162 upregulated DE genes, thirty-five gene functional categories could be identified for the biological process ontology with significant statistics as shown in Appendix Table 4. In general, these DE genes are responsible for several cellular processes, i.e. biosynthesis, protein and macromolecule catabolism, DNA metabolism, actin filament-based process and stress response. Most of these upregulated genes have functions involved in cellular and organic substance biosynthetic processes (94 genes for GO:0044249 and 95 genes for GO:1901576). However, the numbers of genes in the background for these two categories are 496 and 504, accounting for the lowest fold enrichment of 1.21 and 1.2 for GO:0044249 and GO:1901576, respectively. This indicates that the number of DE gene members in each GO term does not reflect the real overrepresentation due to sample size bias. In contrast, two sets of GO terms (1^{st} : GO:0044419, GO:0016032, GO:0044764, GO:0044403 and GO:0051704; 2^{nd} : GO:0000278, GO:0006020 and GO:0007067) show the same number of genes in both the sample and the background (n= 4 for the 1^{st} set and 3 for the 2^{nd} set), resulting in 100 % of background genes in the sample and the highest fold enrichment of 6.37. It reveals that all DE genes with such GO terms have overrepresentation in this gene set, suggesting their possible roles in virulence.

Obviously, from enrichment analysis, most of upregulated DE genes have GO terms involving in key metabolic pathways of the cell. These findings can explain in relevance to the virulence of these strains. For instance, five upregulated GO terms (GO:0030029, GO:0007015, GO:0030036, GO:0008154 and GO:0007010) are responsible for actin filament-based process and cytoskeleton organisation, implying that these three virulent

strains are likely to have better capability of movement, phagocytosis, tissue invasion and surface receptor capping than the nonvirulent Rahman strain. Also, this finding corresponds to the InterProScan result showing the upregulation of actin and other actin-binding domains.

It is interesting that a GO term responsible for 'viral process' (GO:0016032) could be found in the upregulated gene list, making us hypothesise that intestinal viral infection in these virulent strains may help enhance the expression level of this gene set and lead the trophozoites to proliferative and virulent state. Also, other four GO terms: GO:0044419, GO:0044764, GO:0044403 and GO:0051704, which involve in interspecies interaction and parasitism, are also found with 100 % sample frequency and fold enrichment of 6.37, supporting the above hypothesis. Alternatively, these axenically cultured virulent strains might possess evolutionarily some virus-derived genes, intimately affecting to the genome and transcriptome of the parasite and contributing to their pathogenic bahaviour.

As mentioned in Chapter 1, the invasive trophozoites are prone to be battled by host immune defence such as ROS, NO and cytotoxic enzymes [14]. So, it could be explained why these virulent strains have high expression levels of proteins responsible for stress response (GO:0006950 and GO:0033554), compared to Rahman strain.

The ontology terms related to protein and macromolecule catabolism: GO:0030163, GO:0044257, GO:0043632, GO:0019941, GO:0006511, GO:0051603, GO:0009057 and GO:0044265 are listed in the first eight rows of Appendix Table 4 with fold enrichment values of 2.17 to 2.57. The overpresentation of these terms reflects the increased protein turnover activity in these three virulent strains. This is also consistent with the InterProScan result in Figure 2.22, obviously showing the upregulation of proteasome domain as well as proteasome subunit A N-terminal signature.

II. GO terms identified in downregulated gene cluster

As detailed in Appendix Table 7, a total of 997 downregulated DE genes were identified into fourty-four gene functional categories for the biological process ontology with significant statistics. Strikingly, 27 of 44 GO terms are involved in regulation of cellular pathways such as 'regulation of response to stress' (GO:0048583), 'biological regulation' (GO:0065007), 'regulation of nucleoside metabolic process' (GO:0009118), 'regulation of signaling' (GO:0023051), 'regulation of cell communication' (GO:0010646), 'regulation of molecular function' (GO:0065009), 'regulation of cellular catabolic process' (GO:0031329) and 'regulation of phosphate and phosphorus metabolic processes' (GO:0019220 and

GO:0051174). Also, other ontologies are mainly responsible for 'cell communication' (GO:0007154), 'signaling' (GO:0023052), 'protein phosphorylation' (GO:0006468), 'phosphorus metabolic process' (GO:0006793), 'tRNA metabolic process' (GO:0006399), 'macromolecule modification' (GO:0043412), 'phosphate containing compound metabolic process' (GO:0006796) and 'ncRNA metabolic process' (GO:0034660).

To assess underrepresentation throughout all downregulated GO terms, it was found that fold enrichment values vary from 1.23 ('macromolecule modification', GO:0043412) to 2.27 ('tRNA processing', GO:0008033), less variable between terms than the upregulated group. Surprisingly in this cluster, gene functional categories could be generally divided into two related groups. For instance, ontologies of 'signaling' or 'intracellular signal transduction' (GO:0023052 and GO:0035556) and 'regulation of signaling' or 'regulation of intracellular signal transduction' (GO:0023051 and GO:1902531) could be found together in this DE list. Other pairs of associated terms are also found as well, i.e. 'phosphorylation' or 'protein phosphorylation' or 'phosphorus metabolic process' (GO:0016310, GO:0006468 and GO:0006793) and regulation of phosphate and phosphorus metabolic processes (GO:0019220 and GO:0051174). This observation guides that it seems to have reduced expressions of both functional proteins and their associated regulators involved in such particular pathways in virulent strains. Almost all downregulated GO terms are implicated in key controlling pathways of the parasitic cell. Therefore, it is reasonable to hypothesise that less strict cellular control due to reduced expression of such protein members in regulatory pathways would be able to lead trophozoites to the virulent state.

However, there was still redundancy amongst these ontologies since one particular protein could be identified by more than one GO term. This can be overcome by a specialised software named '**REVIGO**', capable of clustering similar GO terms into only a single representative, enabling comprehensive interpretation in three different presentations.

2.3.14 Summarisation and visualisation of enriched gene ontologies

To summarise and interpret biological meanings of GO terms, lists of GO terms with FDR-adjusted *P*-values from the previous enrichment analysis were analysed by an online web server, 'REVIGO'. The REVIGO (**Re**duce and **Vi**sualise **G**ene **O**ntology) is a web server designed with a simple clustering algorithm for summarising long and semantically similar list of GO terms into a cluster with a single representative GO term [124,226-228]. Also, the REVIGO can visualise these non-redundant cluster representatives into three different ways for interpretation, i.e. scatterplot, interactive graph and treemap, discussed further.

The clustering algorithm would provide two values of anticorrelated parameters: uniqueness and dispensability. The 'uniqueness' value refers to a degree of negativeness of average similarity of such term to the whole list. It determines whether such GO term of interest differs or detaches from all other members of list. By contrast, the 'dispensability' value represents a degree of redundancy of such term compared to other semantically similar terms. To eliminate redundancy, semantically close terms with higher values of dispensability would be united into the main cluster represented by a semantically similar GO term with less dispensability and more significant adjusted *P*-value.

For instance, the two upregulated GO terms: 'organelle organisation' (GO:0006996) and 'cellular component organisation' (GO:0016043) have dispensability values of 0.704 and 0 with $\log_{10}$(FDR-adjusted *P*-value) of -2.0768 and -2.5452, respectively. This former term 'organelle organisation' was found to share relatively high semantic similarity with the term 'cellular component organisation' which has lower dispensability and more significant adjusted *P*-value as shown in Appendix Table 9. To reduce redundancy, the term 'cellular component organisation' was therefore chosen as a cluster representative for illustrative purposes.

To determine their closeness, each cluster representative obtained after the clustering algorithm finished was assigned for X and Y coordinates in the scatterplot so that GO terms with more semantic similarities would be closer. This could be accomplished by multidimensional scaling-based visualisation with the pairwise distance matrix. On the plot, each cluster is represented in bubble with different colour and size. Additionally, the column of frequency as listed in Appendix Tables 9 - 13 refers to the percentage of UniProt proteins annotated with a GO term in the underlying Gene Ontology Annotation (GOA) database. The user-provided FDR-adjusted *P*-value and frequency of each GO term are represented by bubble colour and size, respectively.

2.3.15 Many biological process ontologies in protein catabolism, biosynthesis, mitotic cell cycle, DNA metabolism, repair and recombination, stress response as well as actin dynamics are overrepresented in the transcriptomes of virulent strains

As shown in Figure 2.24 and Appendix Table 9, thirty-five GO terms for the 1,162 upregulated DE genes were reduced to twenty-one clusters. The larger size of the bubble does not display a fold enrichment value or a sample frequency of ontology but particularly denotes a higher protein frequency of such GO term in the underlying GOA database, indicating a more general term. The level of statistical significance, i.e. $\log_{10}$(FDR-adjusted *P*-

value), is demonstrated in a continuous range of colour spectrum (red, orange, yellow, green and blue). Semantic similarities between GO terms are associated to their closeness on the scatterplot. For instance, four clusters (light blue and blue bubbles) representing GO terms involved in protein and macromolecule catabolism are grouped together (plot_X = -5.841, plot_Y = 3.702).

Notably, there are upregulation of catabolic and anabolic processes in the transcriptomes of virulent parasites as shown in Figures 2.24 and 2.25. The interconnection between clusters of protein and macromolecule catabolism is consistent with the DGE result showing the upregulation of genes encoding ubiquitin-conjugating enzyme family protein in virulent strains. Similarly, the InterProScan result in Figure 2.22 also confirms the upregulation of proteasome subunit domain (PF00227) and proteasome subunit A N-terminal signature (PF10584). Correspondingly, the interactive graphs and treemaps of component and function ontologies, as shown in Figures 2.27 and 2.29, reveal the upregulation of terms representing proteasomal complex and threonine-type endopeptidase activity, respectively, strongly supporting the rapid protein turnover in virulent parasites.

As discussed before, phagocytosis is a hallmark process for virulent parasites to invade and survive against host immune cells. Also, my RNA-Seq data demonstrate the evidence of increased expression of phagocytosis-related genes such as genes encoding C2 domain-containing proteins, actin and cytoskeleton-associated proteins. Therefore, the results suggest that it is highly possible that phagocytosis would be a potential driving process for protein and macromolecule catabolism in virulent parasites. Ultimately, the interconnection between catabolic and anabolic ontologies in Figure 2.25A suggests that such proteolysis and macromolecule degradation would likely drive the parasites for rapid growth and proliferation by increasing the rate of translation and biosynthetic processes.

It is noticeable that clusters of ontology concerning ‘response to stress’ (GO:0006950), ‘DNA repair’ (GO:0006281) as well as ‘DNA recombination’ (GO:0006310) were found together in the enrichment analysis, pointing out the prospective relationship amongst these terms. Intriguingly, the interactive graph also unveils the close relationship of such three clusters. These three clusters are linked as shown in Figure 2.25A, revealing that there is a significant relationship among these three upregulated clusters.

Intestinal parasites are prone to be continuously attacked by host immune response and strong environmental factors which could make changes to their genomic integrity and stability [229-231]. Structural damage of DNA can cause all types of mutation including

point mutation, insertion, deletion and translocation, requiring cellular DNA repair machineries. Therefore, overrepresentation of such process ontologies implies that these virulent strains have a greater potential to eliminate genomic lesions and maintain their genomic stability under stress conditions. Also, it is apparent that there is upregulation of four clusters showing active mitotic cell division in virulent strains as illustrated in Figure 2.25. Therefore, it seems to explain that upregulated expression of gene clusters involved in DNA repair, recombination and DNA metabolism in the transcriptomes of virulent strains would potentially enable the parasites to correct unwanted genetic damages caused by active mitotic cell division.

Previous sequence analysis by Weedall *et al*., 2012 showed the evidence of gene conversion in virulence-associated genes transcribed for the Gal/GalNAc lectin complex [38]. Gene conversion is the process of non-reciprocal homologous recombination by which one DNA region is replaced by its homologous sequence to have identical sequences after the recombination event [232]. This gene conversion exists favourably amongst regions of multigene family members due to their relatively high sequence homology. So, this finding strongly indicates that homologous recombination (HR) can be present and play a biological role in the *E. histolytica* genome, especially driving the molecular evolution of gene families potentially involved in virulence variation [38,231].

Essentially, HR is a conserved biological mechanism most extensively undergone by organisms to precisely repair DNA double strand breaks and to rescue the break point that interrupts DNA polymerase during DNA replication [233,234]. Also, HR is an important mechanism required for telomere maintenance, meiosis, and sexual reproduction [234-236], but obvious sexual means in *E. histolytica* has not yet been demonstrated before [230,237]. Despite of difficulties in genetic studies in parasitic protists, characterisation of meiotic genes and HR specific genes has been demonstrated in the genome sequence data of many species [237-239]. Some meiotic genes such as *DMC1*; *MND1*; *SPO11* as well as HR specific genes such as *MLH1*; *MSH2*; *RAD21*; *RAD51* were found in the genomic data of *E. histolytica*, strongly suggesting the possible sex and the key mechanism of DNA repair in this species [238,240]. Interestingly, ploidy changes and unscheduled gene amplification previously described in *Entamoeba* species might be driven by the process of DNA recombination [37,241,242]. Also, HR can occur in other human parasitic protozoa, i.e. *Plasmodium*, *Trypanosoma* and *Leishmania* [243-245]. For *T. brucei* and *P. falciparum*, HR was found to be critical to parasite survival by generating antigenic diversity implicated for evasion of the host immune response [243,244].

Recently, Singh *et al*., 2013 have proved that expression of meiotic and HR genes were upregulated under induced stresses, i.e. serum starvation, heat shock, oxygen stress and UV radiation in *E. histolytica* and during encystation in *E. invadens* [230]. Also, HR was directly evidenced in inverted repeat plasmid-transfected trophozoites following different stress conditions in *E. histolytica* and under stage conversion in *E. invadens* [230].

In this study, the close relationships between upregulated clusters involved in DNA repair and recombination as well as response to stress in the three virulent strains (i.e. PVBM08B, HM-1:IMSS and IULA:1092:1) highlight the capability of virulent trophozoites to circumvent their DNA damage under strong stress conditions in the host. Essentially, recombinational DNA repair system can improve the fitness of parasites by allowing increased survival of descendents with repaired DNA. In addition, it is indeed evolutionarily advantageous for parasites because DNA recombination can generate novel genotypes that can resist to host negative selective pressures and rapidly disseminate through host populations. Taken together with the presence of meiosis-related genes in the *E. histolytica* genome, the overrepresentation of gene ontologies related to DNA repair, recombination and stress response emphasises that sex potentially occurs in this parasite.

The enrichment analysis data also suggests that 'inositol metabolism' (GO:0006020) might be partly responsible for virulence characteristics. Principally, phosphoinositides, phosphorylated forms of phosphatidylinositol (PI), play important roles in a vast variety of cellular processes such as proliferation, cytoskeletal rearrangement and membrane trafficking [246]. Phosphatidylinositol 3-kinases (PI3Ks) have catalytic functions in phosphorylating the inositol ring at D3 hydroxyl group and produce active lipid derivatives including phosphatidylinositol 3-phosphate [PI(3)P], phosphatidylinositol 3,4-bisphosphate [$PI(3,4)P_2$], phosphatidylinositol 3,5-bisphosphate [$PI(3,5)P_2$] and phosphatidylinositol 3,4,5-triphosphate [$PI(3,4,5)P_3$]. Towards the biological importance, PI3K signaling is a key regulatory pathway for phagocytosis, cell motility and chemotaxis [247,248].

Blazquez *et al.*, 2008 discovered that chemotaxis towards pro-inflammatory cytokine TNF in *E. histolytica* was inhibited in the presence of PI3K inhibitor, wortmannin (Wm) [249]. The Wm-treated trophozoites were unable to migrate towards TNF due to loss of ability to reorganise the cytoskeleton through PI3K-dependent pathways during chemotaxis. Microarray analysis also revealed the upregulation of the Gal/GalNAc lectin and certain cytoskeleton dynamics-related proteins during chemotaxis towards TNF. Interestingly, both actin (EHI_159150) and actin modulating proteins such as gelsolin repeat protein (EHI_009570) and Cofilin-like protein (EHI_054800) were also transcriptionally upregulated during TNF-induced chemotaxis [249]. This is consistent with

my InterProScan result showing the upregulation of actin dynamics-related proteins. Together with upregulated ontology of inositol metabolism, it is fair to state that the three virulent strains have greater potential to initiate directional cell polarisation, motility and chemotaxis, compared to the nonvirulent Rahman strain.

In addition, it was recently found that PI3K-mediated pathways also affect the proteolytic activity, the phagocytic capacity as well as the ability to develop amoebic liver abscess *in vivo* [250]. Essentially, this observation also emphasises the predominant role of protein kinome and the networks of protein phosphorylation in controlling pathogenesis and virulence in *E. histolytica*.

Two clusters of actin-filament based process (GO:0030029) and chromosome organisation (GO:0051276) are also upregulated in these three virulent parasites, indicating the increase of cell motility, phagocytosis and mitotic cell division. This finding is in accordance with the InterProScan result in Figure 2.22, showing upregulation of the actin cytoskeleton and its modulating proteins. Increased transcription of cytoskeleton-related proteins reflects the role of actin dynamics in regulating many cellular processes in the virulent parasites.

It could be summarised for that increase of cell cycle process indicates rapid proliferation of the virulent parasites. Concomitantly, upregulation of genes responsible for DNA metabolism, repair and recombination might be as a consequence of many mitotic cell divisions. Also, increased translation and biosynthetic processes could be driven by nutrients and energy derived from increased protein and macromolecule catabolism, possibly due to enhanced phagocytosis. Upregulation of actin filament-based process reflects the rapid cytoskeletal dynamics served for the increase of cell motility, phagocytosis and cell division. Conclusively, as the human host can be infected by multiple parasite species and also counteract the parasites with effective immune responses, the enhancement of such above catabolic and anabolic biological processes potentially provides synergy and competitive advantages to the parasites to be better able to rapidly grow and survive under the strong environmental stress in the host.

2.3.16 Downregulation of process ontologies involved in protein phosphorylation, signaling and regulation of response to stimulus indicates less stringency in biological regulations in virulent parasites, possibly leading to host tissue invasion

Contrastedly, the interactive graph of downregulated categories as shown in Figure 2.31A shows the functional network of regulatory process ontologies. Highly similar GO clusters involved in phosphorylation, signaling and regulatory processes are interconnected, indicating a less strict cellular control in these three virulent strains. In addition, phosphotransferase (kinase) function ontologies are found to be downregulated as shown in Figures 2.32 and 2.33.

Essentially, such notable downregulations in regulatory process and kinase function ontologies are consistent with the previous DGE result revealing a larger number of downregulated genes than upregulated genes for the functional annotations responsible for protein phosphorylation and signaling pathways as listed in Table 2.10.

This elaborate network provides us the new evidence that downregulation of ontologies involved in cellular regulatory processes such as protein phosphorylation and signaling seems to trigger these three virulent parasite strains to be a tissue invading form due to their nonstrict cellular control. It is simply reasonable to explain that the less stringency in cellular control enables the parasites to be 'greedy' and prioritise the expression of genes directly responsible for phagocytosis, macromolecule catabolism, biosynthesis, mitotic cell division and DNA metabolism/repair/recombination as shown in Figures 2.24 and 2.25. As explained before, such upregulated biological processes can drive the virulent parasites to rapidly proliferate and subsequently invade the host tissues.

In the light of evolution, the expression of aberrant characters in virulent *E. histolytica* parasites corresponds to the short-sighted evolution hypothesis proposed by Levin and Bull, 1994 [251]. The gist of this hypothesis is that mutant parasites, which possess greater potential to 1) increase their rapid reproduction; 2) invade and proliferate in the host tissues where there is low competition from parasite members and co-infecting species; 3) escape the host immune response, would gain 'short-sighted' local advantage and subsequently enhance their virulence in the host, even though their increased virulence would, indeed, decrease the chance of dispersal to other new hosts [251,252].

The classic examples of this proposed evolutionary mode are bacterial meningitis caused by *Haemophilus influenzae*, *Streptococcus pneumoniae* and *Neisseria meningitidis* and

poliomyelitis caused by the poliovirus [251,252]. Actually, almost all human beings are infected by such pathogens, nevertheless only few develop the disease. Normally, *H. influenzae*, *S. pneumoniae* and *N. meningitidis* colonise in the upper respiratory tract and can infect the new host via the aerosol droplet whilst the poliovirus multiplies in the host digestive tract and disseminate to the new host by ingestion of contaminated food and water [253,254]. However, clinical manifestations are developed by their invasion into the central nervous system where these pathogens lose their capability to transmit. Intriguingly, Levin, 1996 reported that parasites with this short-sighted evolution would show genetic difference and have higher capability to multiply at invasive sites than their ancestor [252].

In other words, the nonvirulent Rahman strain possesses more strict biological regulations than the other three virulent strains. It seems that tight cellular regulations in nonvirulent parasites enable themselves to sense and correctly respond to the environmental stimuli and eventually transmit their offsprings to other hosts. This finding may promisingly explain why asymptomatic cyst passers are more prevalent (approximately 90% of clinical case reports) than invasive cases [3,6]. Correspondingly, the 'Commensal Theory' proposed by Kuenen WA and Swellengrebel NH, 1913 stated that *E. histolytica* normally acts as a gut commensal responsible for multiplication and transmission to the new host and certain unknown stimuli can trigger the trophozoites to be a invasive form which is not a typical stage of the life cycle and no longer able to cause the new infection since cyst production cannot occur within the host tissues [81,255]. Based on the trade-off hypothesis, parasites would tend to decrease their virulence in a compromising way to ultimately improve their chance to reproduce and spread to a new host [256,257]. Therefore, it could be stated that the invasive behaviour of the virulent parasites is not evolutionarily advantageous since their atypical behavior actually reduces their overall reproductive fitness, like 'committed suicide' [81,255].

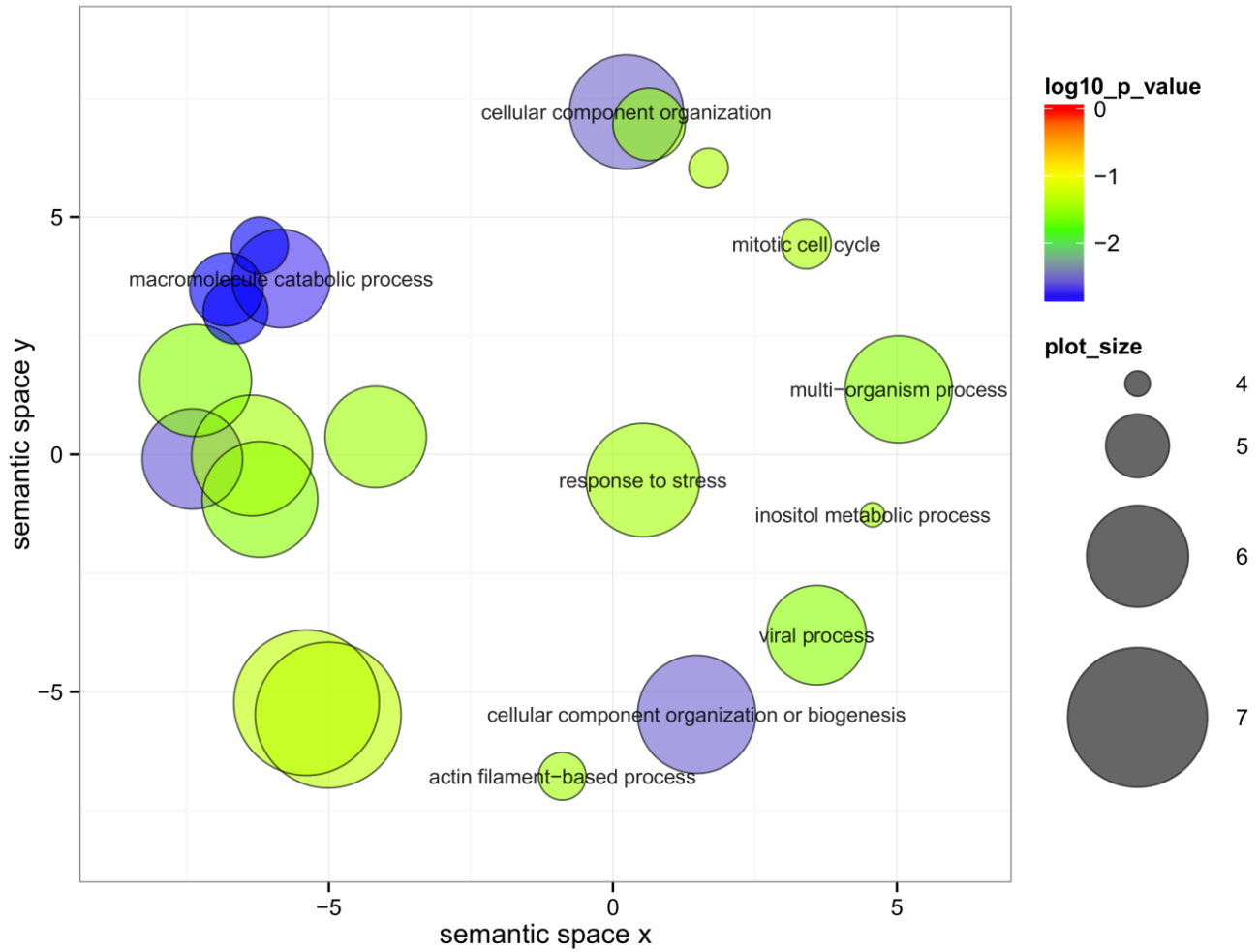


Figure 2.24: **21 cluster representatives of 35 enriched biological process ontologies upregulated in the three virulent *E. histolytica* strains (i.e. PVBM08B, HM-1:IMSS and IULA:1092:1).** Blue and light blue coloured clusters represent GO terms with more significant FDR-adjusted *P*-values than green coloured ones. A larger sized bubble reflects a more general term than a smaller one. Multidimensional scaling was calculated using the pairwise distance matrix. Their closeness on the plot would reflect the semantic similarity. Overall, these 21 upregulated clusters represent several biological processes including biosynthesis, cellular component organisation, cytoskeleton, protein catabolism, cell division and stress response, implying their roles in pathogenesis and virulence.

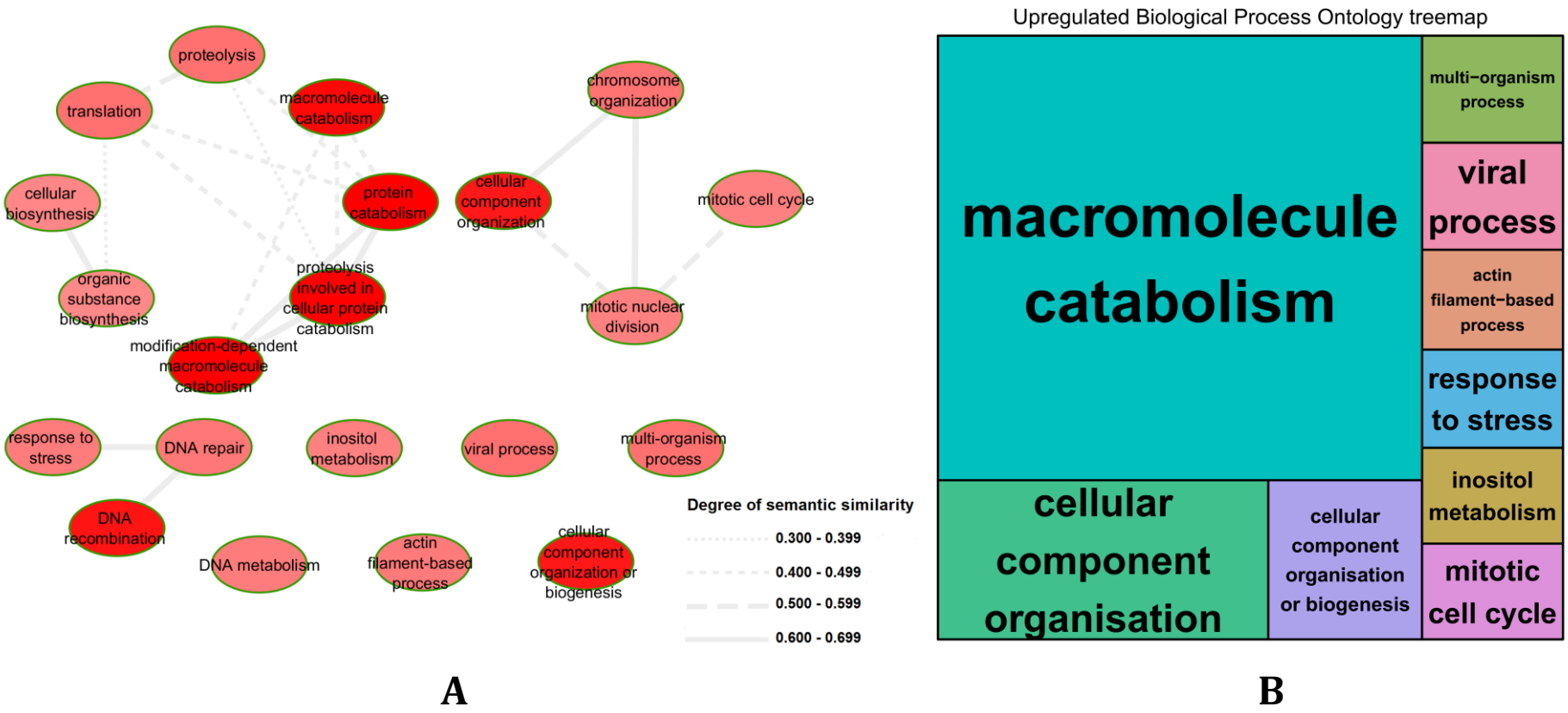


Figure 2.25: Interconnection of 21 representative process ontologies upregulated in the three virulent *E. histolytica* strains. Highly similar GO clusters are linked together, likely to form two interactomes of protein catabolism and cell division (A). It is reasonable to explain that increased cell cycle process is accompanied with DNA metabolism, repair and recombination. Also, increased translation and biosynthetic processes could be driven by nutrients and energy derived from increased protein catabolism, possibly due to increased phagocytosis. Upregulation of actin filament-based process reflects the rapid cytoskeletal dynamics served for the increase of cell motility, phagocytosis and cell division. Different line types represent degrees of semantic similarity. Reddish pink coloured bubbles have more significant FDR-adjusted *P*-values than pink coloured bubbles. As shown in the treemap, the majority of enriched GO clusters are joined into the supercluster of 'macromolecule catabolism', suggesting that catabolic process is favorable in virulent strains (B). Size of each rectangle is adjusted by both its FDR-adjusted *P*-value and the frequency of such GO term in the GOA database.

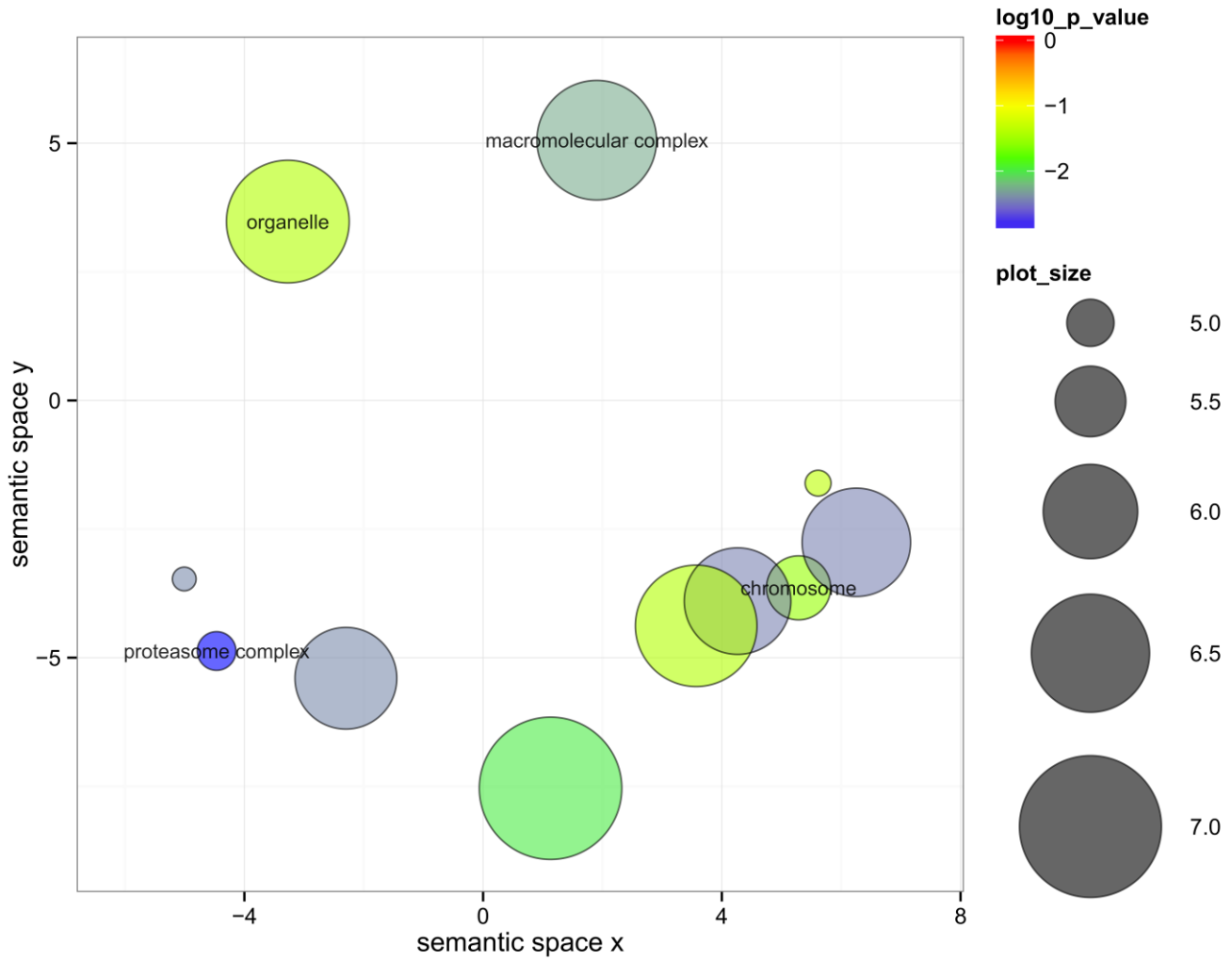
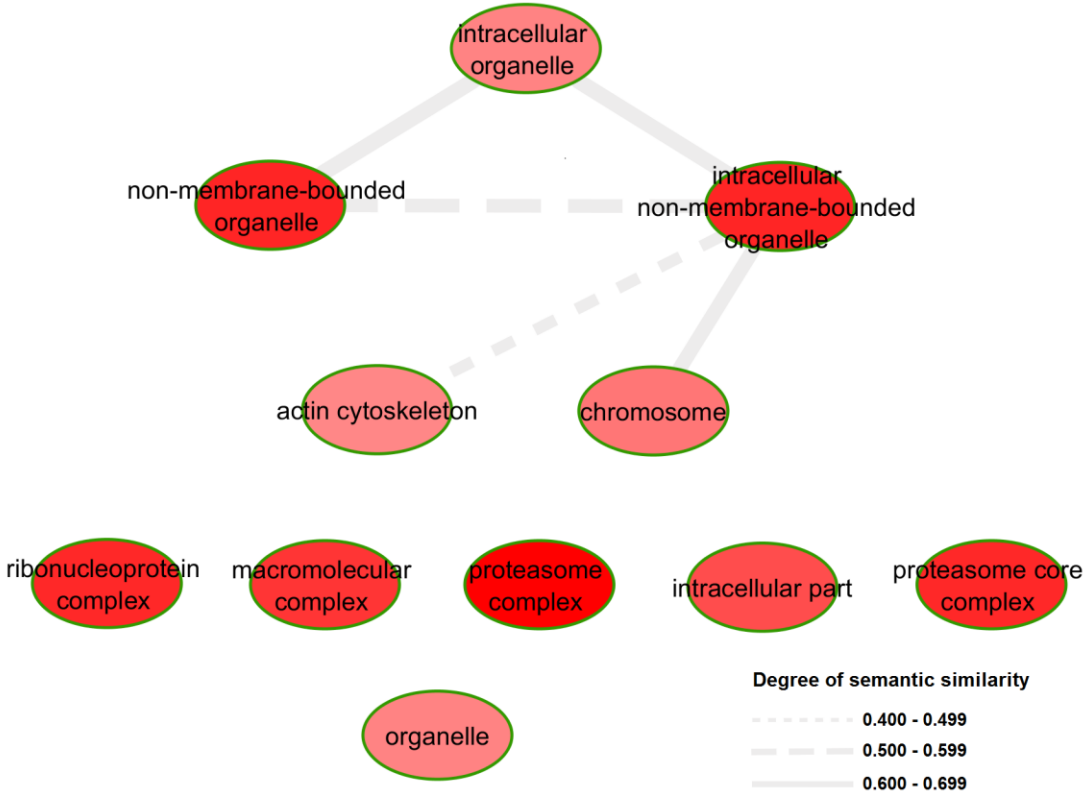
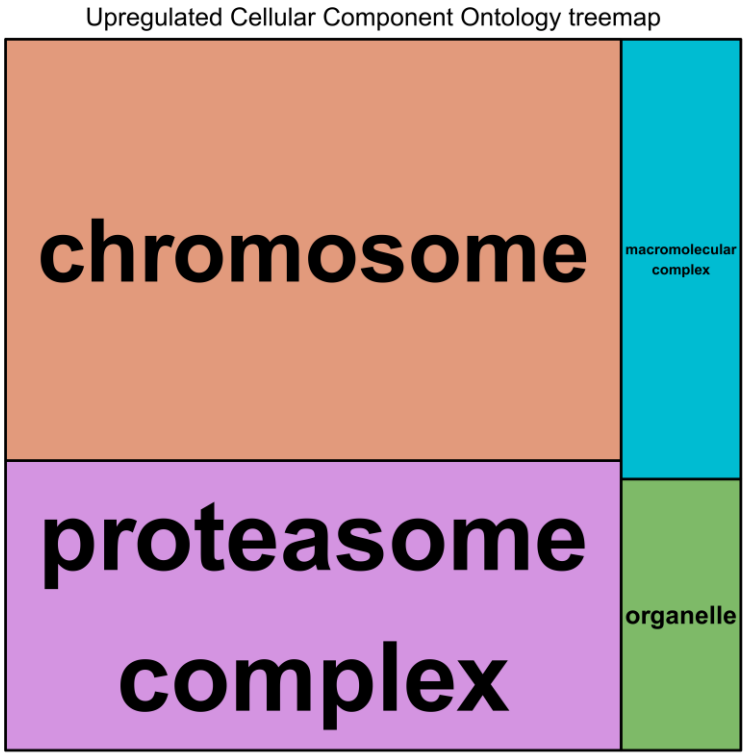


Figure 2.26: **11 cluster representatives of 15 enriched cellular component ontologies upregulated in the three virulent *E. histolytica* strains.** Blue and deep green coloured clusters represent GO terms with more significant FDR-adjusted *P*-values than light green ones. Consistently, cellular localisations of these clusters are mainly associated with biological processes described in the previous plots (Figures 2.24 and 2.25).



A



B

Figure 2.27: Interconnection of 11 representative component ontologies upregulated in the three virulent *E. histolytica* strains. Consistent with the previous interactive graph (Figure 2.25A), chromosome, ribonucleoprotein complex and proteasome complex are main cellular components responsible for mitotic cell division, DNA metabolism/repair/recombination and protein catabolism, respectively (A). Correspondingly, two superclusters of 'chromosome' and 'proteasome complex' in the above treemap indicate high protein catabolism and active cell division in the three virulent strains (B).

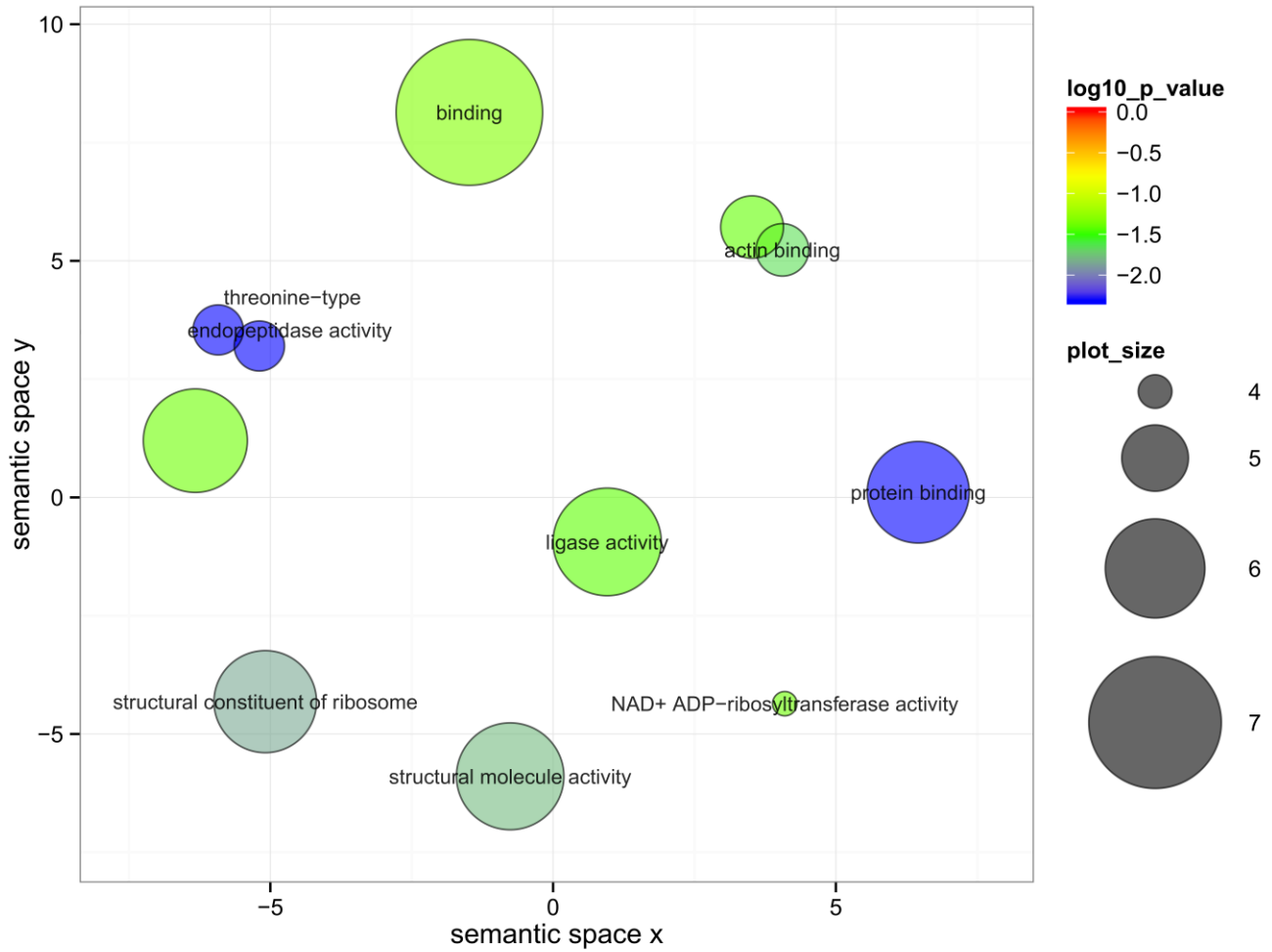


Figure 2.28: **11 cluster representatives of 12 enriched molecular function ontologies upregulated in the three virulent *E. histolytica* strains.** Clusters with blue and deep green colours represent GO terms having more significant FDR-adjusted *P*-value than green coloured ones. Upregulated clusters represent ontologies responsible for a variety of functions including structural component of ribosome, enzymatic activity, cytoskeleton binding as well as protein-protein interaction.

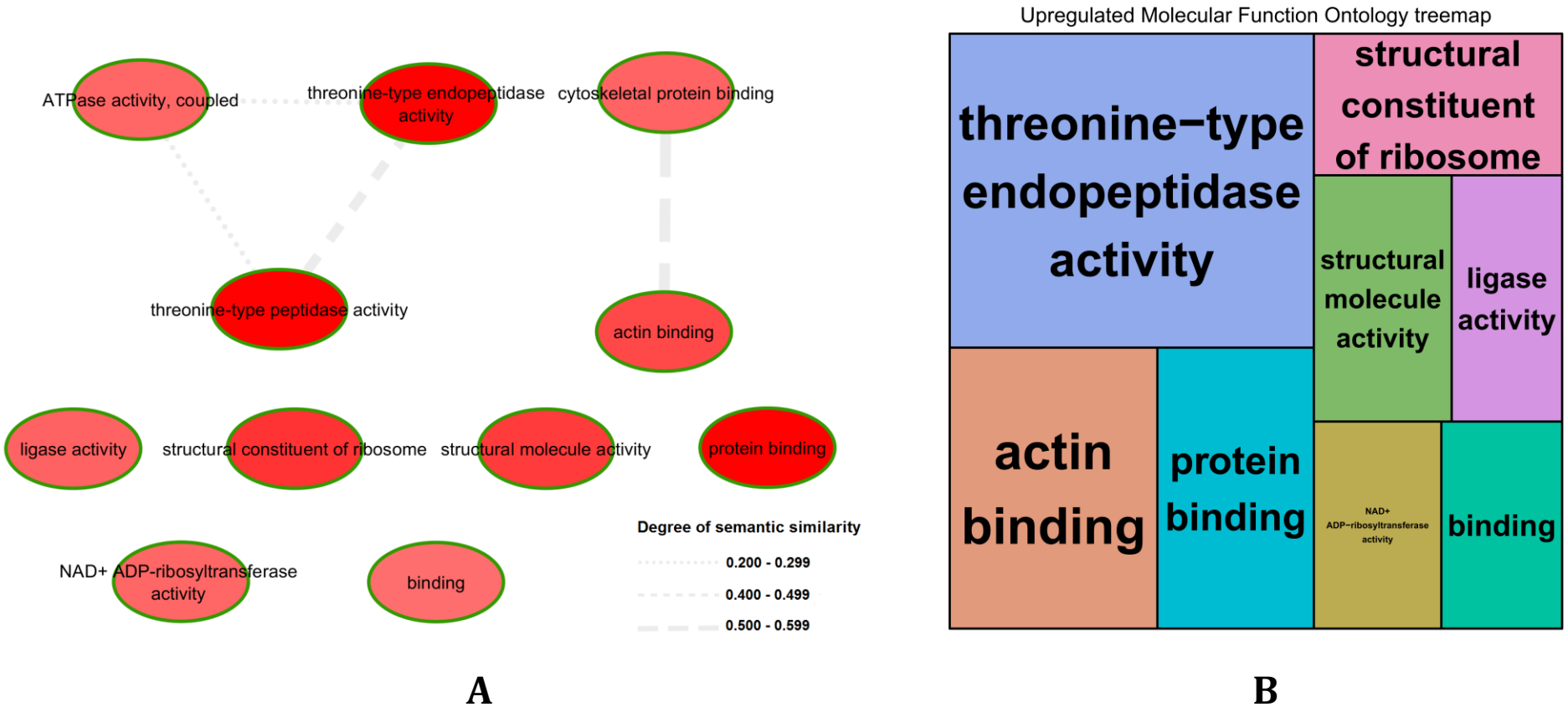


Figure 2.29: Interconnection of 11 representative function ontologies upregulated in the three virulent *E. histolytica* strains. Upregulation of endopeptidase activity, actin and cytoskeletal protein binding functions as well as NAD^+ ADP-ribosyltransferase activity, responsible for cell signaling, gene regulation and DNA repair, are likely to be the molecular basis implicated in parasite virulence (A). The two superclusters in the above treemap contain GO representative clusters responsible for endopeptidase and actin binding activities, inferring increased activities of protein catabolism and cytoskeletal dynamics in virulent parasites (B).

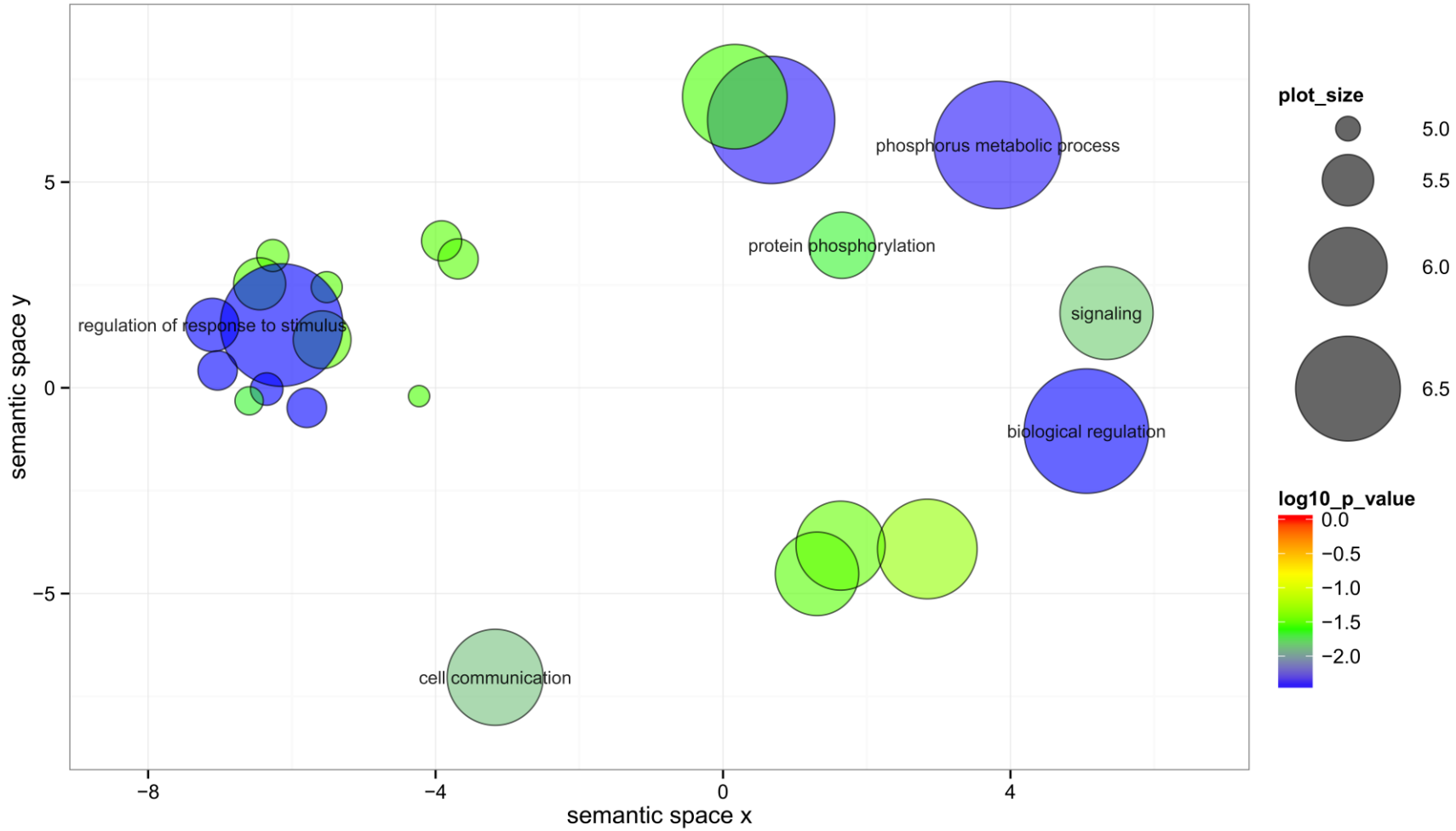
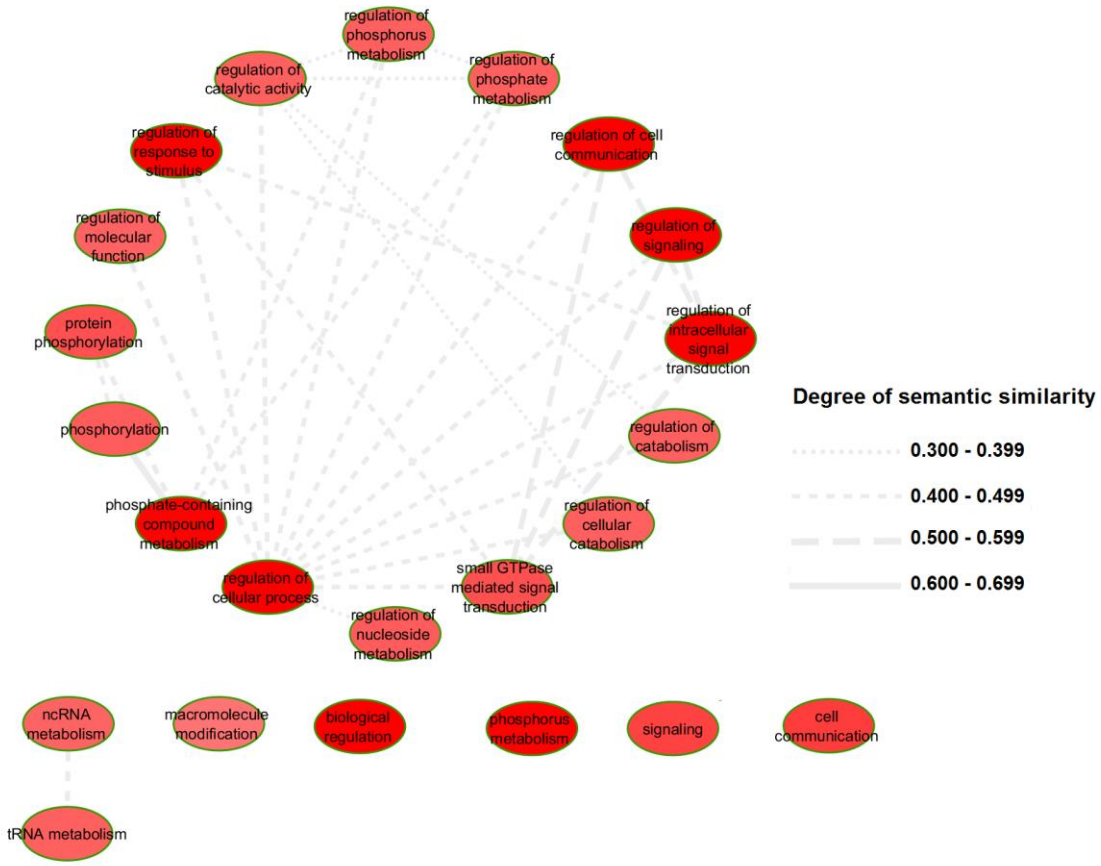
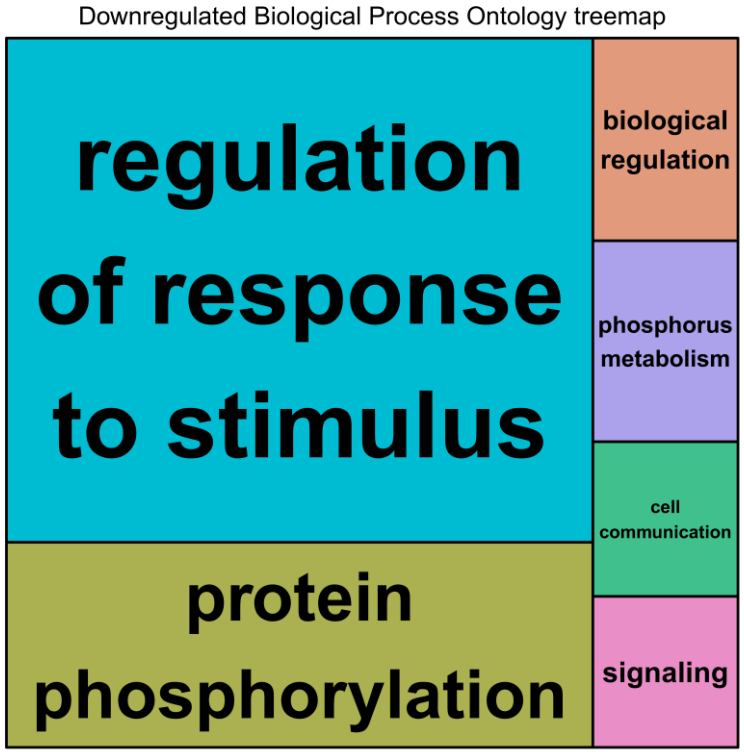


Figure 2.30: **23 cluster representatives of 44 enriched biological process ontologies downregulated in the three virulent *E. histolytica* strains.** Blue and deep green coloured clusters represent GO terms with more significant FDR-adjusted *P*-value than green coloured ones. Mostly, these 23 downregulated clusters represent several regulatory processes involved in signaling, cell communication, nucleoside metabolism, molecular function, cellular catabolism and phosphate metabolism, implying that reduction in the strict cellular control is likely to lead trophozoites to the pathogenic or virulent state.





A **B**

Figure 2.31: Interconnection of 23 representative process ontologies downregulated in the three virulent *E. histolytica* strains. Highly similar GO clusters involved in regulatory processes and and signaling are interconnected and tend to form the functional interaction network, suggesting less stringent regulations and corresponding incorrect response to the host environmental stimuli in virulent strains (A). Notably, two superclusters in the above treemap contain GO representative clusters responsible for regulatory processes and protein phosphorylation, implying that the marked downregulation of such regulation transcripts in virulent strains likely results in an aberrant behavior of virulent parasites (B).

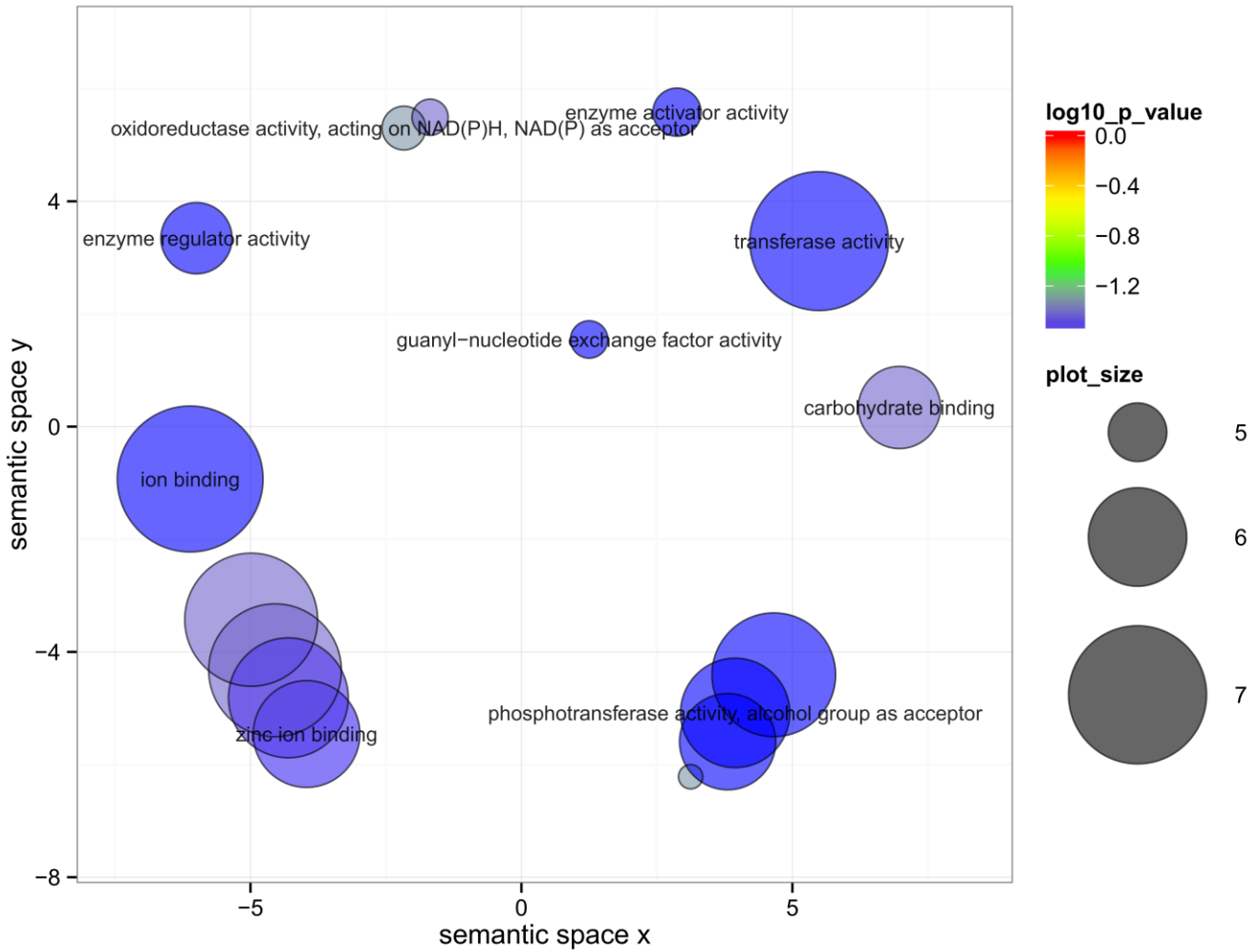
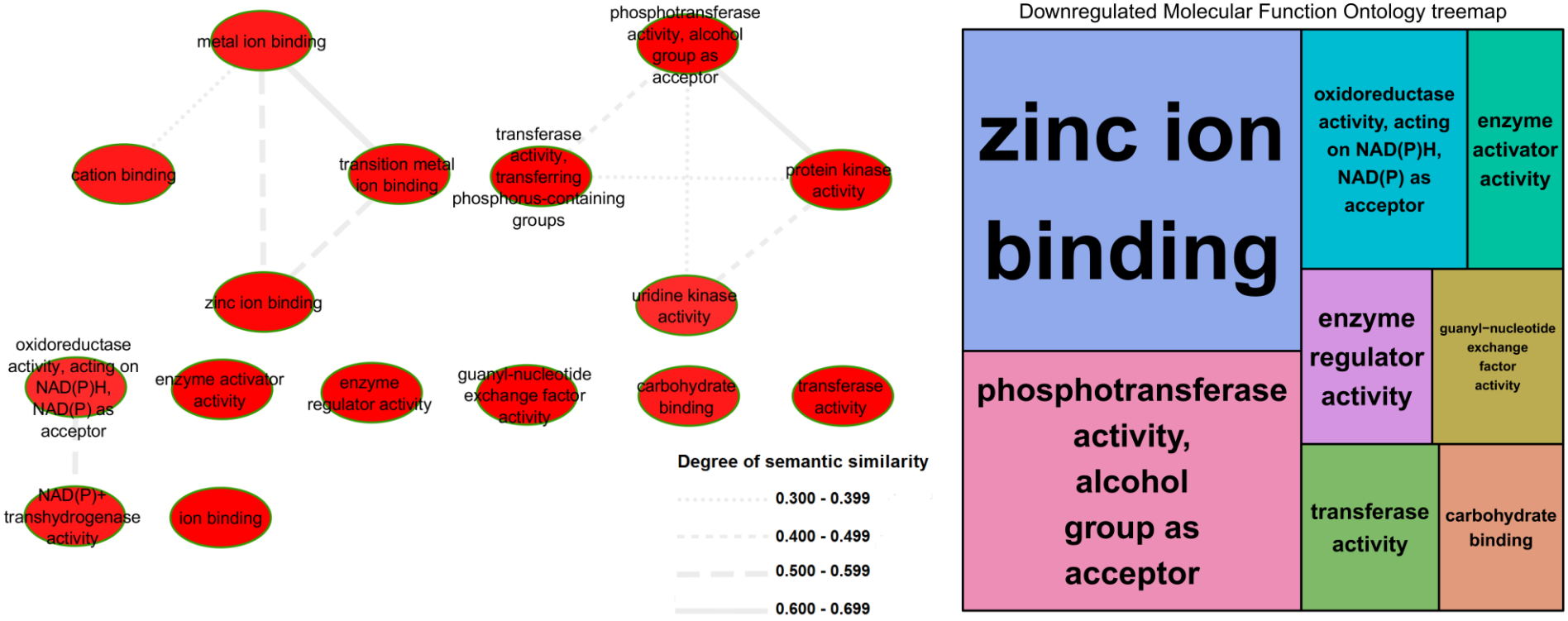


Figure 2.32: **16 cluster representatives of 24 enriched molecular function ontologies downregulated in the three virulent *E. histolytica* strains.** Blue coloured clusters represent GO terms with more significant FDR-adjusted *P*-values than light blue ones. Interestingly, the cluster of enzyme activator activity contains ontologies involved with small GTPases (Ras/Rab) activator activity, indicating reduced activation of Ras superfamily in the virulent parasites.



A **B**

Figure 2.33: Interconnection of 16 representative function ontologies downregulated in three virulent *E. histolytica* strains. Highly similar GO terms responsible for kinase and phosphotransferase activities are linked together, reflecting the reduction of protein phosphorylation and signaling in virulent strains (A). Correspondingly, the supercluster of 'phosphotransferase activity' in the above treemap indicates reduced protein phosphorylation in the three virulent strains (B).

2.4 Concluding remarks

In this study, genome-wide transcriptomic approaches using the Illumina HiSeq RNA-Seq can uncover significant differences in expression profiles between nonvirulent and virulent laboratory-adapted *E. histolytica* strains. Differential gene expression analysis between the nonvirulent Rahman strain and three virulent strains, i.e. PVBM08B, HM-1:IMSS and IULA:1092:1, reveals that transcripts of genes involved in host cell killing and mucosal invasion, nucleic acid interaction and response to oxidative stress are prominently upregulated in the virulent trophozoites.

The cluster of 98 DE genes with a high degree of transcriptional variability among the four strains was identified by the hierarchical clustering analysis based on their relative expression profiles, indicating that this gene cluster is likely to play a role in determining virulence in a strain-specific pattern. Moreover, the gene members in this cluster exhibit a high frequency of sequence polymorphisms, 9.31 SNPs/kb on average, and show the significant positive correlation with their transcriptional variability across the *E. histolytica* strains, reflecting the variable degrees of gene regulation among these polymorphic genes. As such, the identification of the exclusive set of rapidly evolved genes exhibiting transcriptional variation across the strains enables us to better understand the impact of genetic variation on the differential virulence in *E. histolytica* infection.

Protein domain signatures identified by InterProScan also indicate the upregulation of transcripts encoding proteolysis-related domains as well as the co-upregulation of actin cytoskeleton and actin-modulating domains in the virulent strains. Consistently, diverse process ontologies related to protein catabolism, cellular biosynthesis, DNA metabolism, repair and recombination, mitotic cell division, cytoskeletal dynamics as well as response to stress are highly overrepresented as a core metabolism in the virulent strains, indicating the rapid proliferation and active metabolic state are the main drivers of virulence.

Noticeably, the DGE and InterProScan analyses revealed that functionally annotated transcripts involved in protein phosphorylation and G-protein signaling were both upregulated and downregulated as well as constituted a large fraction of the modulated transcripts in the transcriptomes of the virulent strains, indicating the great effect of PK-dependent and G-protein signaling pathways in regulation of diverse biological processes in this parasite. However, the number of signaling-related transcripts is higher in downregulation than upregulation, suggesting the less strict cellular regulations compared to the nonvirulent Rahman strain. Likewise, the striking underrepresentation of ontologies involved in signaling and regulatory processes was observed in the virulent parasites.

Altogether, it could be explained that reduced regulation of sensing and correctly responding to the environmental stimuli potentially enables the parasites to become virulent and subsequently cause the invasive infection.

Invasive trophozoites cannot develop cysts to infect other hosts, resulting in a reduction of reproductive fitness [81,255]. It is therefore unsurprising that asymptomatic *E. histolytica* infection is much more prevalent, accounting for ~90% of worldwide reported cases [3,6]. Hence, it could be argued that the nonvirulent strains are better adapted to their host through improved environmental sensing and gene regulation. In conclusion, my comparative transcriptomic analysis identified a large number of modulated transcripts which potentially contribute to differential virulence among the four laboratory-adapted *E. histolytica* strains. Also, my transcriptomic characterisation can provide a fuller understanding in the molecular basis of physiological differences between nonvirulent and virulent strains as well as the evolutionary perspectives on the spectrum of disease severity in *E. histolytica* infection.

Chapter 3: Analysis of differential gene expression focusing on a representative set of putative virulence-associated genes using NanoString nCounter® technology

3.1 Introduction

In Chapter 2, I carried out the comparative transcriptomics across nonvirulent and virulent *E. histolytica* strains using Illumina RNA-Seq analysis. RNA-Seq can provide high-resolution transcriptomic landscapes of the *E. histolytica* parasite strains and catalog the gene clusters of upregulation and downregulation in relation to their virulence variability.

However, current high-throughput sequencers can provide accuracy with good Phred quality score for cDNA fragments with partial length of the original transcripts. Due to this limitation, cDNA library preparation for RNA-Seq normally requires a step of fragmentation, potentially resulting in a set of cDNA fragments with non-uniform distribution [108,258,259]. This fragment bias can affect to the accurate measurement of transcript abundances. Additionally, several steps during cDNA library preparation such as reverse transcription, PCR amplification as well as adapter ligation may introduce sequence-dependent biases and amplification noises, resulting in the decreased possibility to detect rare transcripts [259,260]. To overcome these obstacles, several methods have been developed such as NanoString technology that can abolish all involved enzymatic reactions mentioned above and apply the specific probes for hybridisation and direct digital detection instead [261]. Alternatively, direct RNA sequencing can be applied to minimise such biases by skipping the PCR step and directly sequencing the RNA molecule [260].

The NanoString nCounter® gene expression (GX) analysis is a novel, robust technology recently developed for simultaneous, multiplexed detection and quantitation of up to 800 transcripts in a single reaction without amplification [261]. Unlike other expression profiling approaches such as genome-wide microarray or quantitative PCR, it provides the digital measurement of target mRNA molecules by directly hybridising the target with specific colour-coded barcodes as illustrated in Figure 3.1. Each specific colour barcode contains a pair of capture and reporter probes with target specific sequences. During solution phase hybridisation, barcoding of mRNA molecules is achieved by annealing with the reporter probe carrying a unique and target-specific colour code at its 5' end, whereas the capture probe will allow the target-probe complex to be attached on a cartridge for downstream data acquisition. After the hybridisation step, the complex is purified from excess probes and then immobilised on the nCounter® cartridge, which is subsequently

placed in the nCounter® digital analyser for reading barcode-specific fluorescent signals and exporting the data in tabulated form. Essentially, the number of times, which the specific colour-coded barcode for a particular gene of interest is counted, refers to the number of target mRNA molecules.

In this present work, the NanoString technology has been applied to verify the validity of the RNA-Seq data previously obtained in the previous chapter by focusing on a representative set of fifty-three virulence-associated genes. As listed in Table 3.1, this representative set consists of thirty-three functional genes that show significant differential expression between nonvirulent and virulent strains in my RNA-Seq data and the other twenty functional genes that were not revealed for expression difference by my RNA-Seq analysis but reported for their differential expression in the previous publications and mostly characterised for their putative roles in *E. histolytica* virulence [20-22,262-266]. For 33 DE genes as mentioned above, 25 and 8 genes were found to be commonly upregulated and downregulated in the virulent strains (i.e. PVBM08B, HM-1:IMSS and IULA:1092:1) relative to the nonvirulent Rahman strain, respectively.

Therefore, expression profilings of these 53 representative genes across the four *E. histolytica* strains by the NanoString technology will provide us much more promising data without any bias due to fragmentation, PCR amplication or enzymatic reactions and enable us to compare the performance and validity with the previous RNA-Seq data. Additionally, it is hoped that the evaluation of transcriptional variability across this representative set of putative virulence-associated genes would potentially summarise and reflect their virulence variation better than the whole transcriptomic scale.

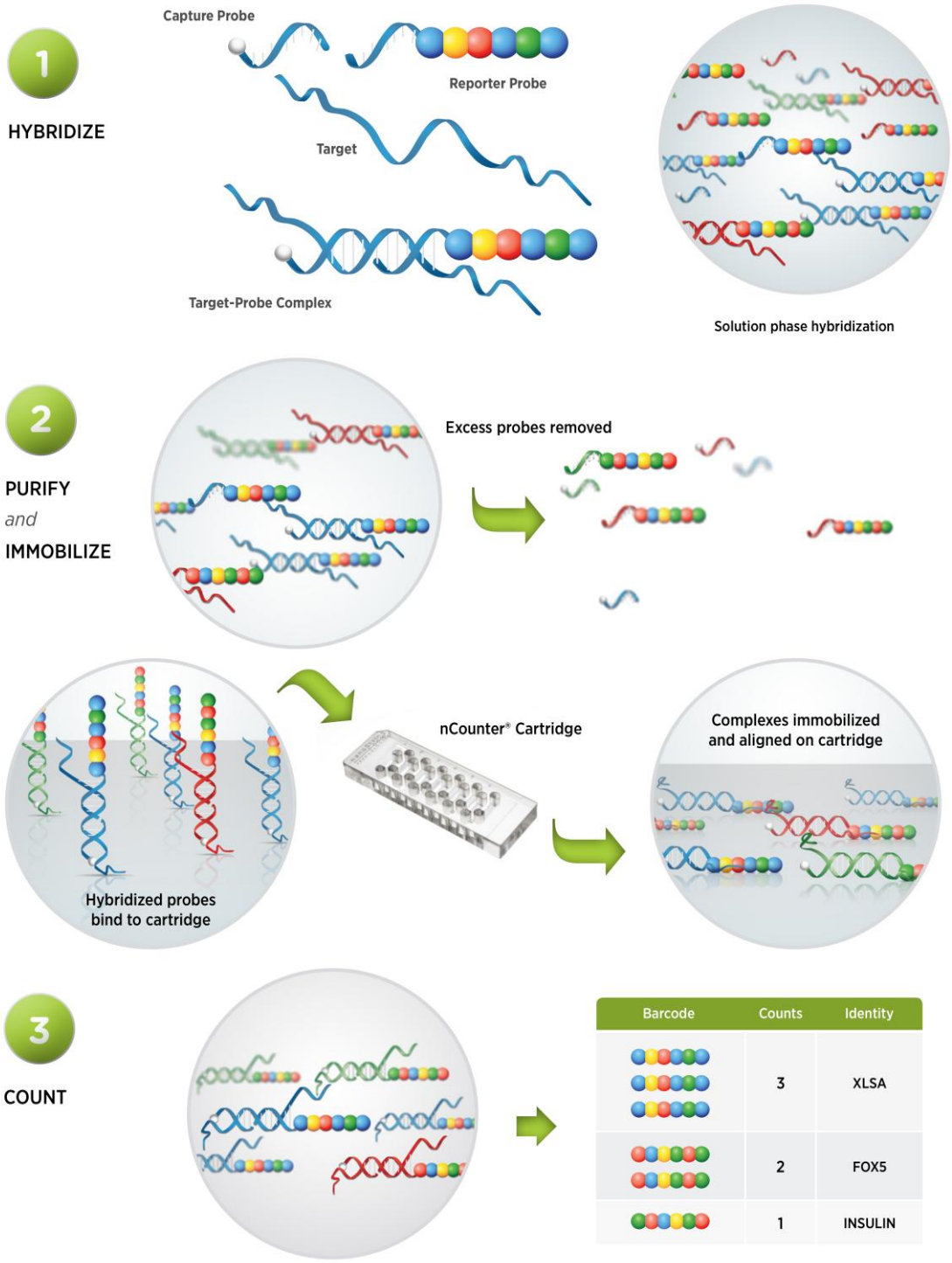


Figure 3.1: Principles and procedures of the NanoString nCounter® GX assay (available online at http://www.nanostring.com).

3.2 Materials and Methods

3.2.1 *E. histolytica* genes chosen for the NanoString nCounter® GX assay

Fifty-three functionally annotated genes and two house-keeping genes: chaperonin-1 60 kDa (cpn60: EHI_178570) and tubulin gamma chain (TUBG: EHI_008240) as listed in Table 3.1 were enrolled for nCounter probe design. Table 3.1 shows the list of these 53 genes with putative functions, consisting of 25 genes (No. 1-25) shown to be upregulated in three virulent strains: PVBM08B, HM-1:IMSS and IULA:1092:1, 8 genes (No. 26-33) downregulated in these three strains as well as an additional set of 20 genes (No. 34-53) previously described as virulence-associated genes.

3.2.2 Strains of *E. histolytica* and total RNA extraction

In this chapter, *E. histolytica* trophozoites were the same strains (i.e. Rahman, PVBM08B, HM-1:IMSS and IULA:1092:1) previously used in the RNA-Seq study (Chapter 2). The experiment was run in triplicate design to prevent any bias of measurements. Mid-log phase trophozoites were collected for 12 samples in total (3 replicate lines for four strains) for RNA preparation. Total RNA isolation was performed using the Trizol® plus RNA purification kit (Invitrogen). The isolated RNA samples were assessed quantitatively using the Qubit® fluorometric assay (Invitrogen) as well as qualitatively using the Agilent 2100 Bioanalyser (Agilent Technologies). The RNA samples were kept at -80 ^{o}C until used for the NanoString analysis.

3.2.3 NanoString nCounter® GX assay and data processing

To directly detect the mRNA expression levels of 53 chosen *E. histolytica* genes, NanoString nCounter® Gene Expression Analysis (NanoString Technologies, USA) was conducted for each sample using a custom designed codeset containing 55 genes including two housekeeping genes: cpn60 (EHI_178570) and TUBG (EHI_008240). Briefly, 100 ng of total RNA for each sample was used for soluble phase hybridisation by incubating overnight with a target-specific codeset of reporter and capture probes as well as 8 pairs of negative control and 6 pairs of positive control probes. Then, the tubes were placed onto the automated nCounter® Prep Station for steps of excess probe removal and immobilisation of target-probe complexes on the nCounter® cartridge. The sample cartridge was transferred to the ncounter® Digital Analyser for digital counting and data collection. Finally, processing of nCounter data was done using the nSolver™ 2.0 Analysis Software (NanoString Technologies).

Raw counts of two housekeeping genes: cpn60 and TUBG were used both for data normalisation. A mean value of the 8 negative control probes was applied for background subtraction in all reactions. A normalisation factor was computed using a geometric mean of the 6 positive control probes and the two housekeeping probes.

3.2.4 Evaluation of concordance between NanoString GX analysis and RNA-Seq results

To assess the performance and validate the normalised NanoString data obtained in this experiment, scatterplots and Pearson correlation analyses were performed using R Statistics software version 3.1.2 (http://CRAN.R-project.org) to determine whether there is concordance of transcript levels between NanoString GX analysis and previous RNA-Seq data [267]. Normalised read count and gene expression fold change of 53 representative genes retrieved from both NanoString assay and RNA-Seq results were plotted against each other as shown in Figures 3.2 and 3.3, respectively. Statistical significance of the Pearson correlation test are considered when *P*-value is less than 0.05.

3.2.5 Agglomerative hierarchical clustering and comparison of the transcriptional profiles between *E. histolytica* strains

Then, normalised nCounter data of the four strains were analysed for agglomerative hierarchical clustering as shown in Figures 3.4 and 3.5. Significant testing for differential expression between two contrasting strains was performed for each contrast using the Student's two tailed t-test and considered statistically significant if an FDR-adjusted *P*-value is less than 0.05.

To explore the transcriptional variation between strains, normalised transcript levels of 53 representative genes in Rahman, PVBM08B and IULA:1092:1 were individually plotted against those of HM-1:IMSS as scatter diagrams, as shown in Figures 3.6, 3.7 and 3.8, respectively. Gene identifiers were labelled for representative genes that were proven by the nSolver™ 2.0 software for their statistically significant upregulation or downregulation in relation to HM-1:IMSS. Also, these expression data of representative gene set were plotted in multidimensional scaling to explore the transcriptional similarity across the four strains as illustrated in Figure 3.9.

Table 3.1: Details of 55 *E. histolytica* genes enrolled for direct digital mRNA detection by the NanoString nCounter® GX assay.

Gene No. 1-25 and Gene No. 26-33 (grey highlight) have been previously shown in RNA-Seq data for their upregulation and downregulation in the virulent strains, respectively. In this study, twenty virulence-associated genes formerly reported are also included as listed in No. 34-53. Two housekeeping genes: chaperonin-1 60 kDa (EHI_178570) and tubulin gamma chain (EHI_008240) are used as references for normalisation. Targ Region = Target Region; Tm_CP = melting temperature of capture probe; Tm_RP = melting temperature of reporter probe.

No.	Accession	Identifier	Targ Region	Tm_CP	Tm_RP	Function	Note
1	EHI_199270	LRR-1	266-365	76	74	leucine-rich repeat protein, BspA family	**A**
2	EHI_180390	AIG1-1	837-936	80	79	AIG1 family protein, putative	
3	EHI_012330	STIRP-1	730-829	75	77	serine-threonine-isoleucine rich protein, putative	
4	EHI_015120	LRR-2	357-456	72	75	leucine-rich repeat protein, BspA family	
5	EHI_025700	STIRP-2	1387-1486	73	81	serine-threonine-isoleucine rich protein, putative	
6	EHI_059860	C2B	313-412	78	84	C2 domain-containing protein	
7	EHI_123850	ARIEL1-1	55-154	75	73	surface antigen ariel1, putative	
8	EHI_144590	PK-1	797-896	78	77	protein kinase domain-containing protein	
9	EHI_050970	CXXC	146-245	77	77	CXXC-rich protein	
10	EHI_182460	DL2	710-809	84	80	dextranase precursor, putative	
11	EHI_004340	STIRP-3	891-990	73	72	serine-threonine-isoleucine rich protein, putative	
12	EHI_191510	LRR-3	966-1065	72	72	leucine-rich repeat protein, BspA family	
13	EHI_185270	DOCK	434-533	80	79	dedicator of cytokinesis domain-containing protein	
14	EHI_022730	SRP54	380-479	79	79	signal recognition particle 54 kDa protein, putative	**B**
15	EHI_108750	Rap1GAP	11-110	74	74	Rap/Ran GTPase-activating protein, putative	
16	EHI_021570	SAT2	814-913	83	78	serine acetyltransferase 1	
17	EHI_148550	TMK52	1003-1102	81	79	protein tyrosine kinase domain-containing protein	
18	EHI_145840	PRDX	474-573	81	79	peroxiredoxin	**C**
19	EHI_188600	DSPP	185-284	80	84	dentin sialophosphoprotein precursor, putative	**D**
20	EHI_049620	NifU	67-166	81	83	Fe-S cluster assembly protein NifU, putative	
21	EHI_138480	ISF	273-372	77	80	iron-sulfur flavoprotein, putative	
22	EHI_105080	Zif	491-590	76	76	zinc finger protein, putative	
23	EHI_176970	Cdc48-like	2084-2183	79	83	cdc48-like protein, putative	
24	EHI_095060	LRR-4	285-384	76	81	leucine-rich repeat protein, BspA family	
25	EHI_117680	TMK10	2979-3078	72	73	tyrosine kinase, putative	

Table 3.1: Details of 55 *E. histolytica* genes enrolled for direct digital mRNA detection by the NanoString nCounter® GX assay. (Continued)

No.	Accession	Identifier	Targ Region	Tm_CP	Tm_RP	Function	Note
26	EHI_176590	AIG1-2	776-875	81	79	AIG1 family protein, putative	
27	EHI_023880	UBE2	52-151	78	77	ubiquitin-conjugating enzyme family protein	
28	EHI_023890	NUDC	216-315	78	81	nuclear movement protein, putative	
29	EHI_023840	RPL38	1-100	76	72	60S ribosomal protein L38, putative	
30	EHI_023870	WD40	575-674	80	81	WD domain-containing protein	
31	EHI_063550	MYB-like	21-120	76	75	myb-like DNA-binding domain-containing protein	**E**
32	EHI_172850	ARIEL1-2	74-173	71	76	surface antigen ariel1, putative	
33	EHI_115720	MBL	510-609	79	79	metallo-beta-lactamase superfamily protein	
34	EHI_025850	MDN	55-154	74	81	midasin	
35	EHI_096770	AT	453-552	79	78	acetyltransferase, putative	
36	EHI_179340	HMG	508-607	79	83	HMG box protein	
37	EHI_079300	LCFA-CoA-L	65-164	80	77	long-chain-fatty-acid--CoA ligase, putative	
38	EHI_060340	CS-3	567-666	81	84	cysteine synthase A, putative	
39	EHI_000900	APPBP1	1096-1195	77	80	ThiF family protein	
40	EHI_164520	ISF-Ps	147-246	77	79	iron-sulfur flavoprotein, putative, pseudogene	**F**
41	EHI_082590	GARP	6-105	83	84	glutamic acid-rich protein precursor, putative	
42	EHI_168240	CP-A5	628-727	82	80	cysteine proteinase, putative	
43	EHI_050570	CP-A4	511-610	77	78	cysteine proteinase, putative	
44	EHI_012270	Hgl2	835-934	79	75	Gal/GalNAc lectin heavy subunit	
45	EHI_077500	Hgl3	1760-1859	78	75	galactose-specific adhesin 170kD subunit	**G**
46	EHI_197460	ROM1	29-128	81	84	peptidase S54 (rhomboid) family protein	
47	EHI_098210	KERP1	192-291	80	79	lysine and glutamic acid-rich protein 1 (KERP1)	
48	EHI_159480	AP-A	96-195	81	77	pore-forming peptide ameobapore A precursor, putative	
49	EHI_026420	Rab5	159-258	81	73	Rab family GTPase	
50	EHI_048410	PK-2	1315-1414	80	82	serine/threonine protein kinase, putative	
51	EHI_019390	CP-Ps	38-137	73	74	cysteine proteinase, pseudogene	**H**
52	EHI_048850	LRR-5	136-235	76	75	leucine-rich repeat-containing protein	
53	EHI_049690	Lgl2	702-801	77	77	galactose-specific adhesin light subunit, putative	
54	**EHI_178570**	**cpn60**	**1139-1238**	**78**	**81**	**chaperonin-1 60 kDa**	**HK**
55	**EHI_008240**	**TUBG**	**370-469**	**83**	**78**	**tubulin gamma chain**	**HK**

Note comments for Table 3.1

A = also targets multiple leucine-rich repeat protein BspA family genes (EHI_041470; EHI_102380; EHI_034610; EHI_134140) > **92%**

B = also targets EHI_004750 putative signal recognition particle protein SRP54 at **100%**

C = also targets mulitple peroxiredoxins & peroxiredoxin pseudogenes (EHI_139570; EHI_172720; EHI_061980; EHI_123390; EHI_121620; EHI_201250; EHI_122310; EHI_114010; EHI_001420) > **92%**

D = also targets EHI_005260 putative surface antigen ariel1 at **98%**

E = also targets EHI_012420 myb-like DNA-binding domain-containing protein at **100%**

F = also targets EHI_189480 putative iron-sulfur flavoprotein at **98%**

G = also targets EHI_042370 putative galactose-specific adhesin 170 kD subunit at **100%**

H = also targets EHI_127470 (cysteine proteinase pseudogene) & EHI_046700 (hypothetical protein pseudogene) > **92%**

HK = Housekeeping gene

3.3 Results and Discussion

3.3.1 Normalised NanoString data show concordance with the previous RNA-Seq study

As shown in Table 3.2, normalised mRNA count for each gene of interest was listed in comparison among the four different strains of *E. histolytica*. Seven negative probes except NEG_F show a very low number of detected transcripts, indicating a very low background signal in this analysis. To validate the performance of the NanoString analysis in comparison with the previous RNA-Seq results, scatterplot analyses of all expression data and the Pearson's correlation tests were conducted using R Statistics software for all individual strains as well as six pairwise comparisons across all four strains enrolled in this study as illustrated in Figures 3.2 and 3.3, respectively.

A significant positive correlation was found between NanoString dataset, represented by $\log_2$(normalised NanoString count) and previous RNA-Seq, represented by $\log_2$(FPKM), with Pearson's correlation coefficients (r) = 0.7759-0.8874 and P-value less than 0.05 in all four strains, as shown in Figure 3.2 (A-D). These positive Pearson's correlation coefficients indicate a linear response of normalised NanoString counts (y-axis) to increasing FPKM (x-axis). Fundamentally, FPKM values were calculated by Cufflinks software (data not shown) to be a comparable parameter and reflect directly to the mRNA transcript level of an interested gene. In addition, all 53 functional genes selected for this NanoString experiment exhibit different FPKM values that are representative of varying the expression levels. Therefore, linearity between $\log_2$-transformed values of normalised NanoString counts and RNA-Seq FPKM demonstrates that the number of mRNA molecules counted by NanoString is promisingly proportional to the expression level of gene encoding such mRNAs.

In addition to a high degree of consistency between NanoString count and FPKM obtained from RNA-Seq, the performance of NanoString for differential expression analysis was also evaluated. As shown in Figure 3.3 (A-F), high correspondence between two sets of gene expression fold change values retrieved from the nSolver™ 2.0 and edgeR analyses can be observed in all six contrast pairs with correlation coefficients ranging from 0.7879 to 0.9179, P-value < 0.05. This siginificantly high concordance between these two transcriptomic platforms indicates that the NanoString analysis has precision in digital detection of mRNA transcripts as well as can provide reliability in differential gene expression analysis for studying comparative transcriptomics across a large range of expression and sample types.

Table 3.2: Normalised nCounter data from total RNA of the four *E. histolytica* strains. Control probesets includes 6 positive control probe pairs (POS_A - POS_F) and 8 negative control probe pairs (NEG_A – NEG_H). All 53 *E. histolytica* genes as well as two references (cpn60 and TUBG) are designated as 'Endogenous' and 'Housekeeping', respectively.

Grouped Data Name	Identifier	Accession	Rahman	PVBM08B	HM-1:IMSS	IULA:1092:1
Positive	POS_A	ERCC_00117.1	9162	10980	14163	9998
Positive	POS_B	ERCC_00112.1	2678	3253	4203	2963
Positive	POS_C	ERCC_00002.1	655	813	1058	762
Positive	POS_D	ERCC_00092.1	156	213	268	164
Positive	POS_E	ERCC_00035.1	31	32	51	33
Positive	POS_F	ERCC_00034.1	3	1	5	2
Negative	NEG_A	ERCC_00096.1	1	1	1	1
Negative	NEG_B	ERCC_00041.1	1	1	1	1
Negative	NEG_C	ERCC_00019.1	1	1	1	1
Negative	NEG_D	ERCC_00076.1	1	1	1	2
Negative	NEG_E	ERCC_00098.1	1	1	1	1
Negative	NEG_F	ERCC_00126.1	122	154	124	95
Negative	NEG_G	ERCC_00144.1	1	1	1	1
Negative	NEG_H	ERCC_00154.1	1	1	1	1
Housekeeping	TUBG	EHI_008240	384	697	645	640
Housekeeping	cpn60	EHI_178570	2857	1574	1702	1715
Endogenous	AIG1-1	EHI_180390	1	119	5382	36
Endogenous	AIG1-2	EHI_176590	2189	1	1	1
Endogenous	AP-A	EHI_159480	311247	484912	534115	386242
Endogenous	APPBP1	EHI_000900	356	1399	2454	1973
Endogenous	ARIEL1-1	EHI_123850	1	1	142	274
Endogenous	ARIEL1-2	EHI_172850	211	93	202	3
Endogenous	AT	EHI_096770	64	392	714	645
Endogenous	C2B	EHI_059860	6	14847	11559	17
Endogenous	CP-A4	EHI_050570	10454	13771	25888	4518
Endogenous	CP-A5	EHI_168240	71100	118415	120135	26845
Endogenous	CP-Ps	EHI_019390	548	1	45	248
Endogenous	CS-3	EHI_060340	82	237	504	532
Endogenous	CXXC	EHI_050970	2	1592	1249	820
Endogenous	DL2	EHI_182460	33	112	730	3904
Endogenous	DOCK	EHI_185270	6	872	363	641
Endogenous	DSPP	EHI_188600	28	89	137	75
Endogenous	GARP	EHI_082590	1783	3365	2315	2662
Endogenous	HMG	EHI_179340	944	1783	2545	1664
Endogenous	Hgl2	EHI_012270	35607	25118	14162	13545
Endogenous	Hgl3	EHI_077500	17588	1294	34775	6961
Endogenous	ISF	EHI_138480	222	1214	1847	2494
Endogenous	ISF-Ps	EHI_164520	344	405	523	1065

Table 3.2: Normalised nCounter data from total RNA of the four *E. histolytica* strains. (Continued)

Grouped Data Name	Identifier	Accession	Rahman	PVBM08B	HM-1:IMSS	IULA:1092:1
Endogenous	KERP1	EHI_098210	3857	3455	3036	4365
Endogenous	LCFA-CoA-L	EHI_079300	2350	8967	20246	12148
Endogenous	LRR-1	EHI_199270	1	766	1162	1
Endogenous	LRR-2	EHI_015120	1	25964	26299	13495
Endogenous	LRR-3	EHI_191510	1	107	165	105
Endogenous	LRR-4	EHI_095060	37	149	206	125
Endogenous	LRR-5	EHI_048850	641	207	99	98
Endogenous	MBL	EHI_115720	600	77	121	56
Endogenous	MDN	EHI_025850	13120	24749	22188	18241
Endogenous	MYB-like	EHI_063550	18253	7369	4223	2984
Endogenous	NUDC	EHI_023890	11490	1569	446	1041
Endogenous	NifU	EHI_049620	4462	20753	28066	13586
Endogenous	PK-1	EHI_144590	19	645	757	156
Endogenous	PK-2	EHI_048410	16585	8424	2237	1514
Endogenous	PRDX	EHI_145840	241324	481232	253053	686689
Endogenous	ROM1	EHI_197460	4260	4150	4723	3441
Endogenous	RPL38	EHI_023840	148	1	1	4
Endogenous	Rab5	EHI_026420	1494	2131	1703	733
Endogenous	Rap1GAP	EHI_108750	13	581	408	370
Endogenous	SAT2	EHI_021570	2	116	238	303
Endogenous	SRP54	EHI_022730	3576	8345	4930	5260
Endogenous	STIRP-1	EHI_012330	24	4237	2810	12
Endogenous	STIRP-2	EHI_025700	1	7811	3791	5
Endogenous	STIRP-3	EHI_004340	4	337	357	11
Endogenous	TMK10	EHI_117680	1	71	107	47
Endogenous	TMK52	EHI_148550	5	158	156	255
Endogenous	UBE2	EHI_023880	25624	3621	1565	2995
Endogenous	WD40	EHI_023870	10453	1421	1239	1282
Endogenous	Zif	EHI_105080	538	6401	5847	6664
Endogenous	Cdc48-like	EHI_176970	2300	19298	18544	20925
Endogenous	Lgl2	EHI_049690	25962	9338	8980	382

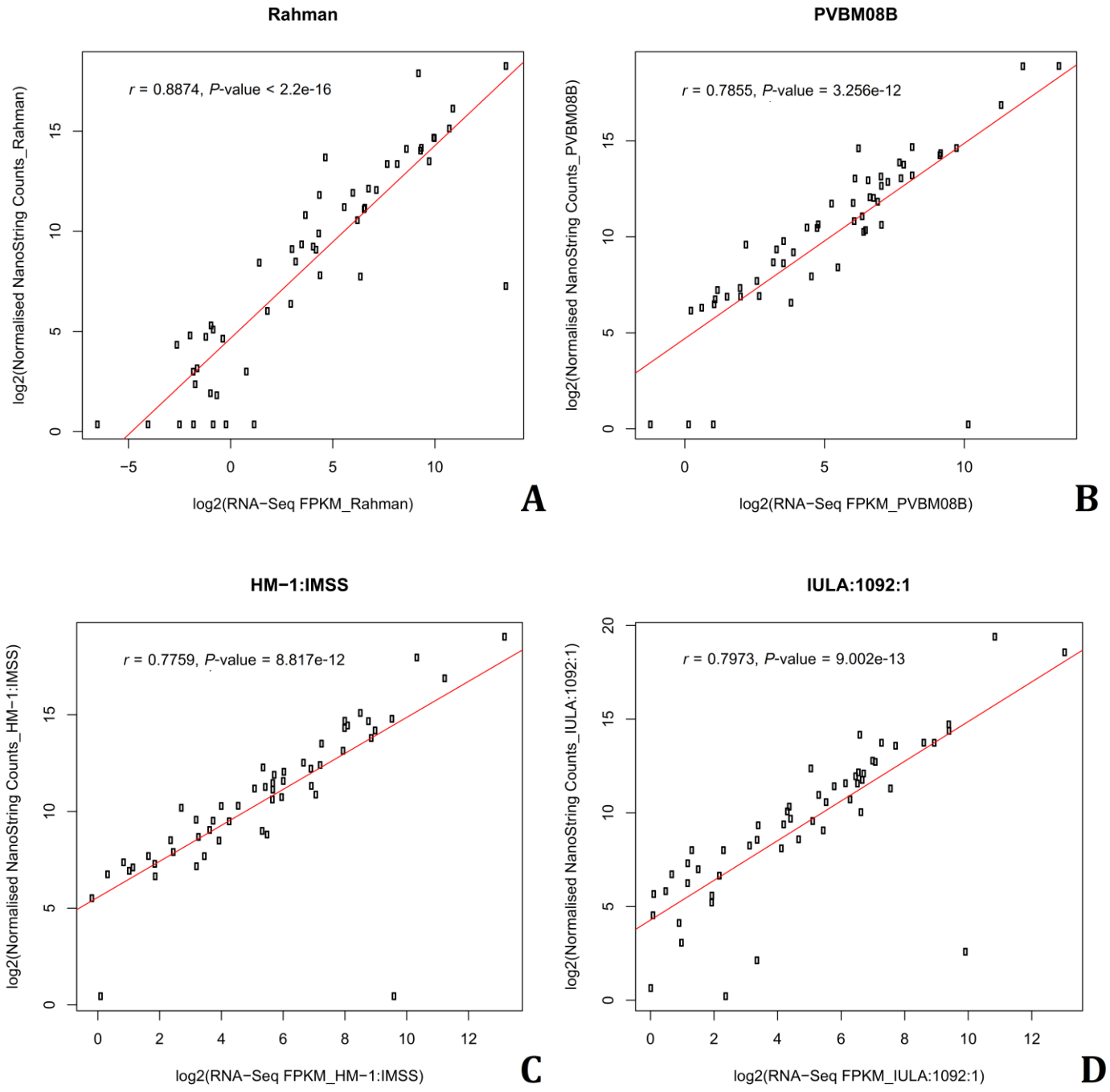


Figure 3.2: **Correspondence of gene expression levels as measured by RNA-Seq and NanoString analysis.** A total of 53 genes in each strain are plotted to reveal a significant positive correlation between two expression data sets that are obtained from RNA-Seq data (log_2FPKM, x-axis) and NanoString analysis (log_2(Normalised NanoString Counts), y-axis). A: Rahman, B: PVBM08B, C: HM-1:IMSS and D: IULA:1092:1, respectively.

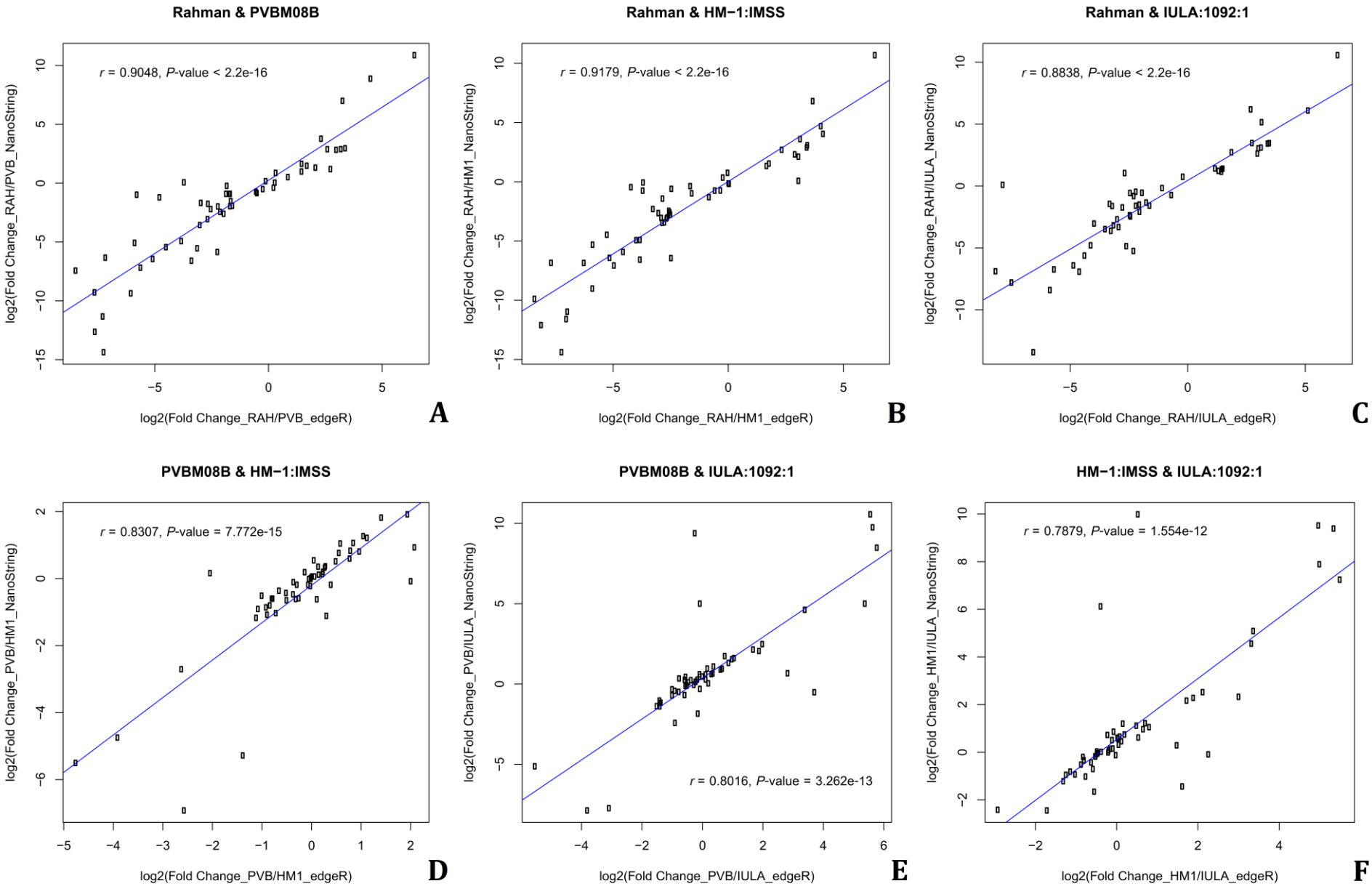

Rahman & PVBM08B
r = 0.9048, P-value < 2.2e-16
log2(Fold Change_RAH/PVB_NanoString)
log2(Fold Change_RAH/PVB_edgeR)
A
Rahman & HM−1:IMSS
r = 0.9179, P-value < 2.2e-16
log2(Fold Change_RAH/HM1_NanoString)
log2(Fold Change_RAH/HM1_edgeR)
B
Rahman & IULA:1092:1
r = 0.8838, P-value < 2.2e-16
log2(Fold Change_RAH/IULA_NanoString)
log2(Fold Change_RAH/IULA_edgeR)
C
PVBM08B & HM−1:IMSS
r = 0.8307, P-value = 7.772e-15
log2(Fold Change_PVB/HM1_NanoString)
log2(Fold Change_PVB/HM1_edgeR)
D
PVBM08B & IULA:1092:1
r = 0.8016, P-value = 3.262e-13
log2(Fold Change_PVB/IULA_NanoString)
log2(Fold Change_PVB/IULA_edgeR)
E
HM−1:IMSS & IULA:1092:1
r = 0.7879, P-value = 1.554e-12
log2(Fold Change_HM1/IULA_NanoString)
log2(Fold Change_HM1/IULA_edgeR)
F

Figure 3.3: **High correlation of fold change transcriptional differences between RNA-Seq and NanoString analysis.**

Scatterplots illustrate a significant positive correlation of gene expression differences (log_2FC) observed in a representative set of 53 genes in each strain contrast, as measured by RNA-Seq (x-axis) and NanoString method (y-axis).

A: Rahman vs PVBM08B; B: Rahman vs HM-1:IMSS; C: Rahman vs IULA:1092:1;

D: PVBM08B vs HM-1:IMSS; E: PVBM08B vs IULA:1092:1; F: HM-1:IMSS vs IULA:1092:1

3.3.2 Agglomerative hierarchical clustering of the nCounter data reveals co-expression of multigene family members

To comprehensively understand transcriptomic profiling across the four strains, normalised nCounter data of all 53 representative genes were clustered by agglomerative hierarchical clustering. In principle, agglomerative clustering will show an extensive hierarchy of clusters that includes a similar pattern of datasets and summarises the relationships between datasets in the form of a heatmap with a dendrogram tree as illustrated in Figure 3.4.

All fifty-three genes could be categorised into two clusters A and B. Thirty-five genes which showed upregulation in the three virulent strains, i.e. PVBM08B, HM-1:IMSS and IULA:1092:1, relative to Rahman were grouped together into cluster A whilst other 18 genes which showed downregulation relative to Rahman were categorised into cluster B. Consistently, 25 upregulated and 8 downregulated genes (see Table 3.1: No. 1-25 and No. 26-33, respectively) previously identified in the RNA-Seq analysis were sorted into cluster A and B, respectively.

Intriguingly, gene family members encoding proteins of related or similar function appear to be similar in their expression profiles across the parasite strains. For instance, genes encoding *E. histolytica* serine-threonine-isoleucine rich proteins (STIRP-1, STIRP-2 and STIRP-3), BspA-like LRRP (LRR-1) and C2 domain-containing protein (C2B) were grouped together in the same subcluster as demonstrated in Figure 3.5. As previously described in Chapter 2, EhSTIRPs were encoded by members of a multigene family and have cytotoxic and adhesive properties associated with virulence. Based on the number of mRNA molecules counted as listed in Table 3.2, there was very low or absent expression of these three gene family members in nonvirulent Rahman and virulent IULA:1092:1 strains. Contrastedly, all these three *EhSTIRP* gene family members were highly expressed in PVBM08B and HM-1:IMSS. This indicates that EhSTIRP expression is confined to the strains with high virulence potential.

In addition to EhSTIRPs, this unusual expression profile could be observed in a set of genes encoding BspA-like LRRPs such as designated LRR-1 (EHI_199270), LRR-2 (EHI_015120) and LRR-3 (EHI_191510) in this NanoString study. As listed in Table 3.2, these three LRRs were absent (mRNA count =1) in expression in Rahman whereas LRR-5 (EHI_048850) was conversely higher expressed (mRNA count = 641) than other three strains. As discussed above, it is possible to explain that there should be certain regulatory

mechanisms for silencing expression of virulence-associated genes in a strain-specific manner, especially in the Rahman strain.

Also, I applied small RNA sequencing to explore expression profiles of sRNAs in *E. histolytica* strains and determine whether such sRNAs contribute to the differential gene expression across the strains. Noticeably, the majority of expressed sRNAs are associated with reduced or lack of expression of virulence-associated genes in Rahman, including all three EhSTIRP members, LRRP members as well as C2B as mentioned above. Therefore, sRNA-mediated regulation potentially plays a crucial role in shaping parasite virulence. This will be explained later in details of Chapter 5.

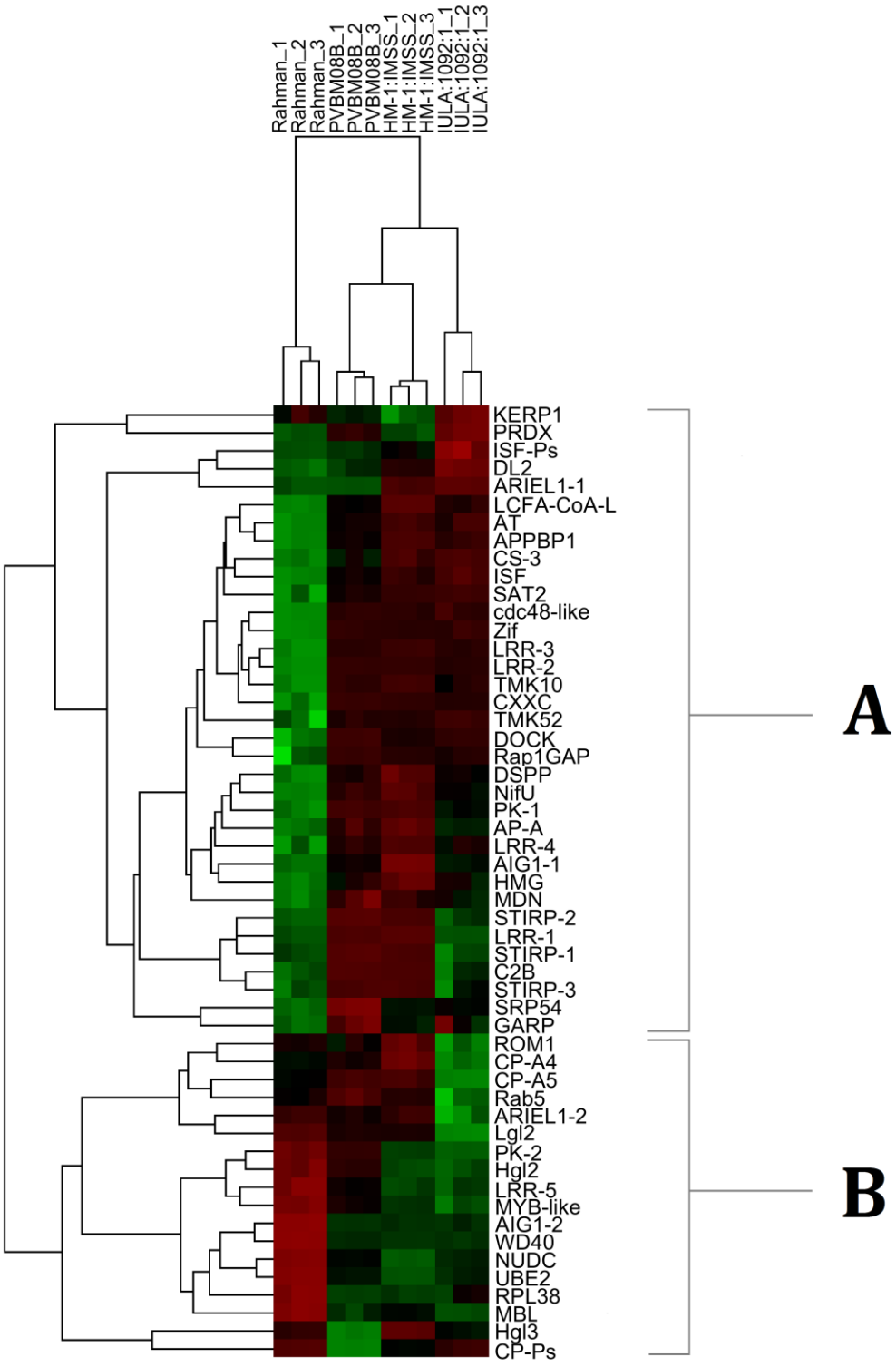


Figure 3.4: Agglomerative hierarchical clustering of 53 chosen representative genes with differential expression across the four *E. histolytica* strains. These 53 genes were categorised into two main clusters: 35 genes for cluster A and 18 genes for cluster B, based on their expression profiles across all four strains. Red colour and green colour spectra represent upregulation and downregulation, respectively.

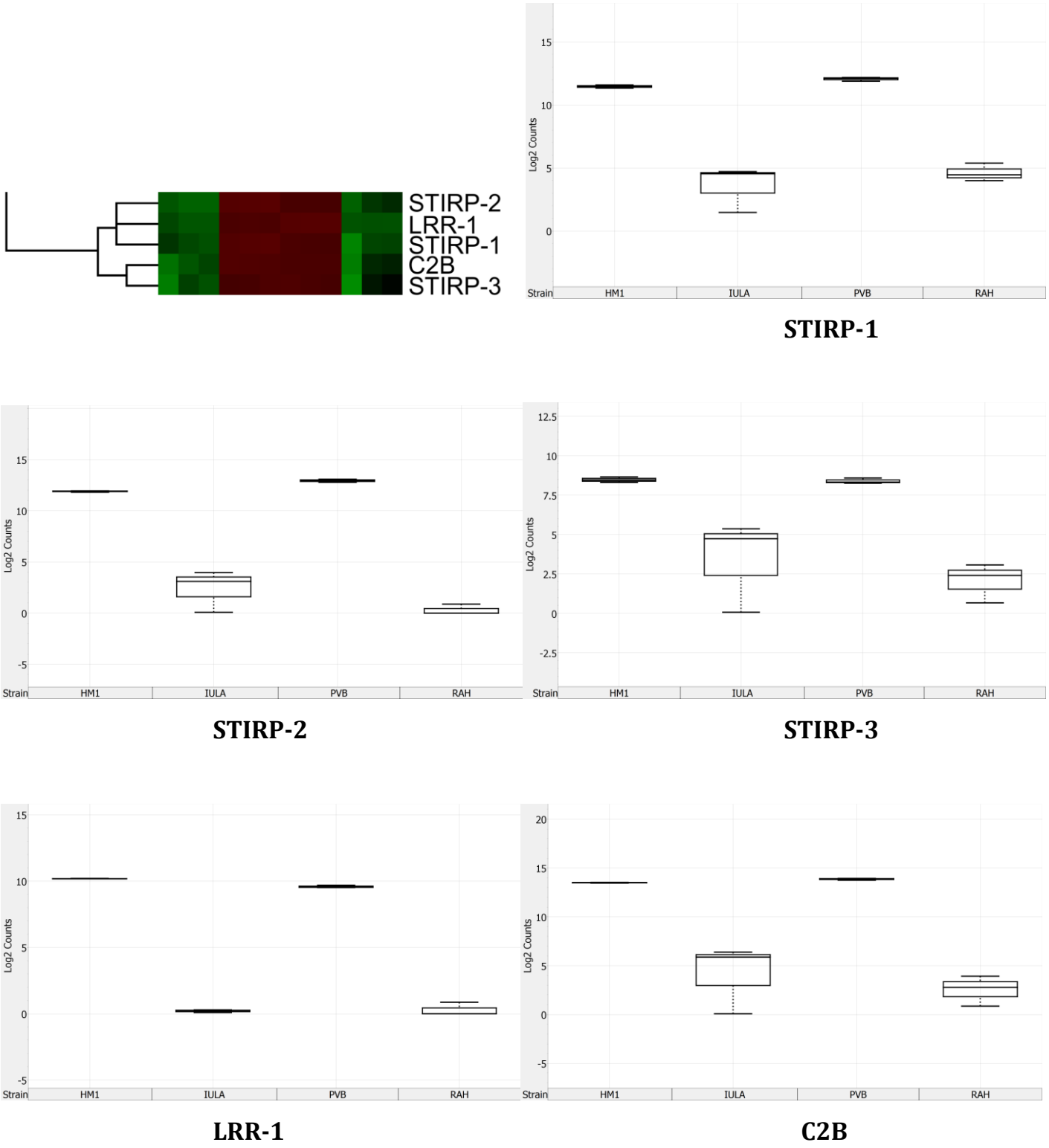


Figure 3.5: Expression levels of five virulence-associated genes in the four *E. histolytica* strains. All these five virulence-associated genes were clustered together in the previous heatmap due to their distinctive expression pattern showing remarkable co-upregulation in the two most virulent laboratory-adapted strains (i.e. HM-1:IMSS and PVBM08B).

3.3.3 Resemblance of expression in HM-1:IMSS and PVBM08B likely reflects the close degree of clinical virulence and outcome

To compare transcriptional profiles of this representative gene set between strains, the log_2-transformed nCounter data of each transcript, i.e. log_2(Counts), of three strains: Rahman, PVBM08B and IULA:1092:1 were plotted on the vertical axis of the scatterplot against the horizontal axis represented by those of HM-1:IMSS. As depicted in Figures 3.6-3.8, the red line running through the red spots of HM-1:IMSS represents the transcript levels measured in HM-1:IMSS. Therefore, gene spots of the other strain located above the line indicate upregulated genes whereas those below the line represent genes with downregulation, relative to HM-1:IMSS. Apparently, the majority of 53 genes with putative functions in HM-1:IMSS were plotted higher in position than Rahman, indicating higher expression levels of such particular genes than those in Rahman as shown in Figure 3.6. However, it is possible to explain that most of the representative genes were chosen based on the RNA-Seq data of upregulated genes in virulent strains, potentially resulting in this experimental biased finding.

Remarkably, most of the PVBM08B gene spots were plotted very close to the red line of HM-1:IMSS as shown in Figure 3.7, suggesting that these two virulent strains are likely to have relatively similar expression levels, especially in virulence-associated genes. It was found that five transcripts: CP-Ps (EHI_019390), ARIEL1-1 (EHI_123850), DL2 (EHI_182460), AIG1-1 (EHI_180390) and Hgl3 (EHI_077500) show significant downregulation in PVBM08B. Interestingly, these 3 of 5 downregulated genes are ARIEL1-1 which is absent in *E. dispar*, AIG1-1 implicated for bacterial killing and Hgl3 responsible for host cell adhesion, implying that higher levels of these three transcripts in HM-1:IMSS than PVBM08B increase the virulence power of HM-1:IMSS trophozoites to survive in microbiome environment and to adhere and invade the intestinal mucosa.

In case of IULA:1092:1 (see Figure 3.8), a total of 11 virulence-associated genes were found to be significantly downregulated in IULA:1092:1 relative to HM-1:IMSS. Of these, CP-A4 (EHI_050570) and CP-A5 (EHI_168240) are the key cysteine proteinases for MUC2 degradation. Also, two EhSTIRPs and the light and heavy subunits of the Gal/GalNAc lectin complex, responsible for host cell adhesion and contact-dependent cytotoxicity, were found to be downregulated. These findings suggest that the less virulence potential in IULA:1092:1, compared to PVBM08B and HM-1:IMSS, is due to the downregulation of key virulence processes.

Originally reported in 1967, HM-1:IMSS was isolated from a colonic biopsy of patient with dysentery in Mexico [24,36]. This strain has been extensively studied for virulence and pathogenesis of amoebiasis as well as widely used as a reference strain in genomic research. In accordance with previous clinical findings, HM-1:IMSS has been characterised as the most virulent strains since it could cause the highest prevalence of ALA occurring in 19% of newborn hamsters injected with 20 amoebic cells and around 90% of hamsters inoculated with 2,000 cells, compared to eleven other strains [71].

Intriguingly, based on the expression profiles of a representative gene set, PVBM08B and HM-1:IMSS exhibit the closest similarity whereas Rahman and IULA:1092:1 are notably different as illustrated in Figure 3.9A. In contrast to their phylogeny in Figure 3.9B, the multidimensional scaling plot shows a clear wide separation between Rahman and IULA:1092:1 and exhibits resemblance in expression between PVBM08B and HM-1:IMSS. This is also consistent with the PCA plot (see Figure 2.8) in Chapter 2, showing a close similarity between PVBM08B and HM-1:IMSS and a wide separation between Rahman and IULA:1092:1. The null hypothesis would be that the strains (i.e. Rahman and IULA:1092:1) that are most similar genetically should have the most similar expression profiles, however this data do not support this. Conversely, it would appear that expression profiles of the two most virulent strains (i.e. PVBM08B and HM-1:IMSS) are most similar to each other.

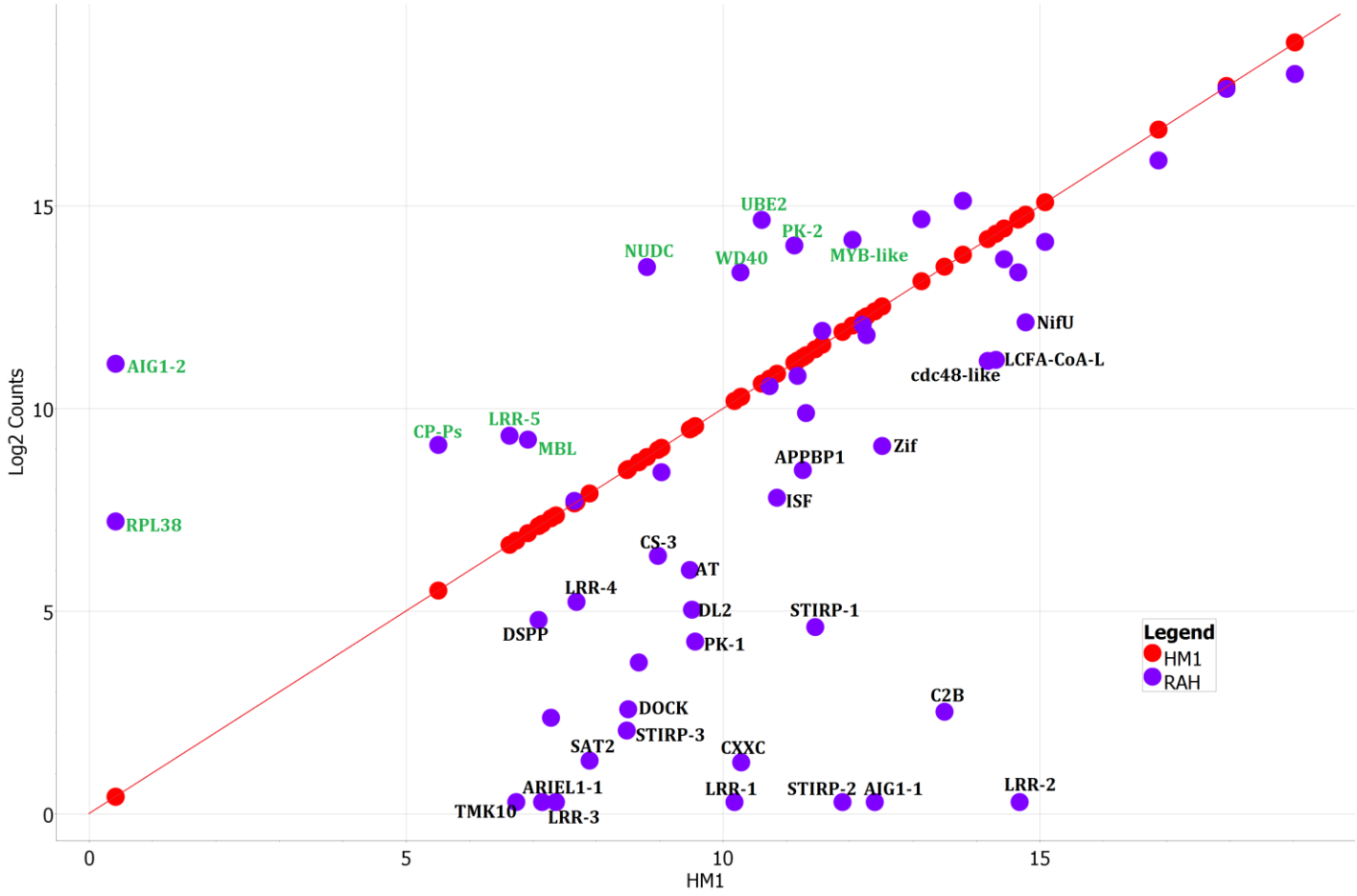


Figure 3.6: **Comparison of the transcriptional profiles between HM-1:IMSS and nonvirulent Rahman.** Log_2-transformed NanoString counts of 53 representative transcripts in Rahman were plotted on the Y-axis against the X-axis, represented by those in HM-1:IMSS. The red line passing through the red spots means the expression levels of transcripts in HM-1:IMSS. Blue spots located above the red line indicate upregulated Rahman genes relative to HM-1:IMSS whereas those below the line represent downregulated genes. Transcript identifiers were designated for the spots with significantly differential expression with different text colours: green for upregulation and black for downregulation. With respect to HM-1:IMSS, 25 and 10 of 53 representative genes in Rahman were found to be significantly downregulated and upregulated, respectively.

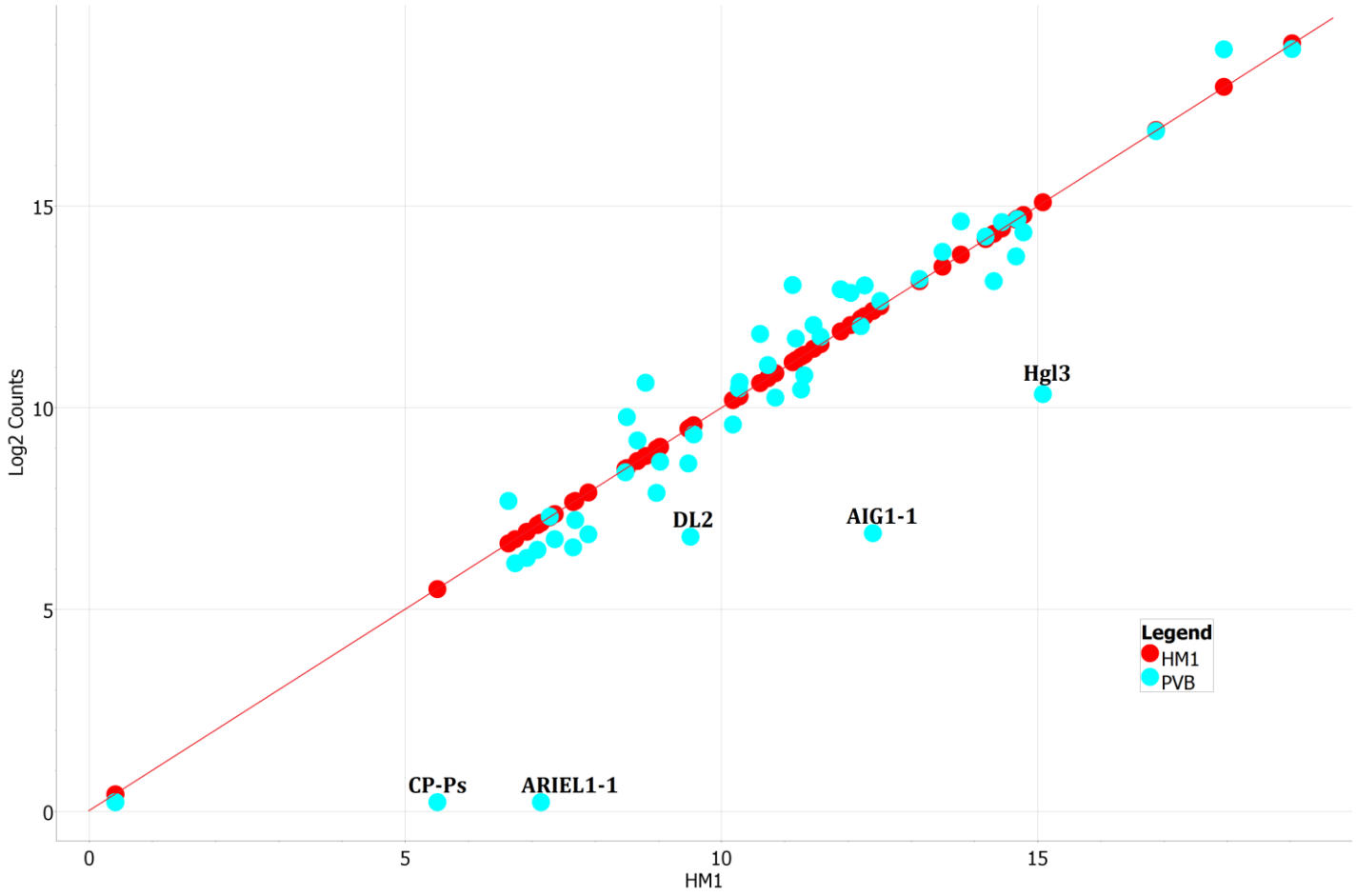


Figure 3.7: Comparison of the transcriptional profiles between HM-1:IMSS and PVBM08B. Log_2-transformed NanoString counts of 53 representative transcripts in PVBM08B were plotted on the Y-axis against the X-axis, represented by those in HM-1:IMSS. Sky blue spots located above the red line indicate upregulated PVBM08B genes relative to HM-1:IMSS whereas those below the line show PVBM08B genes with downregulation. Five genes in black were found to be significantly downregulated in PVBM08B, compared to HM-1:IMSS.

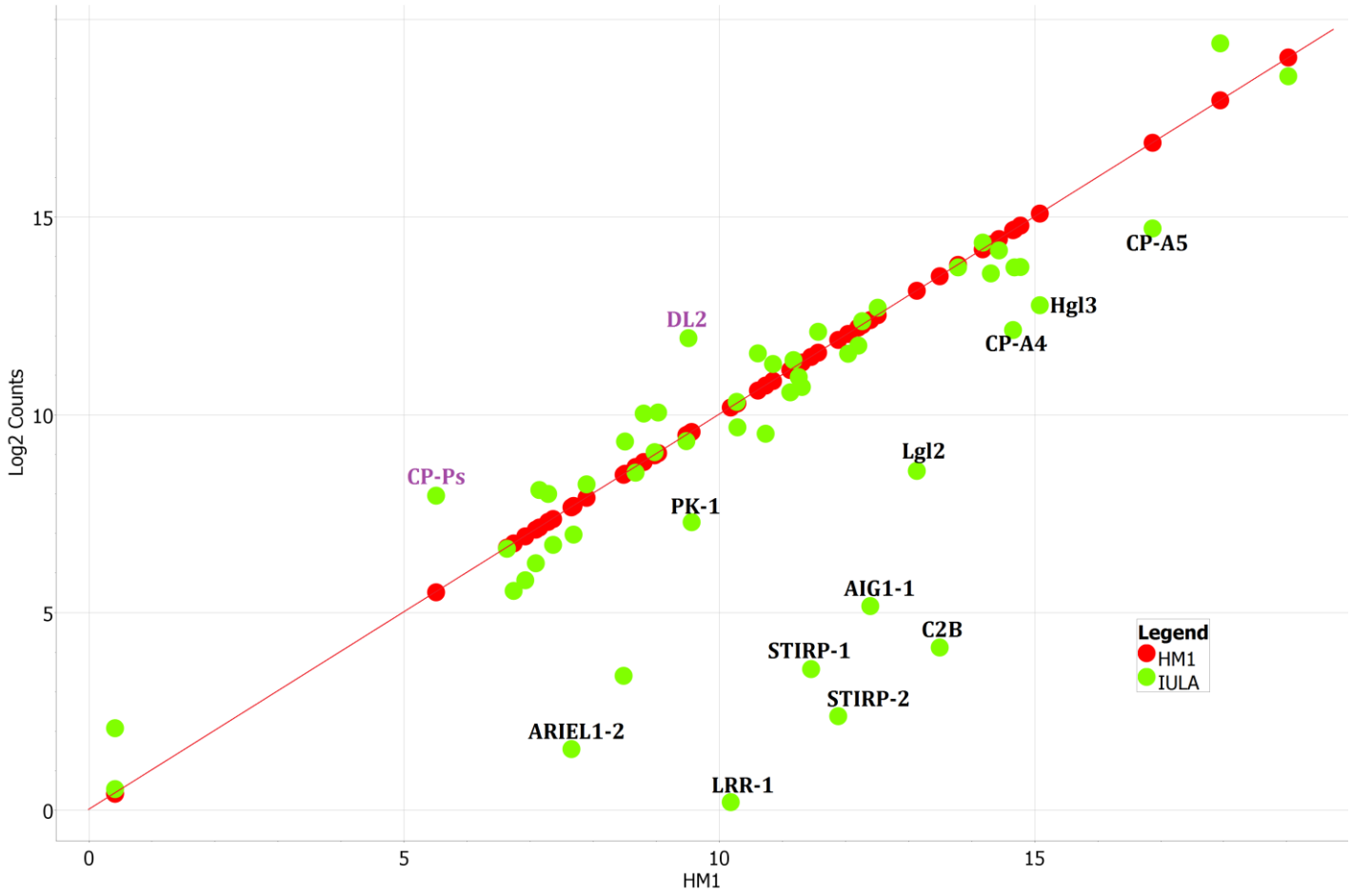
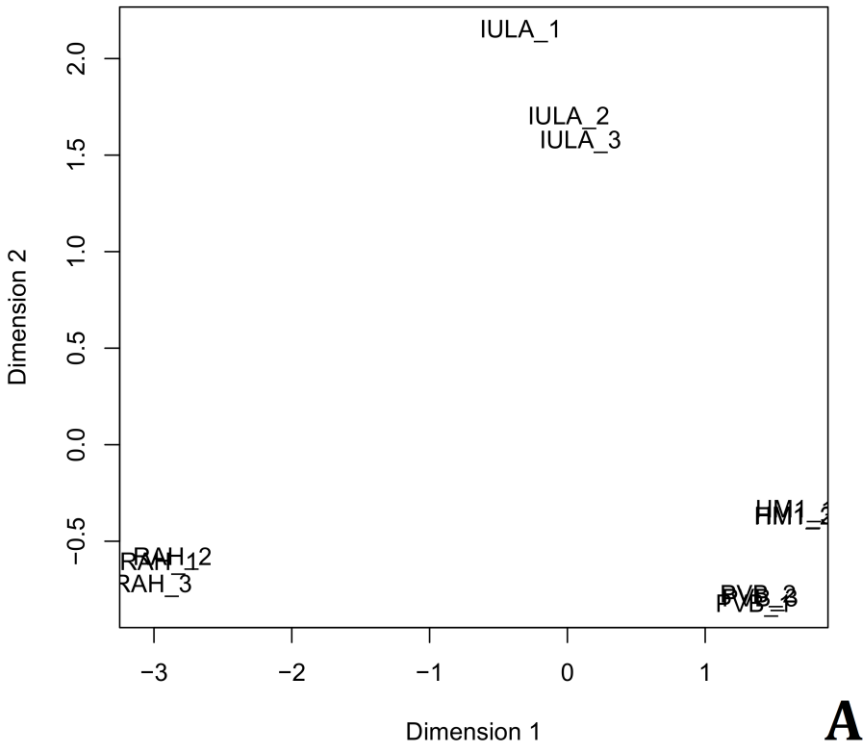


Figure 3.8: Comparison of the transcriptional profiles between HM-1:IMSS and IULA:1092:1. Log_2-transformed NanoString counts of 53 representative transcripts in IULA:1092:1 were plotted on the Y-axis against the X-axis, represented by those in HM-1:IMSS. Green spots located above the red line indicate upregulated IULA:1092:1 genes relative to HM-1:IMSS whereas those below the line represent IULA:1092:1 genes with downregulation. With statistical significance, two genes designated with violet (CP-Ps and DL2) were upregulated whereas eleven black-highlighted genes involved in key processes of virulence were downregulated in IULA:1092:1, compared to HM-1:IMSS.



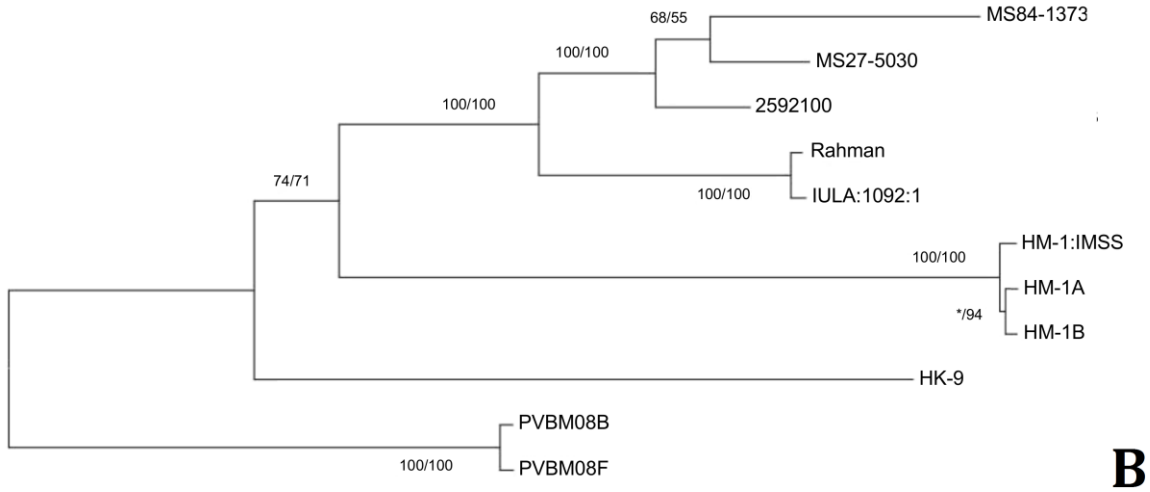


Figure 3.9: Similarity among the four *E. histolytica* strains, based on the NanoString gene expression profiles (A) and the whole genome SNP-based phylogenetic analysis (B). This tree figure (B) is reproduced with permission from Weedall *et al.*, 2012 [70].

3.4 Concluding remarks

In this study, NanoString nCounter® technology was applied to explore the expression profiles of 53 representative genes across the four *E. histolytica* strains. Normalised NanoString data exhibit a high correlation with the previous RNA-Seq study in terms of normalised count and gene expression fold change, indicating the reliability of this highly sensitive and multiplexed hybridisation detection for gene expression analysis. Based on hierarchical clustering analysis, gene family members encoding functionally similar or related proteins, i.e. EhSTIRPs and BspA-like LRRPs, appear to be similar in their expression profiles across the parasite strains. This finding also suggests that these virulence-associated multigene family members potentially share a common mechanism for transcriptional regulation, resulting in their similar relative expression tendencies in each *E. histolytica* strain.

Based on this representative gene set, it is obvious that there is a spectrum of virulence-associated gene expression among these four strains, reflecting their different degrees of virulence. Comparing the representative gene expression profiles between strains, the PVBM08B and HM-1:IMSS strains exhibit resemblance in their expression levels, likely representing their close degree of virulence and clinical outcome. Additionally, it was evidenced that the Rahman and IULA:1092:1 strains which are genetically similar have different expression profiles, implying that there are certain regulatory mechanisms which shape the transcriptomes of these two genetically similar strains in different directions, potentially leading to their different parasite behaviours [70].

Chapter 4: Correlation with the genomic data reveals that gene copy number variation (CNV) influences transcriptomic diversity among *E. histolytica* strains

4.1 Introduction

To understand genetic diversity among *E. histolytica* strains, SNPs have been recently evaluated in the sequenced genomes of ten *E. histolytica* strains [70]. In overall, sequence divergence from the HM-1:IMSS reference genome is quite low, 0.312-0.857 SNPs per kb. Fifty-three genes with five or more nonsynonymous SNPs compared to the HM-1:IMSS reference were identified as highly polymorphic genes [70]. Interestingly, these genes are profoundly involved in the host-parasite interaction, such as EhSTIRP genes, the intermediate chain Gal/GalNAc lectin genes *Igl1* and *Igl2*, the light chain Gal/GalNAc lectin gene *Lgl,* genes encoding BspA-like leucine-rich repeat proteins. As discussed before, the sequence polymorphisms of virulence-associated genes implicated in host-parasite interaction are evolutionarily advantageous for adaptation to the host immune response and parasite survival. However, the evidence of SNPs in this parasite could not explain all of the genotypic variations in relation to biological differences among *E. histolytica* strains due to low average SNP call in each strain as well as a small number of putative highly polymorphic genes.

The genome recharacterisation by Lorenzi *et al*., 2010 revealed a total of 897 protein-coding gene families consisting of 4,564 proteins and constituting ~56% of the *E. histolytica* proteome [25]. The gene families have five members in average, ranging from 2 to 149 members. Of these, seven large gene families with greater than 50 members were identified including families encoding BspA-like LRRPs, kinase domain-containing proteins, WD domain-containing proteins, small GTP-binding proteins, RNA recognition motif domain-containing proteins, RhoGAP domain-containing proteins and a large family of uncharacterised hypothetical proteins. Also, it is noteworthy that TEs account for approximately 20 % of the *E. histolytica* genome. Intriguingly, a considerable number of gene families including virulence-associated families were identified with a high physical association with TEs [25]. For example, 11 of 31 members of the *hsp70* gene family were found to be closely associated with TEs within 1 kb upstream or downstream. As previously reported in *D. melanogaster*, the *hsp70* promoter regions appear to be hotspots of transposition, potentially leading to gene expansion [25,268,269]. Hence, TEs located in close proximity to associated genes potentially drive the amplification and expansion of

gene copy number, ultimately leading to the increase in transcript expression and consequent parasite virulence [25,268,269].

More recently, large differences in coverage depth of genes were observed through the sequenced genomes of *E. histolytica* strains, indicating gene copy number variation (CNV) between strains [70]. Five hundred-fourteen genes in one or more strains were determined as putative high copy number genes with a coverage depth two fold or greater than the average of HM-1A and HM-1B [70]. Large members of these putative high copy number genes are functionally annotated in several biological processes such as 23 ribosomal protein genes for protein synthesis; 14 members of BspA-like leucine-rich repeat protein family implicated for host mucosal invasion; 9 members of AIG1-like protein family involved in bacterial killing; 4 peroxiredoxin genes associated with oxidative stress response; 5 genes encoding protein kinase domain-containing proteins participated in regulatory signaling pathways [70]. It is worth noting that these putative high copy number genes also play an important role in parasite virulence. As reported in Chapter 2, members of these high copy number genes relevant to virulence also displayed higher differential expression in the three virulent strains relative to the nonvirulent Rahman.

A high coverage region spanning over 12.4 kb (positions 21,000 to 33,380) of scaffold DS571330 was also reported and only present in the nonvirulent Rahman strain [70]. This high coverage region includes seven protein-coding genes (EHI_023840, EHI_023850, EHI_023860, EHI_023870, EHI_023880, EHI_023890 and EHI_023900) flanked by TEs as illustrated in Figure 4.7. Therefore, these seven genes are predicted to have upregulated expression levels proportional to their additional copies due to the segmental genome duplication. In addition, many putative missing genes with underbaseline RPKM values (< 1 and < 50% of the reference) were identified among *E. histolytica* strains, indicating that the gene family content is variable among strains [70].

As such, these genomic findings strongly suggest that amplification and loss of gene family members are key dynamic processes for genome plasticity [70]. Essentially, genomic plasticity due to differential copy number and gene family content is more pronounced than sequence polymorphisms and greatly contributes to the genomic diversity among *E. histolytica* strains [29,46,70]. Moreover, the ploidy, haploid chromosome number and chromosome size which are variable under different growth conditions and between life cycle stages potentially contribute to considerable genomic size plasticity among strains [37].

The impact of gene CNVs on the genomic diversity among strains can be seen in other human protozoan parasites such as *Trypanosoma cruzi*, causing Chagas disease in Latin America [269]. A large number of gene CNVs have been previously reported, showing extensive genotypic diversity among *T. cruzi* strains [269]. 'Hotspot regions of CNV' were reported for all chromosomes in *T. cruzi*. Interestingly, gene CNVs in *T. cruzi* are widespread and likely to be focal in highly repetitive regions including large multigene families encoding surface proteins, trans-sialidases, mucins, and mucin-associated proteins. As gene members of the surface protein families share relatively high sequence homology and encode the surface proteins directly subject to the host immune response, the recombination and subsequent variation in gene copy number potentially occur under positive selection due to the host immunological pressure [269,270]. Moreover, substantial expansion and variation of certain genes have also been reported to be associated with different biological characteristics among *T. cruzi* strains. For instance, a total of 37 β-galactofuranosyl transferase genes responsible for the synthesis of complex mucin glycans were found to be located in genomic regions of high CNV and their expansion and variation are consistent with the heterogeneity in the mucin glycan biosynthesis among *T. cruzi* strains [271].

Similar to *T. cruzi*, *E. histolytica* exhibits the highly repetitive genomic structure and gene CNV is a significant major contributor to a high degree of genomic plasticity among *E. histolytica* strains which exhibit variability in their virulence [70]. Therefore, I hypothesised that CNVs may contribute to phenotypic differences and differential clinical virulence among *E. histolytica* strains. To investigate at the genome-wide level whether gene CNVs correlate with transcriptomic diversity among strains, the genomic mapped read data of the four *E. histolytica* strains (i.e. Rahman, PVBM08B, HM-1:IMSS and IULA:1092:1) previously obtained from SOLiD™ library sequencing were used for the pairwise scatterplot analysis in relation the existing transcriptomic data of the four strains in this study.

4.2 Materials and Methods

4.2.1 Whole genomic and transcriptomic data of sequenced strains used in this study

SOLiD™ (**S**equencing by **O**ligonucleotide **Li**gation and **D**etection)-based genomic HTSeq-count data of all 8,333 genes in the four *E. histolytica* strains (i.e. Rahman, PVBM08B, HM-1:IMSS and IULA:1092:1) were kindly offered by Dr. Gareth Weedall, the Liverpool School of Tropical Medicine, Liverpool, UK for this study [70]. Briefly, the Burrows-Wheeler Aligner (BWA) software was applied to map SOLiD™ sequenced reads to the HM-1:IMSS reference genome sequence with mapping parameters as previously described [70,272]. Only uniquely mapped reads were used for downstream analysis. Then, the BAM alignment files for all strains were sorted and transformed to SAM files by the SAMtools software [273]. The HM-1:IMSS genome annotation file (release 2.0, AmoebaDB-2.0_EhistolyticaHM1IMSS.gff file), indicating the locations of 8,333 genes in the genome, was used to count reads aligned to each gene [26]. HTSeq-count software was applied using the sorted SAM files as an input to count reads in features with following options: **-m** <mode> intersection-strict; **-i** <id attribute> Parent; **-t** <feature type> exon; **-s** <stranded> no [119]. Finally, the obtained HTSeq-count data for each strain was normalised by millions of total HTSeq-count reads generated.

For whole transcriptome, HTSeq-count data obtained from RNA-Seq experiment previously described in the 'Materials and Methods' of Chapter 2 were normalised by millions of total HTSeq-count reads generated for each strain.

In this correlation study, the R Statistics software package version 3.1.2 (http://CRAN.R-project.org) was used to plot the genomic data using $\log_2$-transformed values of the ratio between genomic reads per million of total SOLiD™ library reads (RPM) of two contrasting strains, against the transcriptomic data using $\log_2$-transformed values of the ratio between HTSeq-counts per million of total Illumina ScriptSeq™ v2 library reads (RPM) of the same two strains as shown in Figures 4.1-4.6 [267]. Pearson's product-moment correlation tests were conducted to determine whether copy number variation (CNV) correlates with the differential transcript levels between two contrasting strains. Also, percentile rank analysis was performed in each comparison to compare the distribution range between the CNV and the relative expression levels.

4.3 Results and Discussion

4.3.1 Scatterplot analysis between genomic and transcriptomic data reveals that gene copy number variation is associated with differential expression across *E. histolytica* strains, implying the evolution of virulence

In Figures 4.1-4.6, all 8,333 *E. histolytica* genes could be plotted into four quadrants where a centre point locates at x-axis = 0 and y-axis = 0. The majority of genes are plotted towards the zero values (x = 0, y = 0) within the 10^{th} – 90^{th} percentile range on the horizontal axis, inferring that these genes are similar in their gene copy number between contrasting strains. However, transcriptional range was found to be variable across the strains due to its wider 10^{th} – 90^{th} percentile range on the vertical axis.

Comparing between Rahman and PVBM08B, a significant positive correlation with Pearson's correlation coefficient (r) = 0.3544, P-value < $2.2e^{-16}$ was found, indicating that an increase in copy number of a particular gene in Rahman or PVBM08B can upregulate its gene expression level, compared to the other strain. Notably, gene spots in the graph area of quadrant III [x-axis: $\log_2$(Rahman genomic rpm/ PVBM08B genomic rpm] < 0 and y-axis: $\log_2$(Rahman transcript rpm/ PVBM08B transcript rpm) < 0] represent genes that have both lower copy number and lower transcript in Rahman than PVBM08B. In other words, these spots in the third quadrant refer to genes with higher copy number and higher transcripts in PVBM08B. Likewise, spots plotted in quadrant I [x-axis: $\log_2$(Rahman genomic rpm/ PVBM08B genomic rpm) > 0 and y-axis: $\log_2$(Rahman transcript rpm/ PVBM08B transcript rpm) > 0] are designated for genes with higher copy number and higher expression in Rahman than PVBM08B. Taken together, these data in quadrants I and III represent genes whose CNV positively correlates with their expression difference between these two strains and this finding could explain differences in virulence between Rahman and PVBM08B, as well as other phenotypic differences.

Analysing the trend in this scatterplot, it is clear that the linear spread over 10^{th} to 90^{th} percentile is greater on the y-axis (transcript: P_{10} = -0.9503, P_{90} = 1.0374) than that of the x-axis (CNV: P_{10} = -0.4940, P_{90} = 0.4220) as illustrated in Figure 4.1B. It can be inferred that the majority of genes in both Rahman and PVBM08B have a broader range of expression than range of gene copy number.

As shown in Figure 4.2, a positive correlation between CNV and differential expression was found (r = 0.3531, P-value < $2.2e^{-16}$) between Rahman and HM-1:IMSS. Remarkably, a large number of genes were plotted in quadrant III, reflecting the skewed

data distribution. This cluster of genes in quadrant III has lower gene copy number and corresponding decreased expression in Rahman relative to HM-1:IMSS. However, difference in copy number and transcript abundance between these two strains in quadrant III might be partly due to the absence of genes (missing genes) and no corresponding transcripts in Rahman, resulting in the strong negative variables on both axes. On contrary, only few spots are found in quadrant I, referring to a small number of genes with higher copy number and corresponding upregulated transcripts in Rahman relative to HM-1:IMSS. Essentially, this skewed data representation strongly suggests that genome plasticity largely contributes to the transcriptomic difference and phenotypic variability between Rahman and HM-1:IMSS.

Similar to the previous plot, the expression percentile range on the y-axis (P_{10} = -1.1579, P_{90} = 1.1177) is wider than the CNV percentile range on the x-axis (P_{10} = -0.6242, P_{90} = 0.4719) as shown in Figure 4.2B, indicating that most of the genes in these two strains exhibit more variability in expression than that in gene copy number.

Different from the previous two pairs, the scatterplot in Figure 4.3 shows the Pearson's correlation coefficient towards zero (r = 0.0265, P-value = 0.0153), indicating no correlation was found between gene CNV and differential expression in Rahman and IULA:1092:1. Moreover, the majority of genes in these two strains are clustered around the central point (x =0, y =0), pointing out that most of the genes in these two contrasting strains are similar in their gene copy number and transcript abundance. Consistently, difference in the 10^{th}- 90^{th} percentile ranges of the CNV (P_{10} = -0.8431, P_{90} = 0.6229) and relative expression (P_{10} = -1.1924, P_{90} = 1.2521) is narrower than that of the previous pairs, supporting the low variation of gene copy number and transcript abundance in these two strains.

In accordance with the genealogical analysis in Figure 1.5 of Chapter 1, Rahman was clustered very closely together with IULA:1092:1, based on a total of 3,696 SNP sites throughout the genome. This phylogenomic finding reflects the similarity of genomic structure between Rahman and IULA:1092:1. However, I demonstrated the marked transcriptional variation of 53 representative virulence-associated genes between Rahman and IULA:1092:1 as demonstrated in Figure 3.9A of Chapter 3. Altogether, transcriptional variation of virulence-associated genes between Rahman and IULA:1092:1 seems to be not dominated by gene copy number variation but might be influenced by other regulatory elements, e.g. transcription machinery or epigenetic regulations.

Interestingly, a cluster of seven protein-coding genes (EHI_023840, EHI_023850, EHI_023860, EHI_023870, EHI_023880, EHI_023890 and EHI_023900) known to have

segmental duplication found only in the Rahman strain was also found to have both higher copy number and higher transcript levels in Rahman relative to all other strains as highlighted with their AmoebaDB_IDs in Figures 4.1A, 4.2A and 4.3A [70]. In other words, it could be stated that the Rahman segmental duplication is a good example of gene CNV which contributes to differential expression among strains. The details of this segmental genome duplication will be discussed later.

Comparing between PVBM08B and HM-1:IMSS, CNV was found to be associated with transcriptomic variation (r = 0.4051, P-value < $2.2e^{-16}$) with a wider percentile expression range (P_{10} = -0.9233, P_{90} = 0.8369) compared to a CNV range (P_{10} = -0.4723, P_{90} = 0.5488) as shown in Figure 4.4 (A and B). Also, the notable skewed distribution could be seen in quadrant III, representing genes with higher copy number as well as higher expression in HM-1:IMSS than PVBM08B. This indicates that transcriptomic variation between PVBM08B and HM-1:IMSS is dominated by variation of the gene copy number in these two contrasting strains. In other words, phenotypic differences between these two virulent strains are determined in part by higher expression of highcopy number genes in the HM-1:IMSS strain.

Likewise, a significant positive correlation between CNV and expression difference (r = 0.3587, P-value = $2.2e^{-16}$) was found in a contrasting pair of PVBM08B and IULA:1092:1 as shown in Figure 4.5A. Also, a wider variability of relative expression (P_{10} = -0.9746, P_{90} = 0.9838) was demonstrated in most of the genes, relative to their CNV (P_{10} = -0.7683, P_{90} = 0.6493) as illustrated in Figure 4.5B. Different from the previous contrast pair, genes whose CNV positively correlates with their expression were plotted in both quadrant I and III, indicating that copy number expansion has occurred in both these two strains.

For HM-1:IMSS and IULA:1092:1, CNV positively correlates (r = 0.3788, P-value < $2.2e^{-16}$) with expression difference between these two strains and the skewed data distribution could be observed in quadrant I, indicating that expression variability between two strains is due to the variation of gene copy number, higher in HM-1:IMSS than IULA:1092:1. It seems to be that variability of relative expression levels between genes in these two strains is slightly larger than their CNV due to a small difference between expression and CNV percentile ranges as shown in Figure 4.6B.

As explained in all contrasting pairs, it could be argued that genomic plasticity is a main driver of gene expression diversity among *E. histolytica* strains. The positive correlation between CNV and transcriptomic variation could explain phenotypic differences including virulence variability among strains. Also, it is interesting that no correlation was

found in comparison between Rahman and IULA:1092:1 strains that are very genetically similar, however differential expression of virulence-associated genes still exists in these two strains. Hence, this finding suggests that there should be other mechanisms of gene regulation that contributes to their phenotypic differences in addition to gene CNVs.

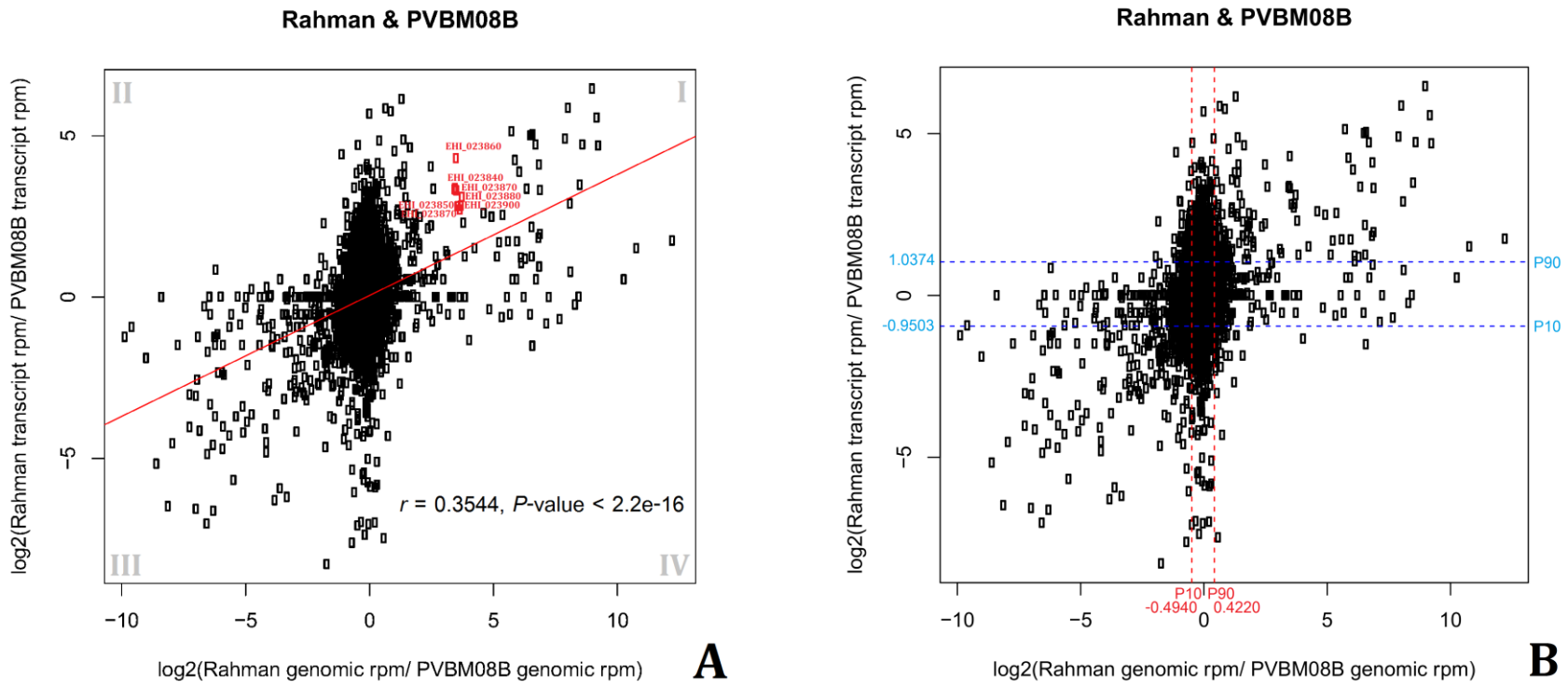


Figure 4.1: **Positive correlation between CNV and relative expression levels in Rahman and PVBM08B.** The Pearson's correlation coefficient (*r*) and the percentile range are shown in Parts A and B of the figure, respectively. In Part B, the range over 10th to 90th percentile on the y-axis (P_{10} = -0.9503, P_{90} = 1.0374; in blue) is wider than that on the x-axis (P_{10} = -0.4940, P_{90} = 0.4220; in red), inferring that most of the genes in these two *E. histolytica* strains have variable expression levels, compared to a percentile range of gene copy numbers. Spots plotted in quadrants I and III obviously represent genes which their CNV correlate positively with their relative expression across these two strains. The AmoebaDB_IDs of 7 genes located on scaffold DS571330 with segmental genome duplication as illustrated in Figure 4.7 are labelled in quadrant I of Part A, revealing that their high copy number obviously contributes to their high transcript level. I = Quadrant I; II = Quadrant II; III = Quadrant III; IV = Quadrant IV.

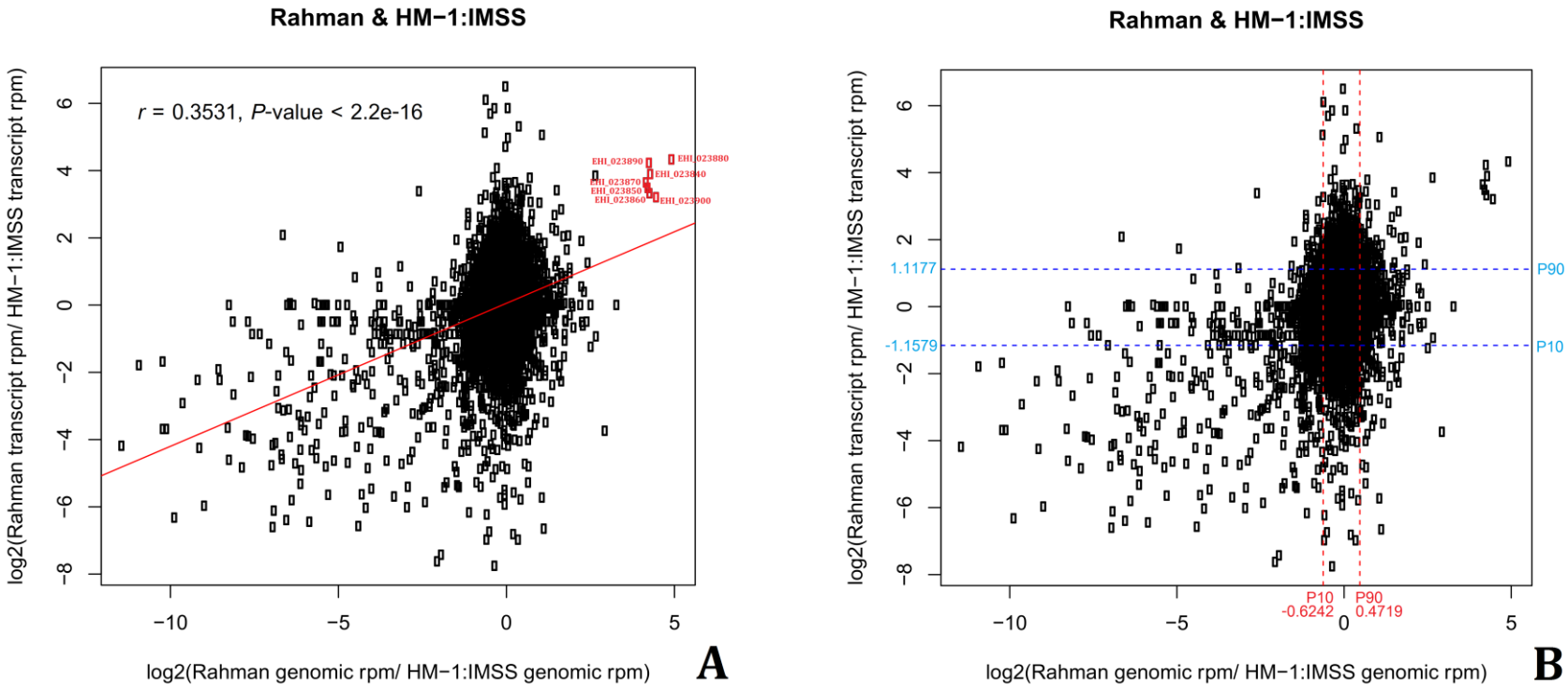


Figure 4.2: **Positive correlation between CNV and relative expression levels in Rahman and HM-1:IMSS.** In quadrant III, it obviously shows up the skewed data representing a lot of genes with lower copy number and lower expression in Rahman than in HM-1:IMSS whereas only few spots are found in quadrant I, referring to a small number of genes with higher copy number and corresponding upregulated transcripts in Rahman relative to HM-1:IMSS. The AmoebaDB_IDs of 7 genes located on scaffold DS571330 with segmental genome duplication as illustrated in Figure 4.7 are also labelled in quadrant I of Part A. In Part B, a wider range of 10th to 90th percentile on the y-axis (P_{10} = -1.1579, P_{90} = 1.1177; in blue) than that on the x-axis (P_{10} = -0.6242, P_{90} = 0.4719; in red) suggests that the majority of genes have more variable expression levels than their CNV.

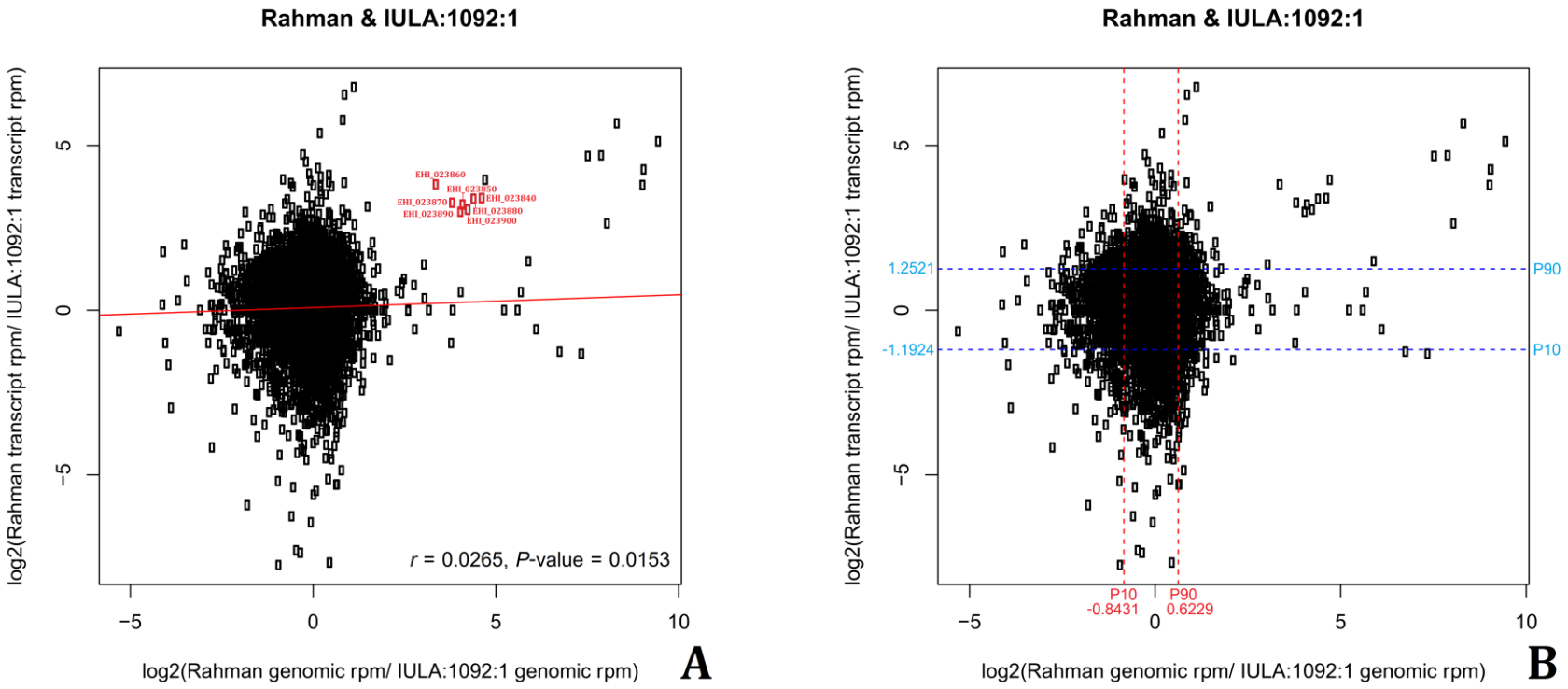


Figure 4.3: **No correlation between CNV and relative expression levels in Rahman and IULA:1092:1.** In Part A, the Pearson's correlation score is towards zero (r = 0.0265, P-value = 0.0153), meaning no correlation between CNV and expression levels in these two strains. As depicted in the plot, the majority of genes are clustered together around the zero value of both the x- and y-axes. This central tendency could be inferred that most of the genes in these two strains are likely similar in their copy number and corresponding transcript level. Narrower difference between percentile ranges of both two axes reflects slightly more variability of relative expression levels between genes relative to their CNV as shown in Part B. The AmoebaDB_IDs of 7 genes located on scaffold DS571330 with segmental genome duplication as illustrated in Figure 4.7 are also labelled in quadrant I of Part A.

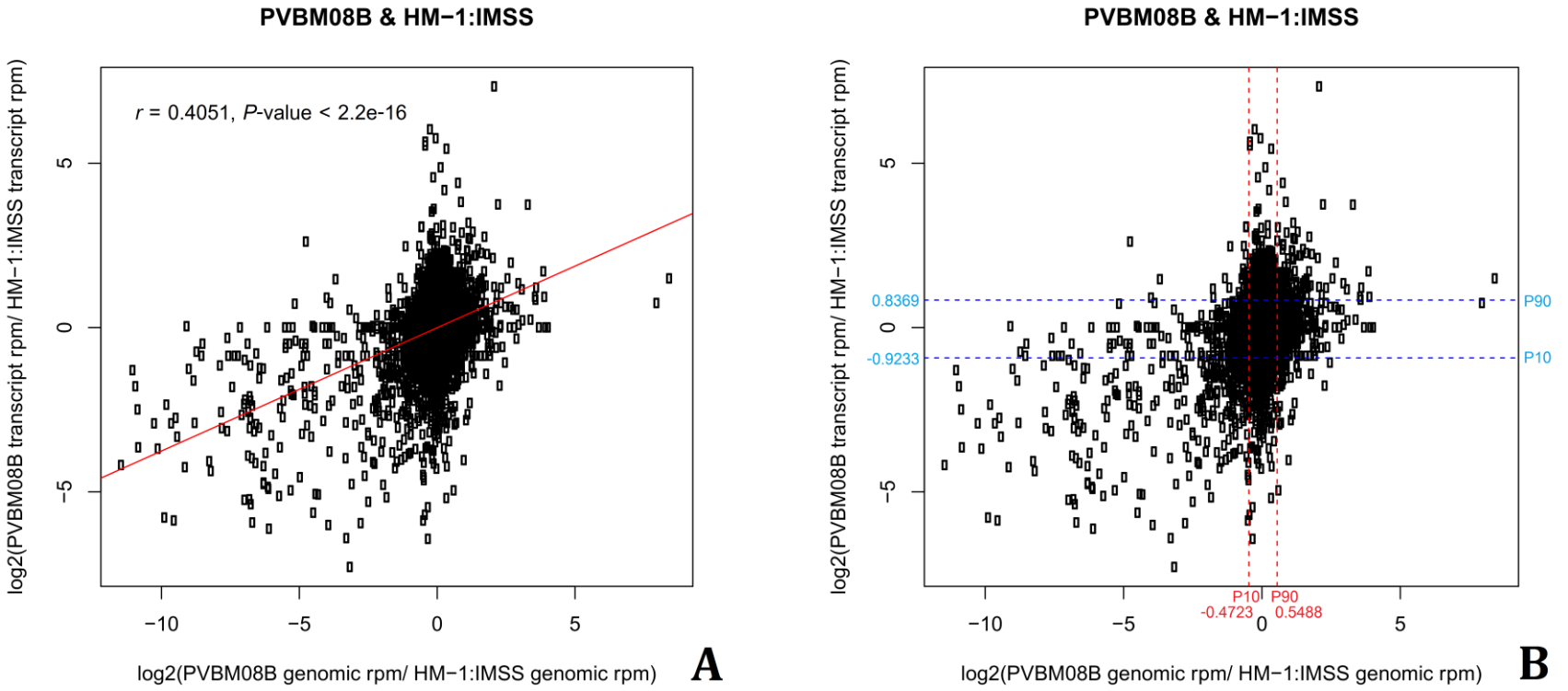


Figure 4.4: **Positive correlation between CNV and relative expression levels in PVBM08B and HM-1:IMSS.** Remarkably, the skewed data could be observed in quadrant III where many spots are located in and represent genes with lower copy number and lower expression in PVBM08B than in HM-1:IMSS. In Part B, a wider expression percentile range on the y-axis (P_{10} = -0.9233, P_{90} = 0.8369; in blue) compared to a CNV range on the x-axis (P_{10} = -0.4723, P_{90} = 0.5488; in red) indicates that the majority of genes exhibit a higher variability in transcript levels than their CNV. Lower CNV and relatively downregulated expression of genes found in the third quadrant likely contributes to a lower degree of virulence in PVBM08B relative to HM-1:IMSS.

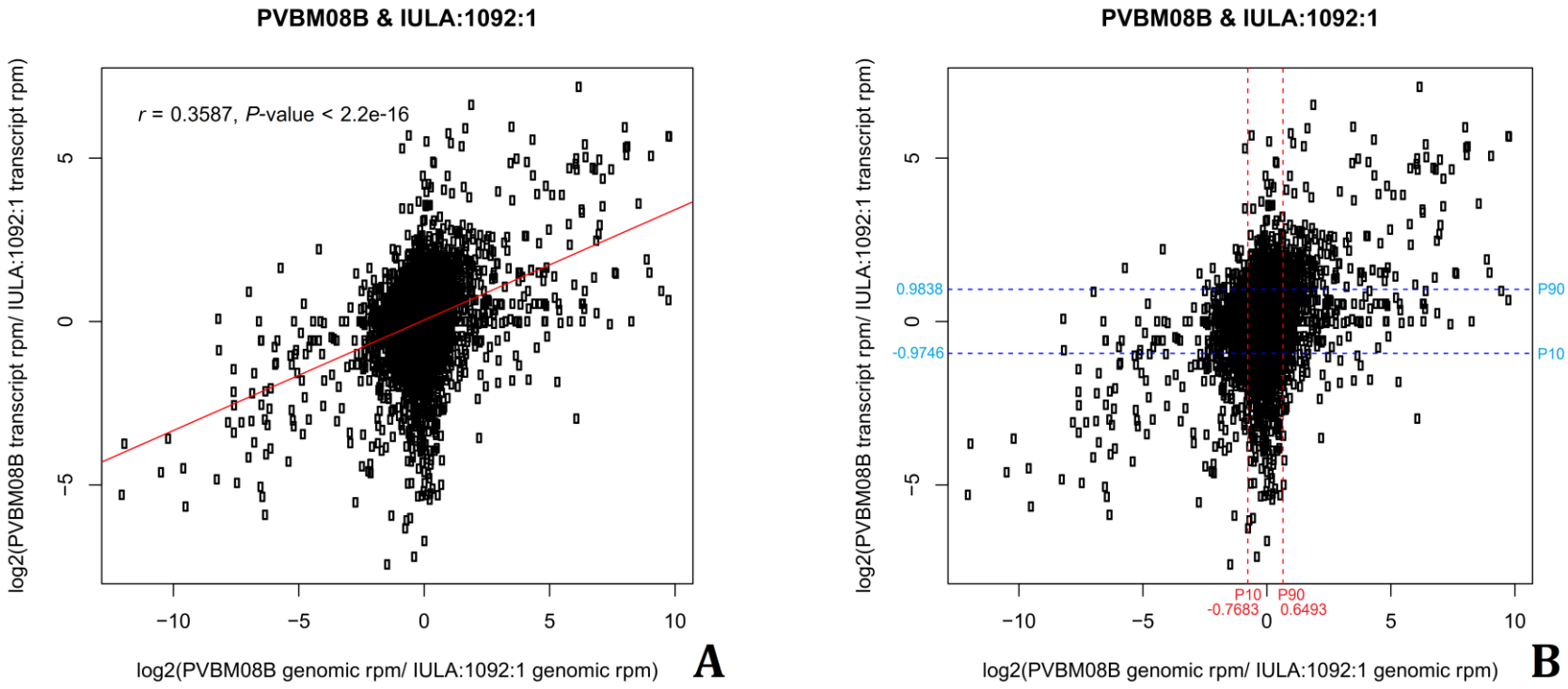


Figure 4.5: **Positive correlation between CNV and relative expression levels in PVBM08B and IULA:1092:1.** As shown in the plots, gene spots plotted in quadrants I and III obviously represent genes whose CNV correlates positively with their relative expression across these two strains. In Part B, a wider range on the y-axis (P_{10} = -0.9746, P_{90} = 0.9838; in blue) than that of the x-axis (P_{10} = -0.7683, P_{90} = 0.6493; in red) suggests that the majority of genes have more variable expression levels than their CNV.

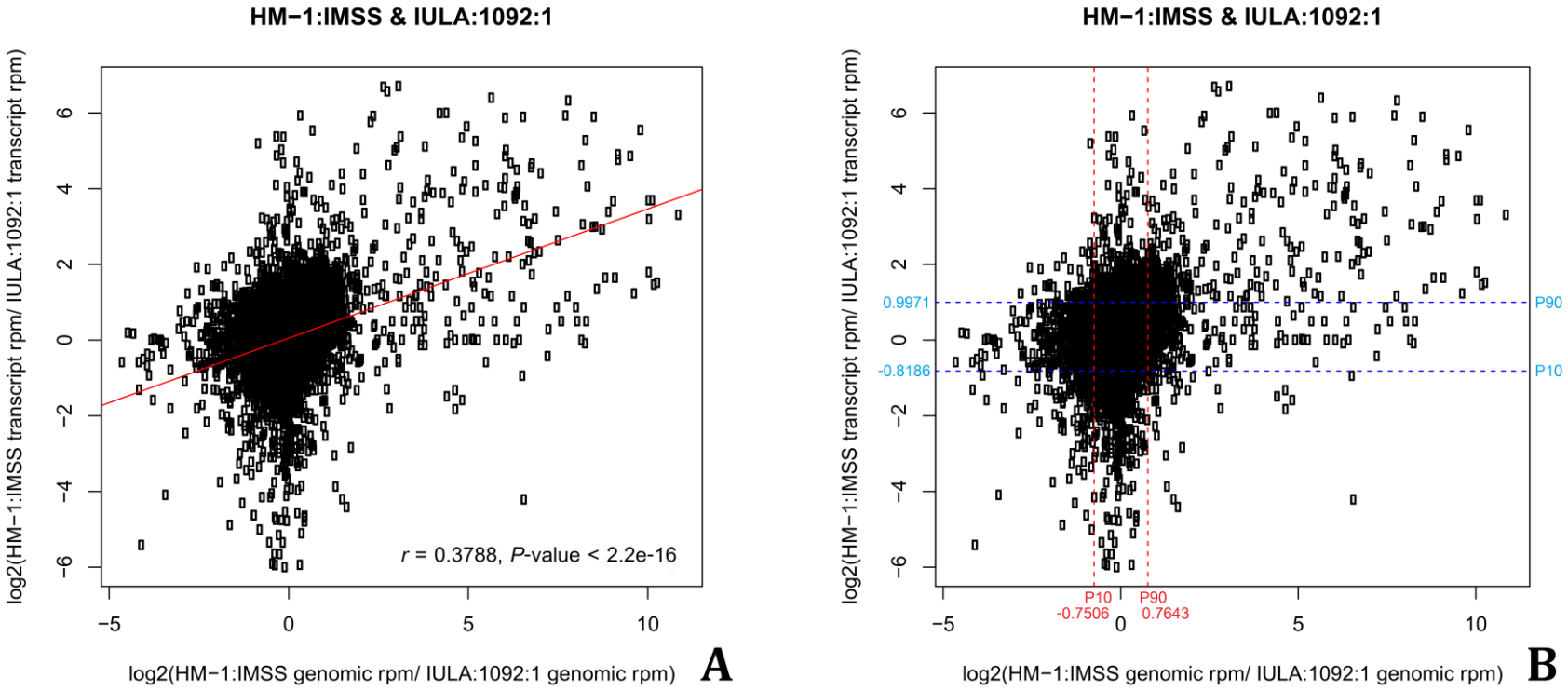


Figure 4.6: **Positive correlation between CNV and relative expression levels in HM-1:IMSS and IULA:1092:1.** Obviously, several gene spots disperse in quadrant I, indicating the skewed data distribution of genes with higher copy number and higher expression in HM-1:IMSS than in IULA:1092:1. Excluding these gene spots in the first quadrant, a slight difference between a percentile range (10^{th} to 90^{th}) of the y-axis (P_{10} = -0.8186, P_{90} = 0.9971; in blue) and the x-axis (P_{10} = -0.7506, P_{90} = 0.7643; in red) indicates slightly more variability of relative expression levels between genes compared to their CNV as shown in Part B. Thus, it implies that higher CNV and relatively upregulated expression of genes found in the first quadrant likely contributes to a higher degree of virulence in HM-1:IMSS compared to IULA:1092:1.

4.3.2 Expression of genes located in the part of scaffold DS571330 in Rahman are enhanced due to the segmental genome duplication process, implying the potential of functionality

After improvement of genome assembly and annotation of *E. histolytica* by Lorenzi *et al*., 2010, novel features were discovered such as the presence of segmental genome duplication of scaffold regions, up to 16 kb, with specific attributes as well as the high association of protein families such as BspA-like protein family, AIG1-like family and Hsp 70 family with repetitive elements [25]. As repetitive elements make up approximately 20% of the *E. histolytica* genome and have a tendency to form large TE clusters, it is likely to contribute to genomic instability of this parasite, including partial genome duplication [25,153]. Four types of segmental genome duplication (D1-D4) have been previously reported in the HM-1:IMSS genome [25]. D1 and D2-types of segmental gene duplication are found to be flanked by 2.3 kb and 1.2 kb inverted repeats (IRs) respectively whereas D3 and D4-types are in close proximity to TEs, mostly EhLINE1 without any flanking IRs. Interestingly, D3 contains a set of genes implicated in a variety of cellular processes, suggesting the functionality of this type of duplication [25].

As demonstrated in Figure 4.7, independent Rahman SOLiD™ and 454 sequencing experiments previously done by Weedall *et al*, 2012 reveals the presence of an expanded region found only in the nonvirulent Rahman strain and spanning over 12.4 kb (positions 21,000 to 33,380) of scaffold DS571330 [70]. This amplified region includes a cluster of seven genes encoding a 60S ribosomal protein L38 (EHI_023840) for protein synthesis, a hypothetical protein (EHI_023850), a protein kinase domain-containing protein (EHI_023860) implicated for phosphorylation and signaling, a WD domain-containing protein (EHI_023870) involved in protein-protein interaction and signal transduction, a ubiquitin-conjugating enzyme family protein (EHI_023880) responsible for proteosomal degradation, a nuclear movement protein (EHI_023890) and a hypothetical protein (EHI_023900). This duplicated segment is also flanked by three repetitive elements: *Entamoeba* repeat element 2 (ERE2) and long interspersed nuclear elements (EhLINE1 and EhLINE2 retrotransposons) [70]. It is interesting that repeat clusters, including repetitive elements in this case, frequently mark the synthenic breakpoints between *E. histolytica* and *E. dispar*, reflecting their adaptive role in generating the genomic diversity among *Entamoeba* species and strains [25,70,153]. Also, this putative segmental genome duplication in the Rahman strain is similar to the D3 type segmental genome duplication previously described in the HM-1:IMSS strain [25].

As illustrated in Figure 4.7, the plot shows the profile of coverage depth in $\log_{10}$-transformed scale across the entire length of scaffold DS571330. The ratio of coverage depth between Rahman and HM-1B in both unduplicated and duplicated regions was calculated to estimate the copy number of the duplicated region. As expected, the median ratio in the duplicated region was 25.0 while in the unduplicated region was just 1.1, indicating that the cluster expansion occurred many times in the Rahman strain [70]. However, from this genomic observation one can only infer this as evidence of its potential for functionality but we can not imply more about its functional role beyond the genomic relevance.

Integrating this observation with the transcriptomic data in Chapter 2, Table 2.9 shows significantly lower transcript levels ($\log_2$FC ≤ -2, FDR-adjusted *P*-value < 0.05) of these seven genes in PVBM08B, HM-1:IMSS and IULA:1092:1, compared to nonvirulent Rahman. Thus, it is likely that this segmental genome duplication of scaffold DS571330 in the Rahman strain contributes to its higher expression levels of such genes located in the expanded segment than those of the three virulent strains. Consistently, the correlation between CNV and mRNA abundance of these seven expanded genes across all four strains as shown in Figure 4.8 confirms that higher transcript abundance of these genes in the nonvirulent Rahman strain is as a result of higher gene copy number due to the segmental genome duplication.

In *E. histolytica*, ribosomal RNA repeats exist exclusively in high-copy-number circular episomal plasmids with varying sizes (15-25 kb) between *E. histolytica* strains [241,274-277]. Also, coding sequences for hemolysins were found to be within inverted rRNA repeats on the episomal plasmid [278]. This finding implies that such hemolysin-coding sequences are much higher in gene copy number than those if they are located on the chromosome and their increased copies potentially result in large amounts of hemolysins, which may be associated with increased virulence and amoebic invasion [279]. This illustrates how gene copy number variation may contribute to differential phenotypes such as virulence via its effect upon the expression of key virulence-associated genes. Therefore, very high copy number of the gene cluster on scaffold DS571330 in Rahman might provide some selective advantages to the parasite, possibly selected by long-term axenic cultivation. Episomal plasmids can play a pivotal role in gene amplification [279]. Whether this, or tandem duplication, is the mechanism of expansion of the amplified segment of the Rahman genome remains to be determined.

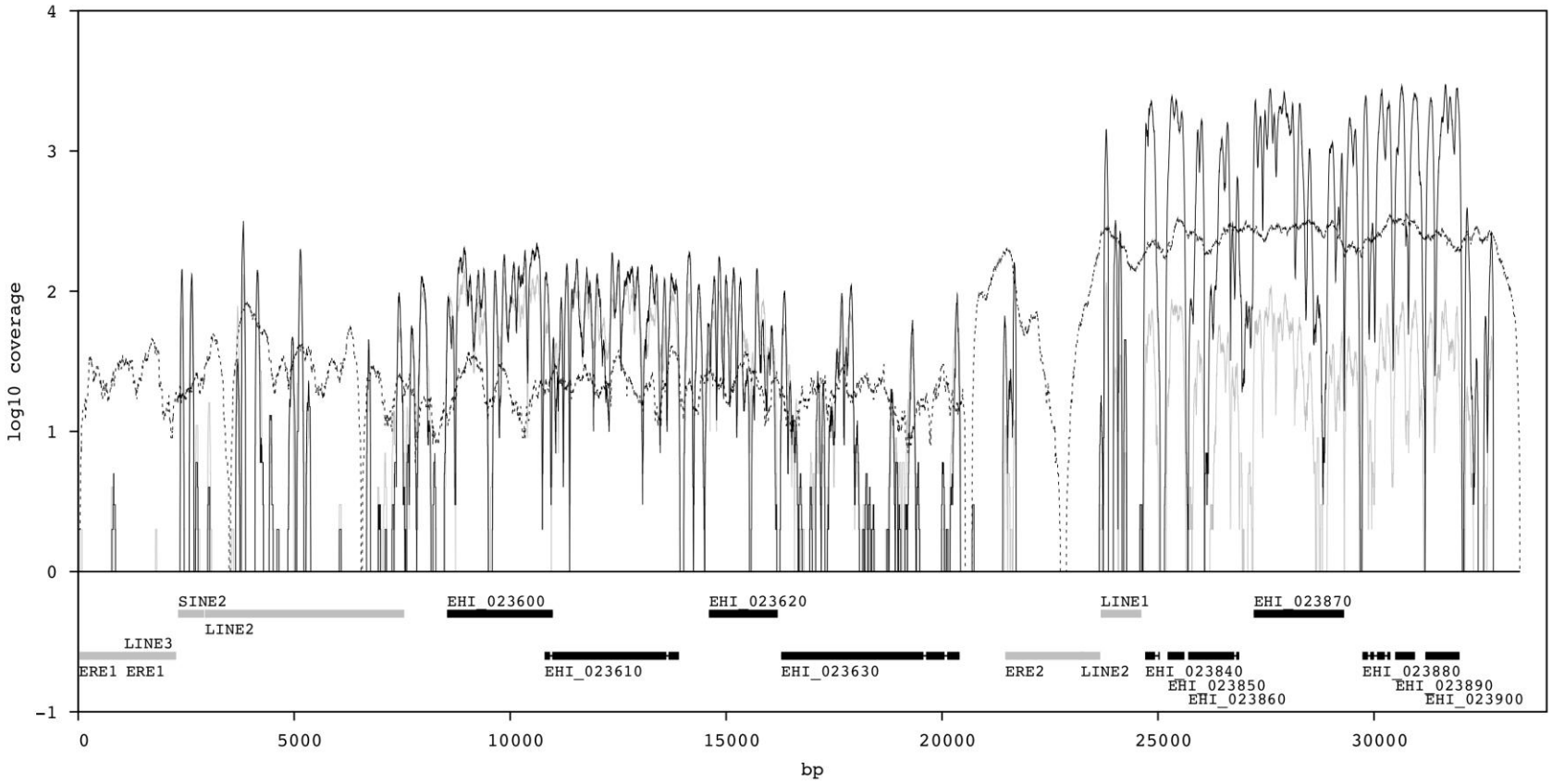
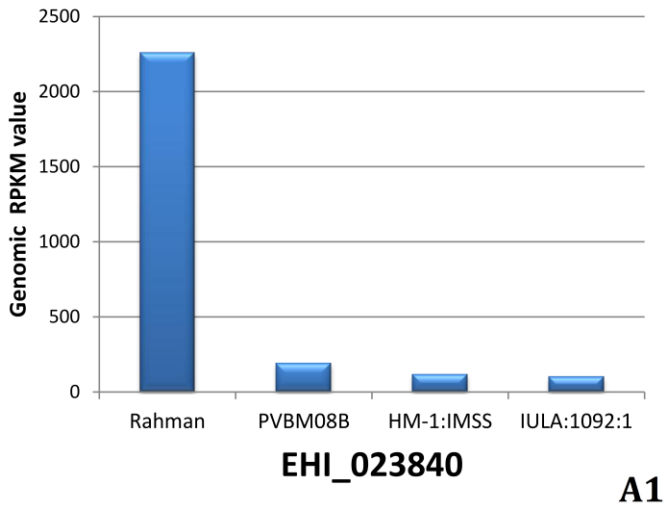
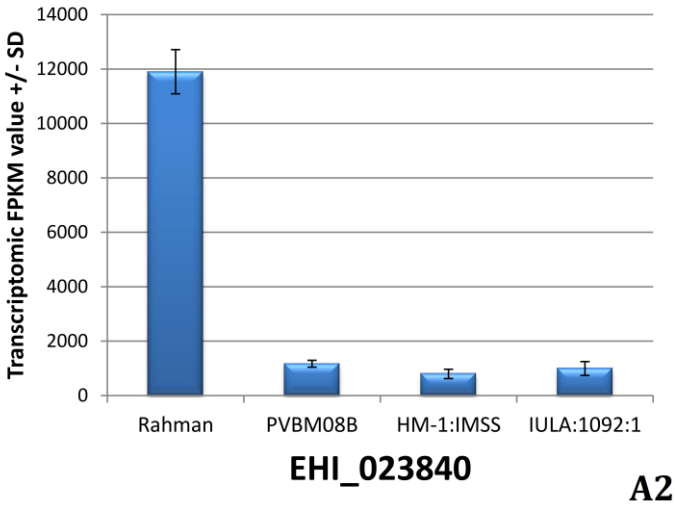


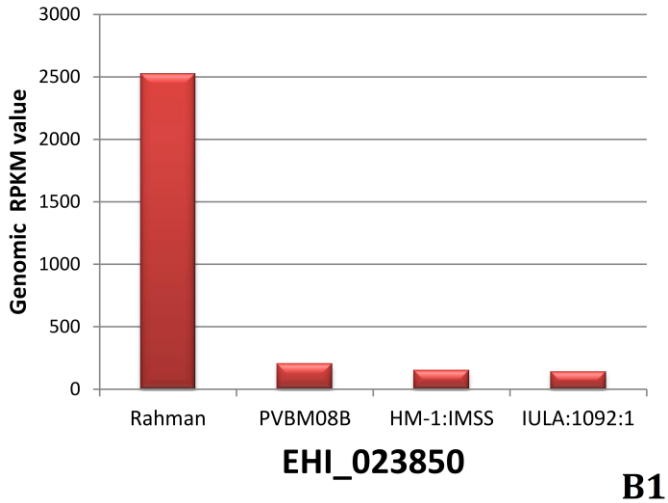
Figure 4.7: Segmental genome duplication on scaffold DS571330 in the nonvirulent *E. histolytica* Rahman strain. High coverage region spans over seven genes (EHI_023840, EHI_023850, EHI_023860, EHI_023870, EHI_023880, EHI_023890 and EHI_023900) with flanking repetitive transposable elements. The SOLiD™ and 454 coverage data in the Rahman strain are represented by the black and dashed lines, respectively. The HM-1B SOLiD™ coverage data represented by the grey line is used as a control. This plot is reproduced with permission from Weedall *et al.*, 2012 [70].



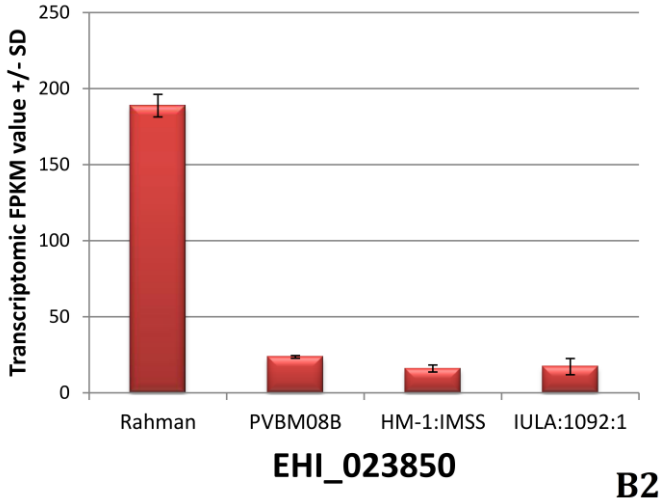
A1



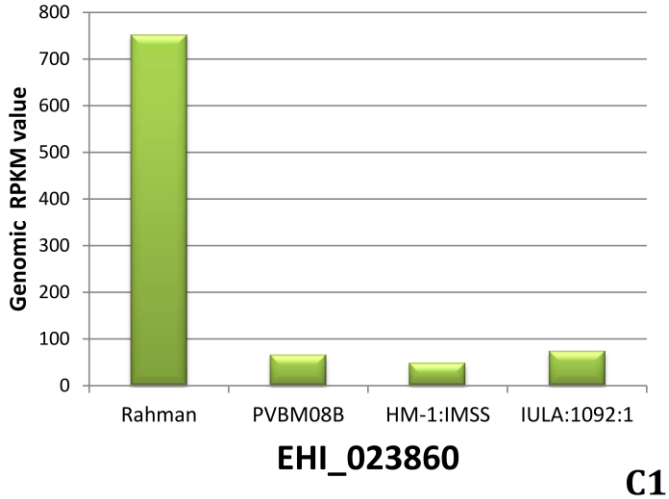
A2



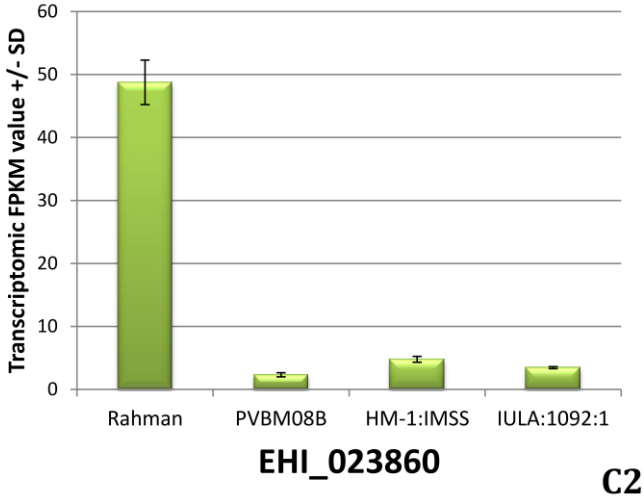
B1



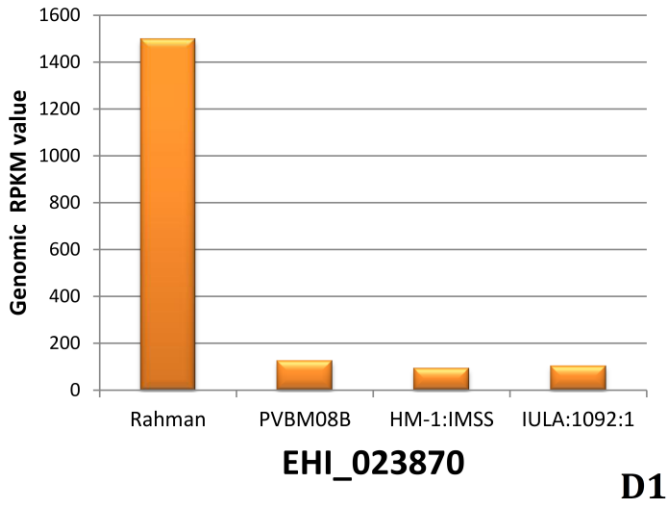
B2



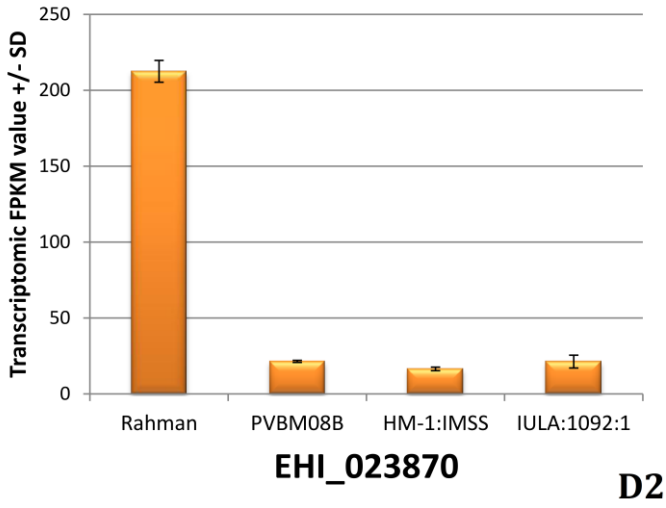
C1



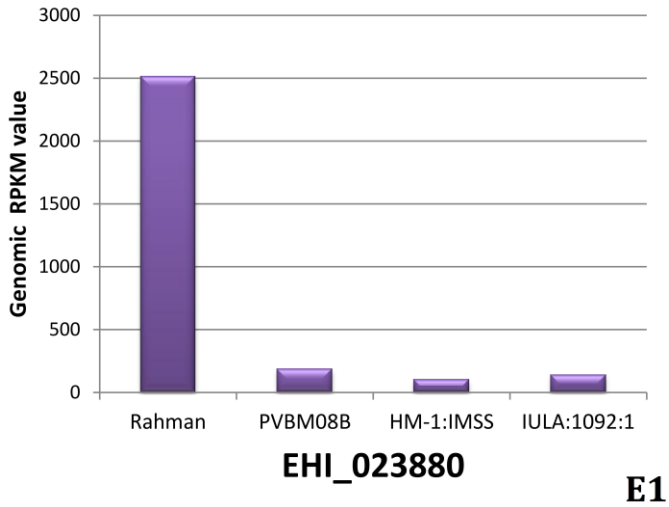
C2



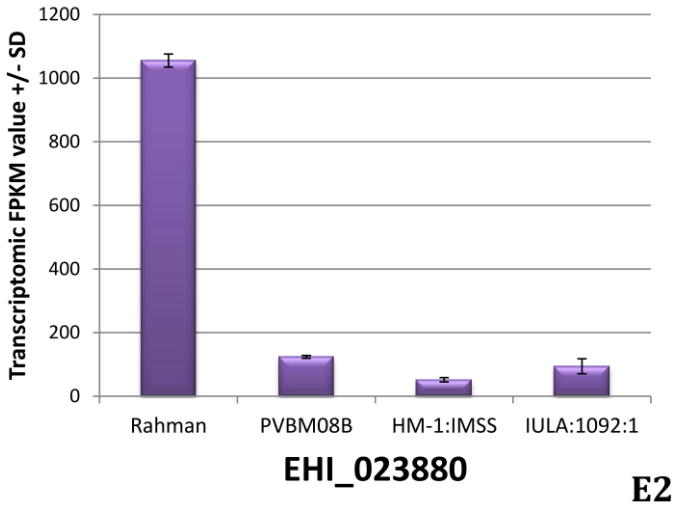
EHI_023870 **D1**



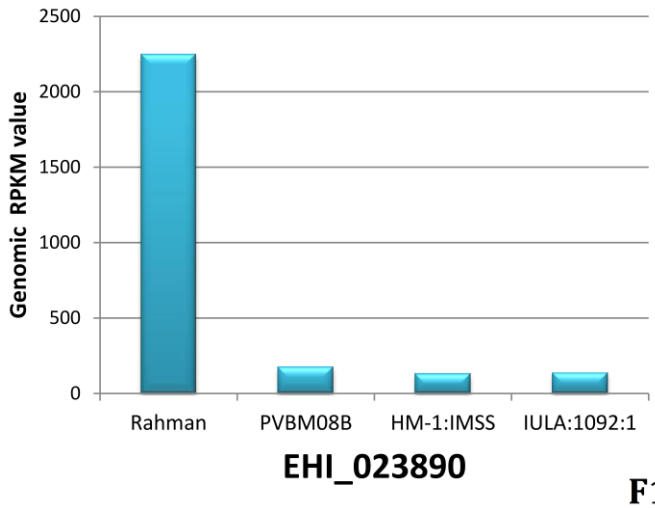
EHI_023870 **D2**



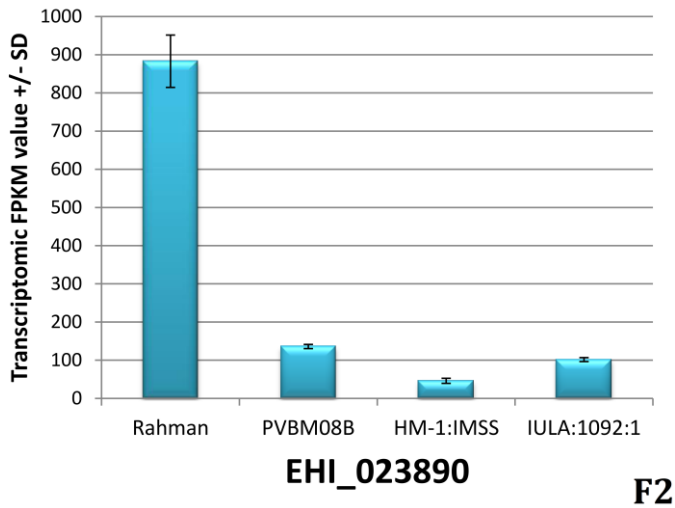
EHI_023880 **E1**



EHI_023880 **E2**



EHI_023890 **F1**



EHI_023890 **F2**

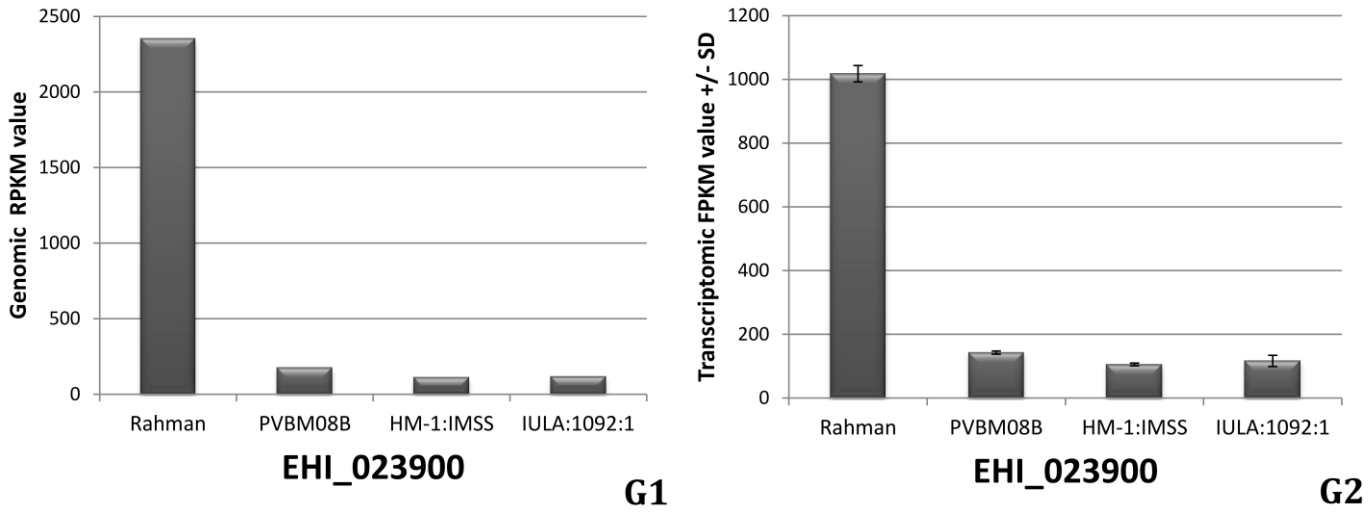


Figure 4.8: Correspondence between genomic copy number variation and differential transcript abundance of seven protein-coding genes located on scaffold DS571330 in the four *E. histolytica* strains. Genomic copy number variation of seven genes (EHI_023840, EHI_023850, EHI_023860, EHI_023870, EHI_023880, EHI_023890 and EHI_023900) across all four strains are represented by genomic RPKM values as plotted in Parts A1, B1, C1, D1, E1, F1 and G1, respectively. Differential transcript abundance of such seven genes across strains are represented by transcriptomic FPKM ± standard deviation (SD) as plotted in Parts A2, B2, C2, D2, E2, F2 and G2, respectively. Notably, all seven expanded genes in the nonvirulent Rahman strain exhibit higher in their gene copy number and corresponding transcript level than those in the other strains.

4.4 Concluding remarks

E. histolytica is believed to be almost exclusively asexual through binary fission with rare frequency of meiotic recombination but conversely shows the complex multiclonal population structure and variable biological features among strains [29,46]. This finding raised the hypothesis that biological diversity among parasite strains should be as consequences of genomic diversity. SNPs are rather limited throughout the *E. histolytica* genome [46,70]. Conversely, there appears to be a lot of gene CNVs among the genomes of the *E. histolytica* strains, reflecting a high degree of genomic plasticity and variability in gene family content [70]. Moreover, this present data show that patterns of CNV contribute to differential expression profiles, therefore we can extrapolate that differences in gene copy number between genomes could contribute to the variation in phenotypic characteristics, including virulence, among parasite strains.

The high repetitiveness of the *E. histolytica* genome could lead to genomic structural diversity, such as segmental genome duplications, resulting in gene copy number variation. Such genome plasticity can also be seen in other human protozoan parasites such as *Trypanosomes* and *Leishmania*, suggesting that CNV is not uncommon and also is a potentially important mechanism of generating genetic diversity and regulating gene expression levels in almost exclusively asexual parasite groups [70,241,269,280,281].

Chapter 5: Analysis of the small RNA transcriptome and its potential role in regulating gene expression, especially of virulence-associated genes

5.1 Introduction

Gene silencing is an epigenetic cellular process for control of specific gene expression found in most eukaryotic organisms [87-91]. RNAi is a key regulatory mechanism for post-transcriptional gene silencing that inhibits gene expression in a sequence-specific manner [87-91]. This regulatory process is typically due to small interfering RNA species (siRNAs) that are double stranded RNA molecules with lengths of 20-25 bp. These siRNAs play an important role in RNA interference by complementary base pairing with expressed mRNA transcripts, causing subsequent mRNA degradation and translational inhibition [91,282]. Also, this specific type of small RNA species can function in concert with associated proteins to target the genomic loci for transcriptional gene silencing [283].

In *E. histolytica*, experimental gene silencing has been achieved by using exogenous dsRNA and siRNA, suggesting the presence of a functional sRNA-mediated silencing machinery in the parasite [284-287]. Zhang *et al*., 2008 demonstrated the presence of three endogenous sRNA populations with sizes of approximately 27, 22 and 16 nt in *E. histolytica* [85]. The 27 nt endogenous sRNA population constitutes the majority of the overall sRNA transcriptome. This distinctive sRNA expression could be commonly found in trophozoites of the reptilian parasite *E. invadens* and the nonvirulent *E. dispar*, suggesting a conserved mechanism that exists throughout the *Entamoeba* species [85].

Interestingly, the 27 nt sRNA population possesses 5'-polyphosphate termini and was significantly enriched in the sRNA fraction co-immunoprecipitated with *E. histolytica* Argonaute-2 (EhAGO2-2, EHI_125650), indicating the association with Argonaute protein in the formation of the RNA-induced silencing complex (RISC), which is responsible for post-transcriptional gene silencing in concert with siRNA or miRNA [85,93]. The 5'-polyphosphate termini can be specifically found in secondary siRNAs expressed in *Caenorhabditis elegans* [288]. Essentially, such secondary siRNAs in *C. elegans* are Dicer-independent and synthesised by RNA-dependent RNA polymerase (RdRp) incorporating the nucleotide with 5'-triphosphate terminus for the first base position whilst other siRNAs are processed by RNase III Dicer and exhibit the typical 5'-monophosphate and 3'-hydroxyl termini [288,289]. In addition, *C. elegans* secondary siRNAs were shown to play a role in the RNAi mechanism by 5'-biased antisense base-pairing to a target mRNA. Therefore, the

structural similarity of 5'-polyphosphorylated sRNAs in *E. histolytica* suggests that these sRNAs might participate in a similar gene regulatory mechanism.

Likewise, these distinctive sRNAs with 5'-polyphosphate in *E. histolytica* were shown to map predominantly antisense to the 5' end of target genes and there is a negative correlation between their abundance and target gene expression levels, strongly suggesting a regulatory role in the siRNA pathway [85,86,92]. Most recently, Zhang *et al.*, 2013 demonstrated that these 5'-polyphosphorylated sRNAs associated with EhAGO2-2 play an important role in regulating virulence-associated gene expression in a strain-specific manner [92].

The miRNAs are small non-coding RNAs with 21-23 nt in length that play a key role in regulation of gene expression in cellular proliferation and development [290,291]. These miRNA molecules occur in many organisms including animals, plants, and viruses [292,293]. These regulatory molecules can recruit the RISC to block mRNA targets with partial antisense complementarity and cause translational repression, mRNA degradation and mRNA deadenylation [290,291,293]. In contrast to siRNAs, miRNAs, especially in animals, can target many different mRNA transcripts with incomplete base pairing whilst siRNAs specifically regulate their complementary mRNA transcripts with perfect matches and induce gene silencing only in a specific gene target [294].

Putative miRNAs have been identified in other human protist parasites including *G. lamblia* and *T. vaginalis* [97-100]. Also, genes encoding proteins involved in miRNA- and siRNA-mediated machineries have been identified in the *E. histolytica* genomic data, providing evidences that both siRNA- and miRNA-associated regulatory mechanisms are likely to exist in this parasite [95,295]. As previously mentioned, the 27 nt antisense sRNAs with 5'-polyphosphate termini have been proven to be the siRNAs responsible for gene silencing in *E. histolytica* trophozoites. Therefore, it is possible for miRNAs to be expressed and play a crucial role in post-transcriptional gene silencing in this parasite.

I hypothesised that differential virulence among *E. histolytica* strains could be potentially regulated by a miRNA-mediated mechanism. Potential miRNA-regulated genes in *E. histolytica* have been reported by De *et al*., 2006 and Mar-Aguilar *et al*., 2013 [94,95]. However, these studies were performed only in the HM-1:IMSS reference strain. Therefore, the previously reported information of novel predicted miRNAs cannot elucidate their biological relevance to differential virulence among strains due to lack of miRNA expression data in other strains. Therefore, the experiments in this chapter were designed using sRNA libraries which were size-selected at 21-23 nt, most likely representing the expected size of

miRNAs and prepared from the four *E. histolytica* strains to enable us to compare the difference of sRNA levels between strains.

Essentially, the aims of this chapter are to explore the differences in the sRNA transcriptomic landscapes among the four *E. histolytica* strains and to investigate the possible roles of antisense sRNAs in parasite gene regulation using the deep sequencing data of the size-fractionated sRNA libraries. To scrutinise the presence of miRNAs and their putative roles in relevance to virulence, novel miRNA candidates were also predicted in all size-fractionated sRNA datasets by the specialised miRDeep2 software package.

5.2 Materials and Methods

5.2.1 Strains of *E. histolytica* and small RNA preparation

Four strains of *E. histolytica* used in the RNA-Seq experiment in Chapter 2 were revived from cryopreserved stocks and maintained in axenic LYI-S-2 media as described in the 'Materials and Methods' of Chapter 2. Briefly, after 60 hrs of culture, the mid-log phase trophozoites were harvested and washed in PBS solution. Then, the fresh trophozoites were immediately used for small RNA extraction using the mirVana™ small RNA isolation kit (Life Technologies, USA). These sRNA enriched samples were then verified qualitatively using an Agilent 2100 Bioanalyser with the Small RNA chip (Agilent Technologies) and quantitatively using the Qubit® fluorometric assay (Invitrogen). Samples of qualified sRNA were stored at -80 ^{0}C until used for small RNA library construction.

5.2.2 Small RNA library construction, size-selection and single-end sequencing

Small RNA libraries were constructed following the protocol of the NEBnext® multiplex small RNA library preparation kit as described in Figure 5.1, with different NEB small RNA index primers to label each of the four strains. Obtained cDNA libraries were checked for their profiles using an Agilent 2100 Bioanalyser with the High Sensitivity DNA chip (Agilent Technologies). Size selection was then performed at 150 bp using 3% Pippin prep at range 125-160 bp. Size-fractionated samples were purified using Agencourt® AMPure XP magnetic beads (Beckman Coulter, USA) with final elution in 20 µl of TE buffer. Their sizes were confirmed again by the High Sensitivity DNA chip as shown in Figure 5.2. Size-selected adaptor-ligated cDNA samples of all four strains were pooled together for single-end sequencing (1x50 bp) on the Illumina MiSeq platform. Sequencing was conducted at the Centre for Genomic Research (CGR), University of Liverpool.

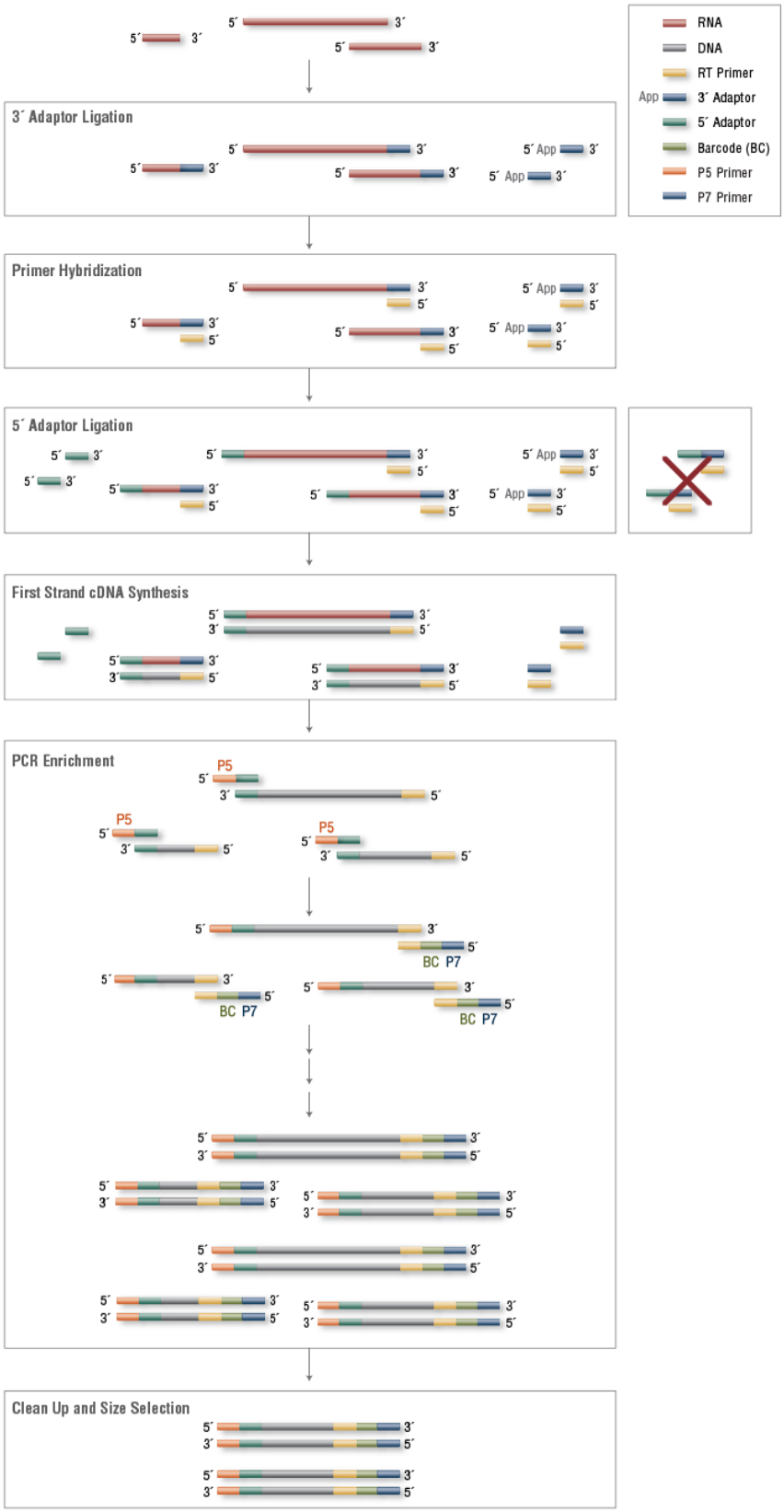
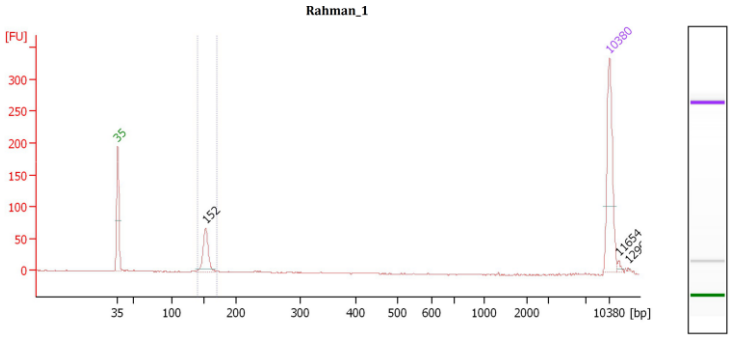


Figure 5.1: Principles and procedures of the NEBNext® multiplex small RNA library preparation for Illumina sequencing in this study (available online at https://www.neb.com).

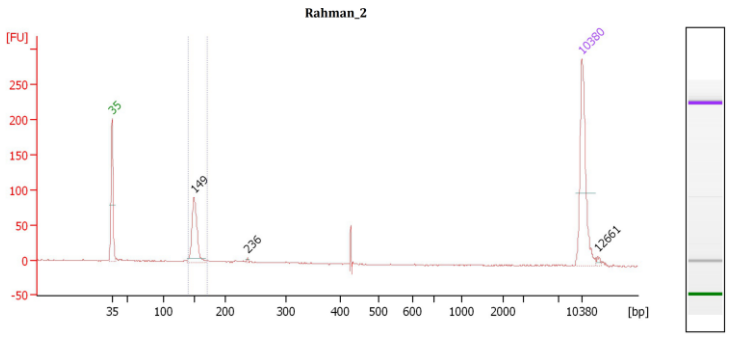
Briefly, freshly extracted sRNA sample was ligated with the 3' SR adaptor. After the 3' ligation reaction, adaptor-ligated sRNAs were hybridised with the reverse transcription primer and followed by the 5' adaptor ligation catalysed by T4 RNA Ligase 1.

Then, 5' and 3' adaptor-ligated sRNA annealed with the reverse transcription primer was used as a template for the first strand cDNA synthesis by the reaction of SuperScript III reverse transcriptase.

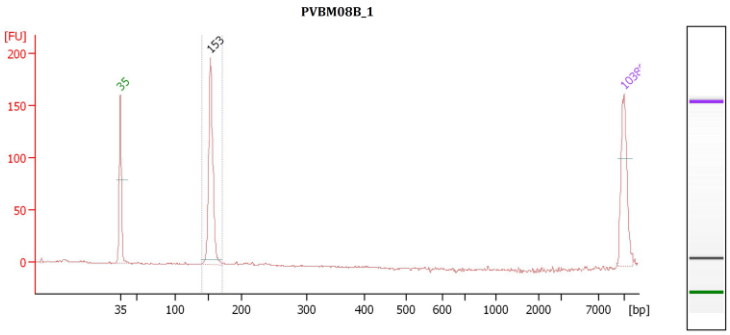
Finally, the reverse transcription reaction mix containing adaptor-ligated cDNAs of each library sample was enriched by PCR amplification using the Multiplex SR primer for Illumina platform and the Index (X) primer specific for each library. The amplified cDNA solution was purified and subsequently size-fractionated by 3% Agarose Pippin Prep.


Rahman_1
[FU]
35
152
10380
11654
[bp]

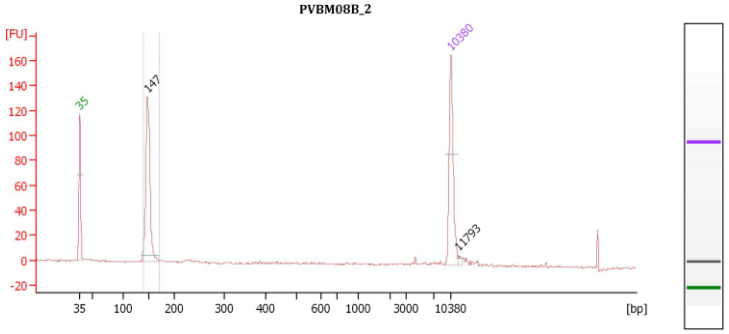
A


Rahman_2
[FU]
35
149
236
10380
12661
[bp]

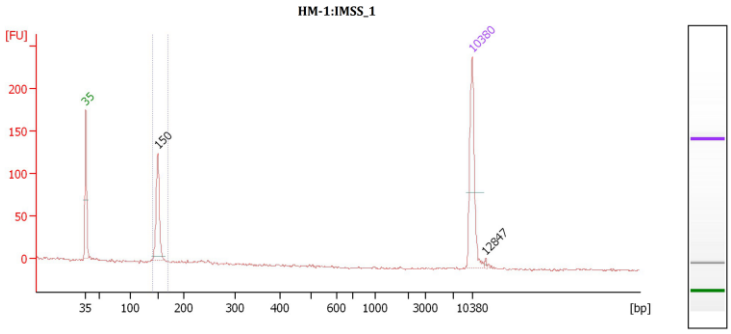
B


PVBM08B_1
[FU]
35
153
[bp]

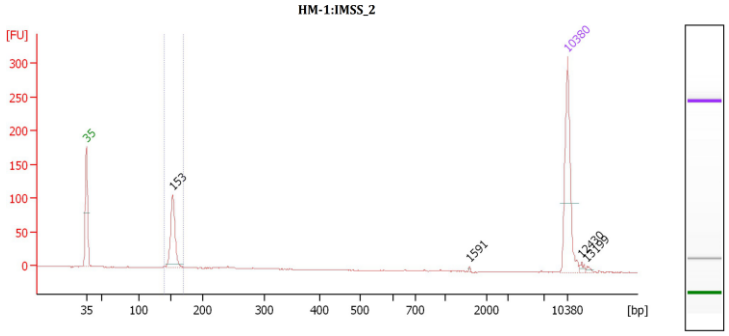
C


PVBM08B_2
[FU]
35
147
10380
11793
[bp]

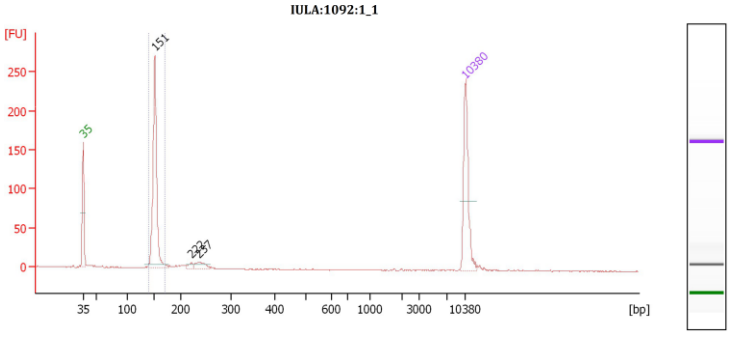
D


HM-1:IMSS_1
[FU]
35
150
10380
12847
[bp]

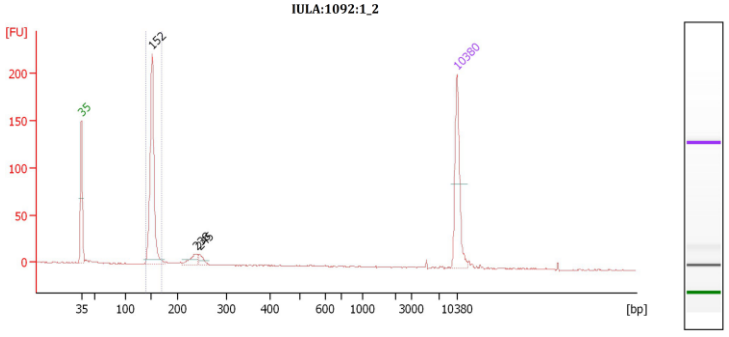
E


HM-1:IMSS_2
[FU]
35
153
1591
10380
[bp]

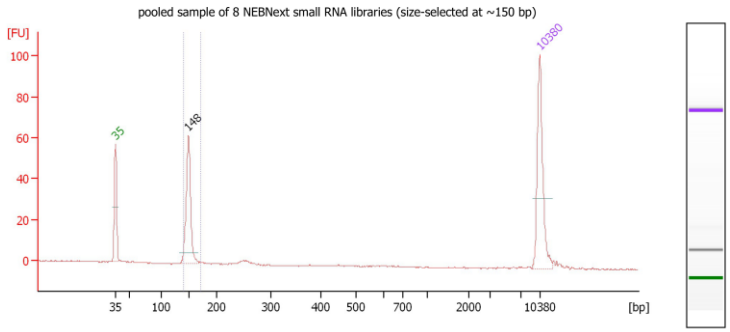
F


IULA:1092:1_1
[FU]
35
151
10380
[bp]

G


IULA:1092:1_2
[FU]
35
152
10380
[bp]

H



I

Figure 5.2: The peak of cDNAs at approximately 150 bp in each sRNA library after the size selection by 3% Agarose Pippin Prep and in final pooled sample of all sRNA libraries. The peaks of approximately 150 bp in all sRNA libraries (graph A-I) are expected to contain adaptor-ligated miRNAs as recommended by the NEB protocol. The graph A-I represent the following sRNA libraries: A = Rahman_01 sRNA library; B = Rahman_02 sRNA library; C = PVBM08B_01 sRNA library; D = PVBM08B_02 sRNA library; E = HM-1:IMSS_01 sRNA library; F = HM-1:IMSS_02 sRNA library; G = IULA:1092:1_01 sRNA library; H = IULA:1092:1_02 sRNA library; I = the final pooled sRNA library sample.

5.2.3 Bioinformatics Pipeline

I. Read processing and quality assessment of the raw sequence data

Raw sequences were obtained in the form of Fastq formatted files. The 3' ends of reads matching adaptor sequences were trimmed using Cutadapt 1.1. Trimming by Sickle version 1.2 with a minimum window quality score of 20 was also done to remove low quality sequence. Reads with a length less than 10 bp after trimming were removed. The total number of raw reads as well as the percentage of single-end trimmed reads were summarised in Figure 5.3 and Table 5.1, respectively. Read lengths after removing adaptor and low quality base in all library samples were illustrated in Figures 5.4-5.8.

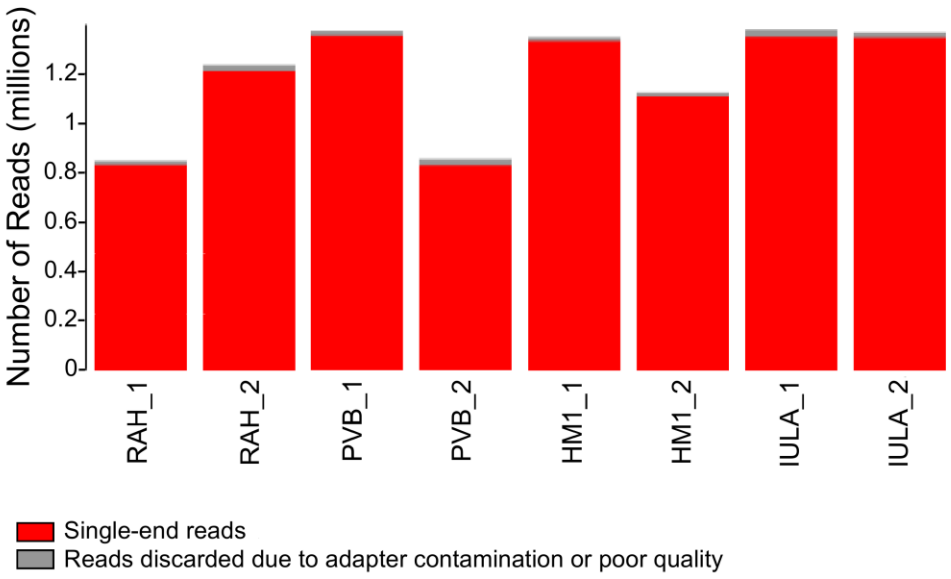


Figure 5.3: The total number of short reads in millions retrieved from each library of the four strains.

Table 5.1: Summary of short sequence read data before and after adapter removal and low Phred score trimming.

Sample	Raw reads	Trimmed R1 reads
Rahman_1	838,812	822,633 (98.07%)
Rahman_2	1,226,260	1,200,153 (97.87%)
PVBM08B_1	1,362,185	1,342,943 (98.59%)
PVBM08B_2	845,170	819,213 (96.93%)
HM-1:IMSS_1	1,337,752	1,322,571 (98.87%)
HM-1:IMSS_2	1,115,217	1,099,742 (98.61%)
IULA:1092:1_1	1,368,808	1,339,467 (97.86%)
IULA_1092:1_2	1,357,693	1,336,227 (98.42%)

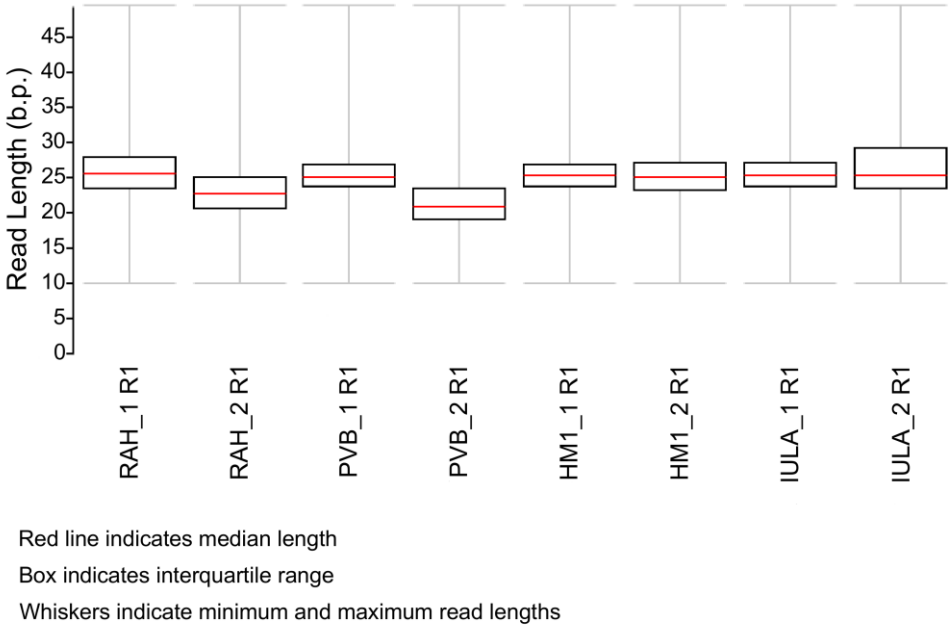


Figure 5.4: Read length distributions after adaptor and low base quality trimming. Only forward unpaired read is represented as R1 read.

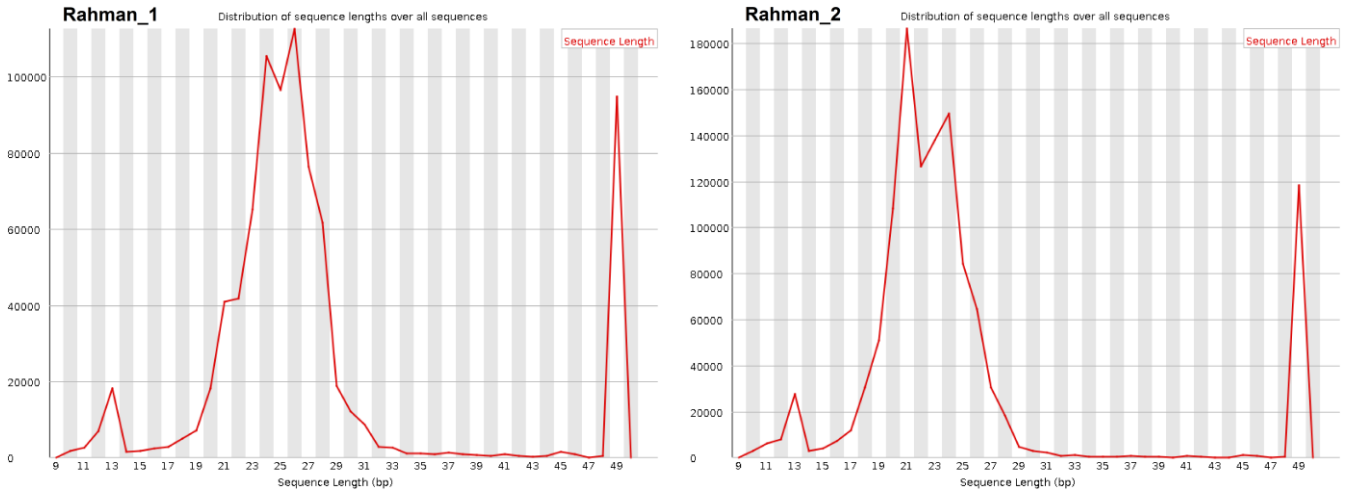


Figure 5.5: **Sequence length distribution of adaptor-trimmed cDNAs in two Rahman biological replicates.** The peaks of sequence length are at 23-28 nt and 20-24 nt for the 1st and 2nd replicates, respectively.

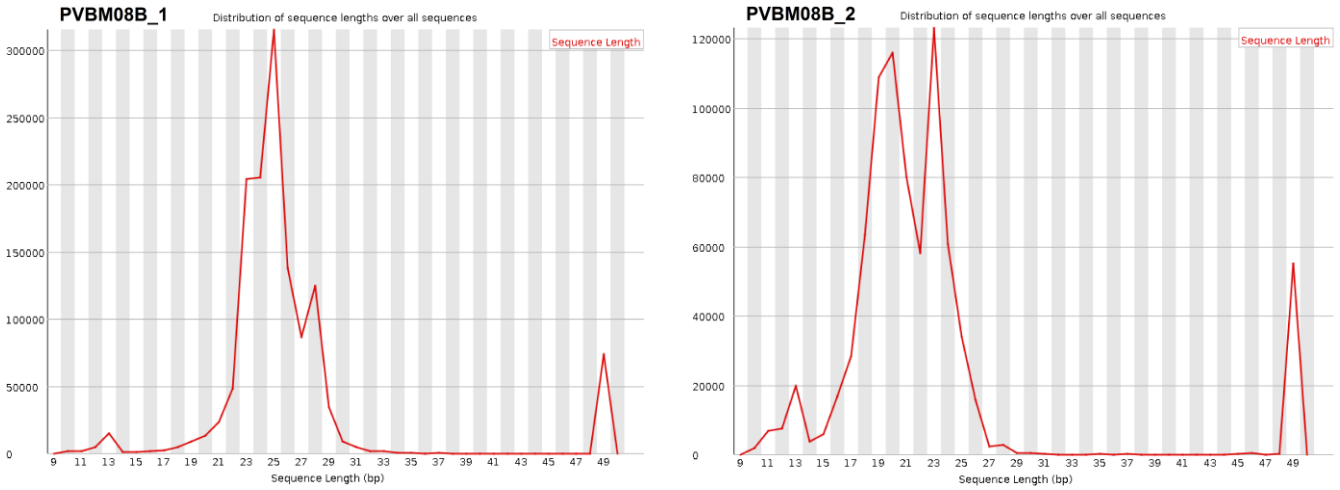


Figure 5.6: **Sequence length distribution of adaptor-trimmed cDNAs in two PVBM08B biological replicates.** The peaks of sequence length are at 23-26 nt and 18-24 nt for the 1st and 2nd replicates, respectively.

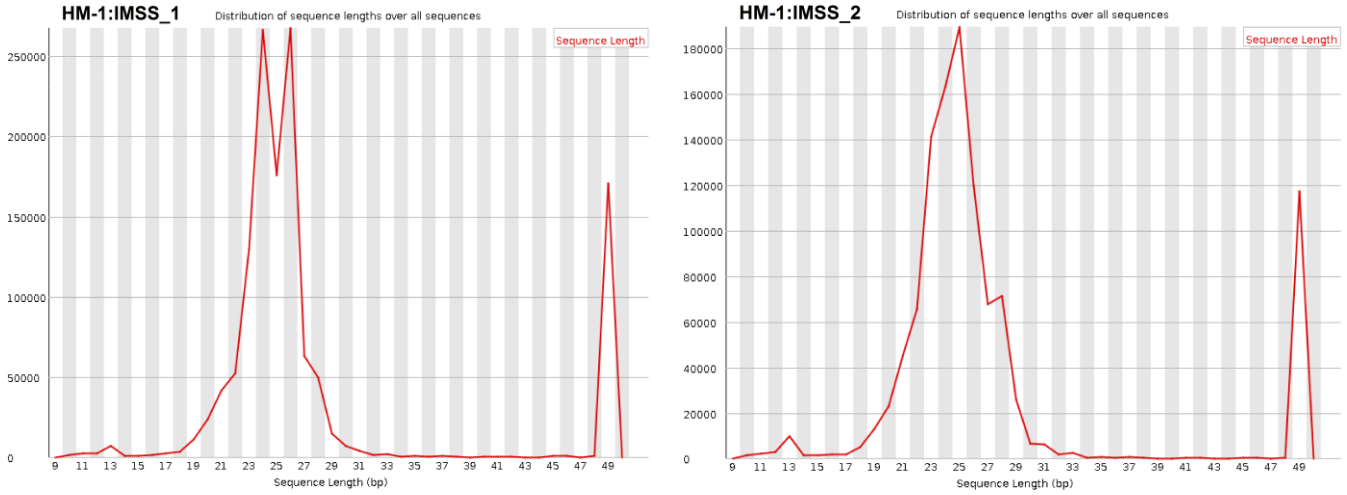


Figure 5.7: **Sequence length distribution of adaptor-trimmed cDNAs in two HM-1:IMSS biological replicates.** The peaks of sequence length are at 23-26 nt and 23-26 nt for the 1st and 2nd replicates, respectively.

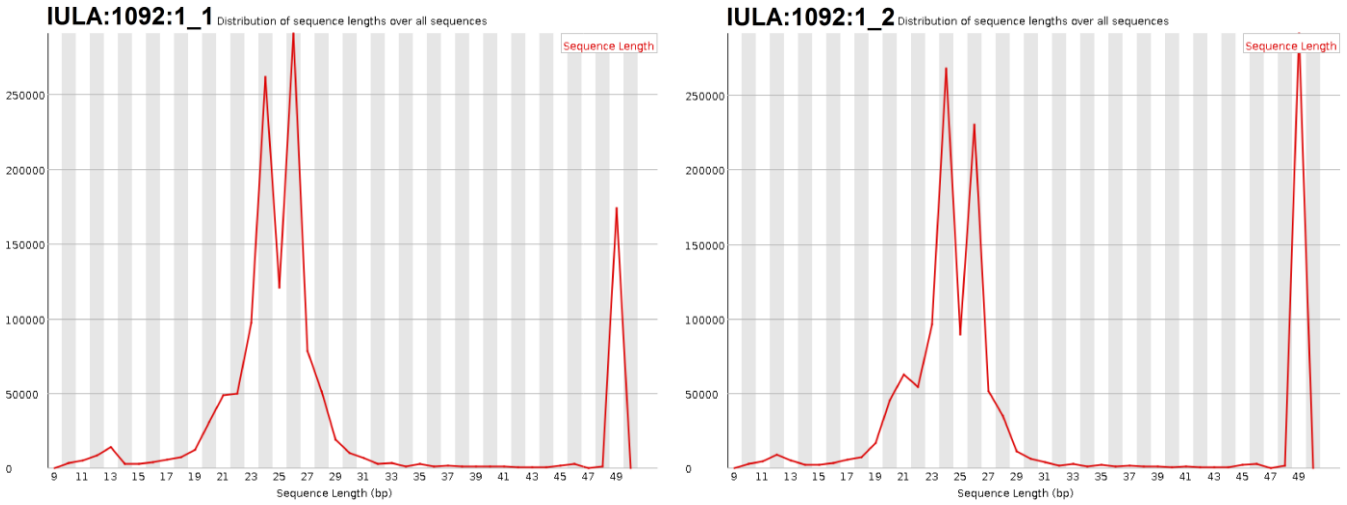


Figure 5.8: **Sequence length distribution of adaptor-trimmed cDNAs in two IULA:1092:1 biological replicates.** The peaks of sequence length are at 23-26 nt and 23-26 nt for the 1st and 2nd replicates, respectively.

II. Mapping of reads to the reference genome sequence and statistical testing for difference between strains

Bowtie2 version 2.2.2 (http://bowtie-bio.sourceforge.net/bowtie2/) was used to map trimmed reads to the *E. histolytica* HM-1:IMSS reference sequence (release 2.0, http://AmoebaDB.org/common/downloads/release2.0/EhistolyticaHM1IMSS/fasta/data/AmoebaDB-2.0_EhistolyticaHM1IMSS_Genome.fasta) [296]. The HM-1:IMSS genome annotation file (release 2.0, AmoebaDB-2.0_EhistolyticaHM1IMSS.gff file), indicating the locations of 8,333 genes in the genome, was used to count reads aligned to each gene [26]. The number and percentage of total short read mapping and uniquely mapped reads are listed in Table 5.2.

Aligned read counts were sorted into sense and antisense orientation using HTSeq-count (release 0.6.1, http://www-huber.embl.de/users/anders/HTSeq/doc/count.html) with the following parameters: **-m** <mode> intersection-strict; **-i** <id attribute> Parent; **-t** <type> exon; **-s** <stranded> yes/no: for sense and antisense direction, respectively. The percentage of genes with five different ranges of normalised mapped antisense sRNA transcripts (antisense sRNA reads per kilobase of exon per million of total mapped reads) in all four strains were summarised in Table 5.4 and Figure 5.10. Only antisense sRNA reads for each gene obtained from the HTSeq-count data would be used for statistical testing for difference between strains by edgeR analysis [115].

Then, genes with no mapped antisense sRNA in all library samples were removed and the subset of only genes with mapped antisense sRNAs in each library were analysed for both 'within-group' and 'between-group' variations as plotted in Figures 5.11 and 5.12, respectively. Also, a sample correlation heatmap was done using Pearson's correlation coefficients (r) to reveal transcriptomic variability within a sample group and between different groups as depicted in Figure 5.13. To assess the overall variation among all 8 sRNA library samples, the $\log_2$-transformed values of raw antisense sRNA reads in all samples were applied for the principal component analysis as plotted in Figure 5.14.

By fitting to the NB model, the dispersion plot was constructed to calculate the common, trended and tagwise dispersions as shown in Figure 5.15. The likelihood ratio (LR) test was applied to determine the difference between $\log_2$FC values of two contrasting strains [120]. Smear plots were drawn to unveil the relationship of the fold change differences ($\log_2$FC) and the average antisense sRNA levels ($\log_2$CPM) for each contrast as demonstrated in Figure 5.16. The distribution of *P*-values for each contrast was shown in Figure 5.17. Statistical significance was indicated when an FDR-adjusted *P*-value less than

0.05. Also, the number of sRNA target genes showing significant differences in mapped antisense sRNA levels in each contrast was summarised in Table 5.5.

To demonstrate the correlation pattern between target gene expression and sRNA abundance in a specific orientation, expression levels of target genes which have normalised antisense sRNAs greater than 50 reads per kilobase of exon per million of total mapped reads were plotted individually against their mapped sRNA abundance for each strain as shown in Figures 5.18A, 5.19A, 5.20A and 5.21A for antisense direction and in Figures 5.18B, 5.19B, 5.20B and 5.21B for sense direction. The 20 most prevalent functionally annotated genes which have normalised antisense sRNAs > 50 were ranked in order for each strain as summarised in Tables 5.6-5.9.

Also, Venn diagrams were constructed to show the number of sRNA target genes which exhibited significantly higher levels of mapped antisense sRNAs between Rahman and other three strains as depicted in Figures 5.22-5.25 and to assess the contribution of sRNAs to their differential gene expression as depicted in Figure 5.26. The sRNA target genes which have markedly high antisense sRNA levels in Rahman were summarised in Table 5.10. Then, from Table 5.10, the 1st subset of sRNA target genes with higher mRNA expression levels in all three virulent strains and the 2nd subset of sRNA target genes with no differential mRNA expression among the four strains were detailed in Tables 5.11 and 5.12, respectively. Finally, comparison to whole transcriptomic data in the same strain was visualised by the Integrative Genomics Viewer (IGV) to explore the biological implications as illustrated in Figures 5.27-5.30.

Table 5.2: **Summary of number and percentage of total and uniquely short read alignments to the *E. histolytica* HM-1:IMSS reference genome using Bowtie2 software version 2.2.2.**

Strain_Replicate	Number of total short reads generated	Number of total short reads mapped to reference	Percentage of total short read mapping	Number of uniquely mapped reads	Percentage of uniquely mapped reads
Rahman_1	822,633	436,209	53.03%	332,159	40.38%
Rahman_2	1,200,153	659,060	54.91%	485,931	40.49%
PVBM08B_1	1,342,943	1,041,751	77.57%	946,464	70.48%
PVBM08B_2	819,213	515,166	62.89%	447,296	54.60%
HM-1:IMSS_1	1,322,571	537,654	40.65%	477,431	36.10%
HM-1:IMSS_2	1,099,742	749,493	68.15%	612,707	55.71%
IULA:1092:1_1	1,339,467	267,325	19.96%	203,618	15.20%
IULA:1092:1_2	1,336,227	349,637	26.17%	294,541	22.04%

III. Novel putative miRNA prediction by the the miRDeep2 software

In order to explore the existence of miRNAs in the *E. histolytica* transcriptome, the miRDeep2 software (https://www.mdc-berlin.de/8551903/en/) was used to process and predict the novel putative miRNA candidates from the small RNA sequencing data with high accuracy, based on the miRNA biogenesis as described in Figure 5.9 [297]. Two perl scripts. collapse_reads_md.pl and mapper.pl, were applied to process the short read sequence data before analysing and scoring by the miRDeep2 core algorithm with miRDeep2.pl, as detailed in Table 5.3.

Briefly, all of the size-selected sRNA read sequences in each library sample obtained in the FASTA format were collapsed for their identical read sequences and summarised for the number of reads for each unique sequence using the collapse_reads_md.pl script. Then, the collapsed read file for each library was aligned against the same *E. histolytica* HM-1:IMSS reference genome sequence used for the whole transcriptomic mapping in Chapter 2, using the mapper.pl script. Finally, the output files from the previous two steps were identified for both known and novel miRNA candidates in comparison to mature miRNA and stem-loop pre-miRNA sequences of the free-living sister species, *Dictyostelium discoideum* (available at http://www.mirbase.org/ftp.shtml) using the miRDeep2.pl script [298]. Novel miRNA precursors with multiple loops and/or energetic instability with non-significant randfold *P*-value were eliminated [299].

The details of novel predicted miRNA sequences with estimated probability of true positives, significant randfold *P*-values and genomic coordinates were summarised in Table 5.13. Predicted secondary structures of potential miRNA precursors with the relative nucleotide positions of the mature miRNA strand, star sequence and loop portion as well as the number of counts for each portion were demonstrated in Figures 5.31-5.34.

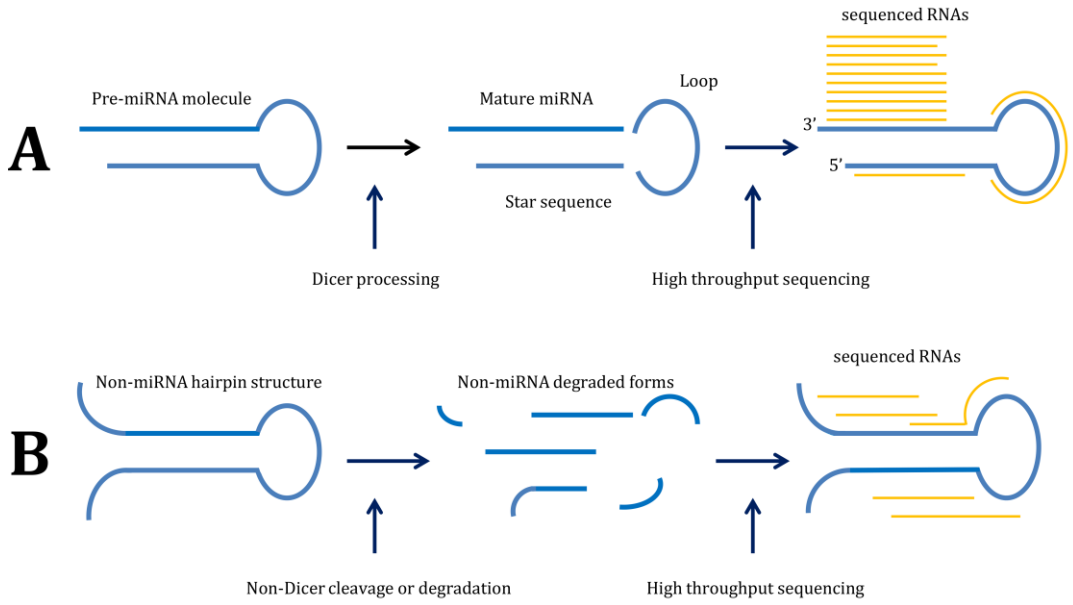


Figure 5.9: Principles of putative novel miRNA detection based on the miRNA biogenesis. Typically, the hairpin structure of the primary miRNA transcript is processed into a stable stem-loop pre-miRNA molecule in the nucleus by the microprocessor complex (Drosha/Pasha) and then transported to the cytosol via the exportin transporter (A) [291]. Then, such pre-miRNA precursor is further cleaved by the ribonuclease III enzyme Dicer to produce the mature miRNA, more abundant than loop and star sequences. Different from the pre-miRNA, non-miRNA transcribed RNA hairpins can provide shorter degraded forms without specific pattern by non-Dicer processing or random degradation (B). Based on the characteristic miRNA biogenesis, the sequencing reads generated from real miRNAs will show high read frequency in mapping to their pre-miRNA precursor, corresponding to its Dicer processing. The figure is redrawn and adapted from Friedländer MR, *et al.*, 2008 [297].

Table 5.3: Summary of perl scripts and their functions in the miRDeep2 analysis.

Perl scripts	Input	Output	Function
collapse_reads_md.pl	reads.fa	reads_collapsed.fa	Collapse all reads with the same sequence to show only once in fasta file. The number of reads for each unique sequence will be indicated in form of '_uniqueNo_xNum'. For example, 1_x20 represents sequence no.1 with twenty reads.
mapper.pl	reads_collapsed.fa	reads_col_vs_genome.arf	Process and map reads to the reference genomic sequence. Option used in this chapter: -p indexed genome : map to indexed reference genome -t arf file : provide output of mapped reads in an arf file
miRDeep2.pl	reads_collapsed.fa genome.fa reads_col_vs_genome.arf mature miRNA.fa* other_mature miRNA.fa* precursor_miRNA.fa* N.B. '*' = optional	Report.log file consists of 1. a spreadsheet 2. a html file	This perl script possesses the 'wrapper function' to identify both known and novel miRNAs from deep sequencing data. The algorithm provides the results with overall information of the predicted miRNA precursors including predicted structure, minimal free energy and the total confidence score for all parameters as shown in Table 5.13.

5.2.4 Validation of the predicted miRNA candidate using qPCR analysis

The expression of predicted miRNA obtained from the previous miRDeep2 analysis was validated using a quantitative polymerase chain reaction (qPCR) method. In this study, predicted miRNA candidate (miR-Rah1) with consensus mature sequence 5' AGAUGGAUUAGAAAAGACGGUUGU 3' as listed in Table 5.13 was chosen for validation. The miRNA-specific forward primer is identical in nucleotide sequence to the predicted miR-Rah1 as detailed above. Briefly, 1 ng of sRNA-enriched samples previously extracted from each *E. histolytica* strain were directly tagged and reverse transcribed to cDNA using the QuantiMir™ Reverse Transcription kit (System Biosciences, USA). Then, obtained QauntiMir™ cDNAs were analysed using Power SYBR Green qPCR mastermix (Applied Biosystems, USA) with universal reverse and miRNA-specific forward primers. qPCR reaction was run in triplicate on the LightCycler® 480 Instrument II (Roche Life Science, USA) with the following conditions: 50 ^{0}C for 2 min, 95 ^{0}C for 10 min and 40 cycles of 95 ^{0}C for 15 sec and 60 ^{0}C for 1 min. Melting curve analysis was conducted after finishing the amplification step. The qPCR amplification curve and the details of crossing point in each RNA sample were demonstrated in Figure 5.35 and Table 5.14, respectively.

5.3 Results and Discussion

5.3.1 Small RNA transcriptome profiling of the four *E. histolytica* strains from axenic culture

To test the hypothesis that sRNAs, and/or miRNAs play a role in post-transcriptional gene regulation in *E. histolytica*, these RNA populations were sequenced using the next-generation sequencing technology. After adaptor removal and quality trimming, the obtained short sequence reads in each sample library were verified for their size distribution as demonstrated in Figures 5.5-5.8 for Rahman, PVBM08B, HM-1:IMSS and IULA:1092:1, respectively. It shows that the majority of adaptor-removed cDNAs range from 20 to 28 nt in Rahman, 18 to 26 nt in PVBM08B and 23-26 nt in HM-1:IMSS and IULA:1092:1, indicating that these cDNA populations are most likely to contain miRNA-derived cDNA molecules. However, a small peak of 13 nt was also seen in all libraries that might be partial degradation fragments of miRNAs generated during library preparation.

It was not possible to calculate the percentage of small RNAs with different RPKMs or to correlate the small RNA data with the genomic data because there was no available miRNA database in *E. histolytica* to use as a reference for miRNA identification. However, after using Bowtie2 to map short sequence reads to the HM-1:IMSS reference genome, sRNAs mapped to each gene were counted by the HTSeq-count software and sorted individually into sense and antisense orientation. Sense sRNAs could be degraded mRNA transcripts, therefore only antisense sRNA reads were analysed for their possible biological roles in gene regulation.

As shown in Table 5.4 and Figure 5.10, it is noticeable that across the four strains, most *E. histolytica* genes have no mapped antisense sRNA transcript and only a small fraction of genes (11.35-28.99 %) showed normalised antisense sRNAs greater than 50 reads, possibly suggesting that antisense sRNAs might play a role in regulating a particular set of genes.

Differential expression of sRNAs was analysed using the edgeR package for difference in the number of antisense sRNAs mapped to a particular target gene between strains. In contrast to the inter-library transcriptomic variation, both 'within-group' and 'between-group' variations of sRNA transcriptomes among library samples are more pronounced as shown in Figures 5.11 and 5.12, respectively. However, sRNA transcriptomes within the same strain are less variable than those between different strains, indicating that inter-strain differences are greater than biological variation between individuals of the

same strain. The Pearson's correlation-based heatmap in Figure 5.13 reveals a wide range of the Pearson's correlation coefficients (*r*) from 0.487 to 1.000, reflecting larger differences in antisense sRNA transcriptomic profiles among the parasite strains compared to the inter-library transcriptomic variation in Chapter 2.

Consistently, the principal component analysis shows the clear separation of all sRNA libraries among the four strains, indicating that the inter-strain differences are strong. As demonstrated in Figure 5.14, the large intra-strain biological variation could be seen in Rahman and HM-1:IMSS strains due to the high difference in total HTSeq-count library size between the two replicates of the same strains as represented on the 1st component axis. The second component reveals more intra-strain variation in HM-1:IMSS and PVBM08B, consistent with the poorer correlation between two replicates demonstrated in the pairwise scatterplots in Figure 5.11.

As shown in Figure 5.15, the dispersion plot exhibits high inter-library variation (high tagwise dispersions on the y-axis) especially in a small number of target genes with average high antisense sRNA levels (high $\log_2$CPM on the x-axis). Also, the common dispersion of all sRNA transcriptomes is equal to 0.2796, much higher than that of the whole transcriptomes (0.0095) in Chapter 2. One plausible interpretation is that a high variability of antisense sRNA transcript levels among samples could be found within genes with high tagwise dispersions, implying that such particular target genes were not equally regulated across all strains due to very different levels of antisense sRNAs mapped to a particular gene in each strain. In other words, it possibly suggests that antisense sRNAs are not equally expressed among the strains and there should be a unique set of target genes which are potentially regulated by a different set of antisense sRNAs in each strain.

In contrast to the DGE analysis in Chapter 2, there are a small number of target genes showing significant difference in mapped antisense sRNA levels between two strains of contrast with a FDR corrected *P*-value < 0.05 as summarised in Figures 5.16 and 5.17 and Table 5.5. In agreement with the previous dispersion plot, it seems to be that differences in antisense sRNA levels between strains are present in a unique set of genes in the parasite transcriptome, possibly implying that differential expression of such target genes between strains might be regulated by these antisense sRNAs.

Table 5.4: Categorisation of all 8,333 *E. histolytica* genes into five groups based on their mapped antisense sRNA transcript level. All of 8,333 genes are categorised into 5 groups by different ranges of normalised HTSeq-count (antisense sRNA reads per kilobase of exon per million of total mapped reads) as follows: zero; low = less than 5; moderate = between 5 and 25; high = between 25 and 50; very high = greater than 50, respectively. The number of genes and corresponding percentages in each strain are shown below.

Strain / normalised HTSeq-counts	Rahman		PVBM08B		HM-1:IMSS		IULA:1092:1	
	No. of genes mapped	Percentage	No. of genes mapped	Percentage	No. of genes mapped	Percentage	No. of genes mapped	Percentage
0	2,837	**34.05 %**	5,317	**63.81 %**	3,942	**47.31 %**	4,232	**50.79 %**
< 5	236	**2.83 %**	0	**0 %**	101	**1.21 %**	0	**0 %**
5-25	2,829	**33.95 %**	541	**6.49 %**	1,802	**21.62 %**	516	**6.19 %**
25-50	1,485	**17.82 %**	882	**10.58 %**	1,083	**13.00 %**	1,169	**14.03 %**
> 50	946	**11.35 %**	1,593	**19.12 %**	1,405	**16.86 %**	2,416	**28.99 %**

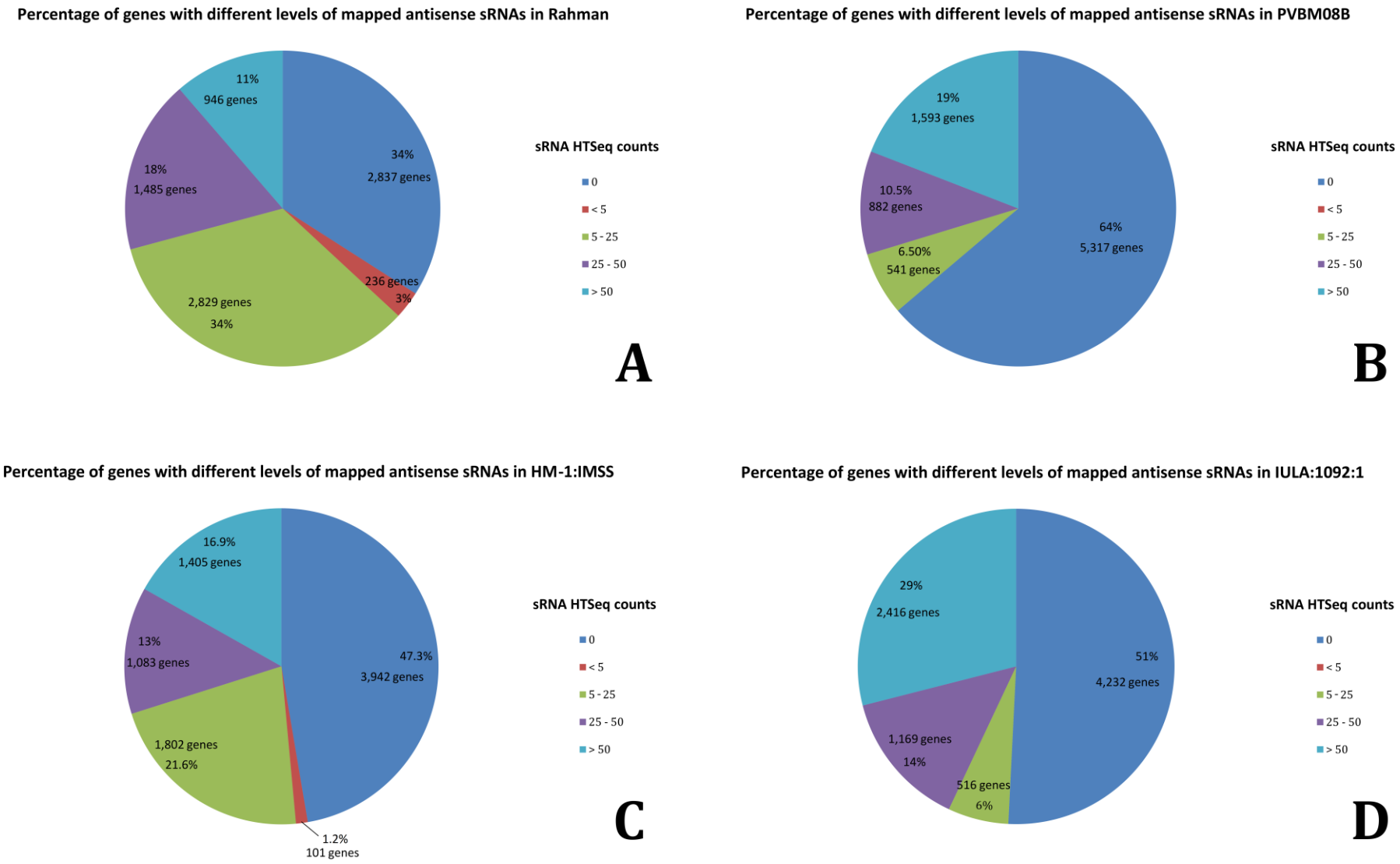


Figure 5.10: Percentage of genes with different antisense sRNA levels in Rahman (A), PVBM08B (B), HM-1:IMSS (C) and IULA:1092:1(D). Most of the genes (34-64%) in all four strains have no mapping with 21-23 nt antisense sRNA molecule as shown above.

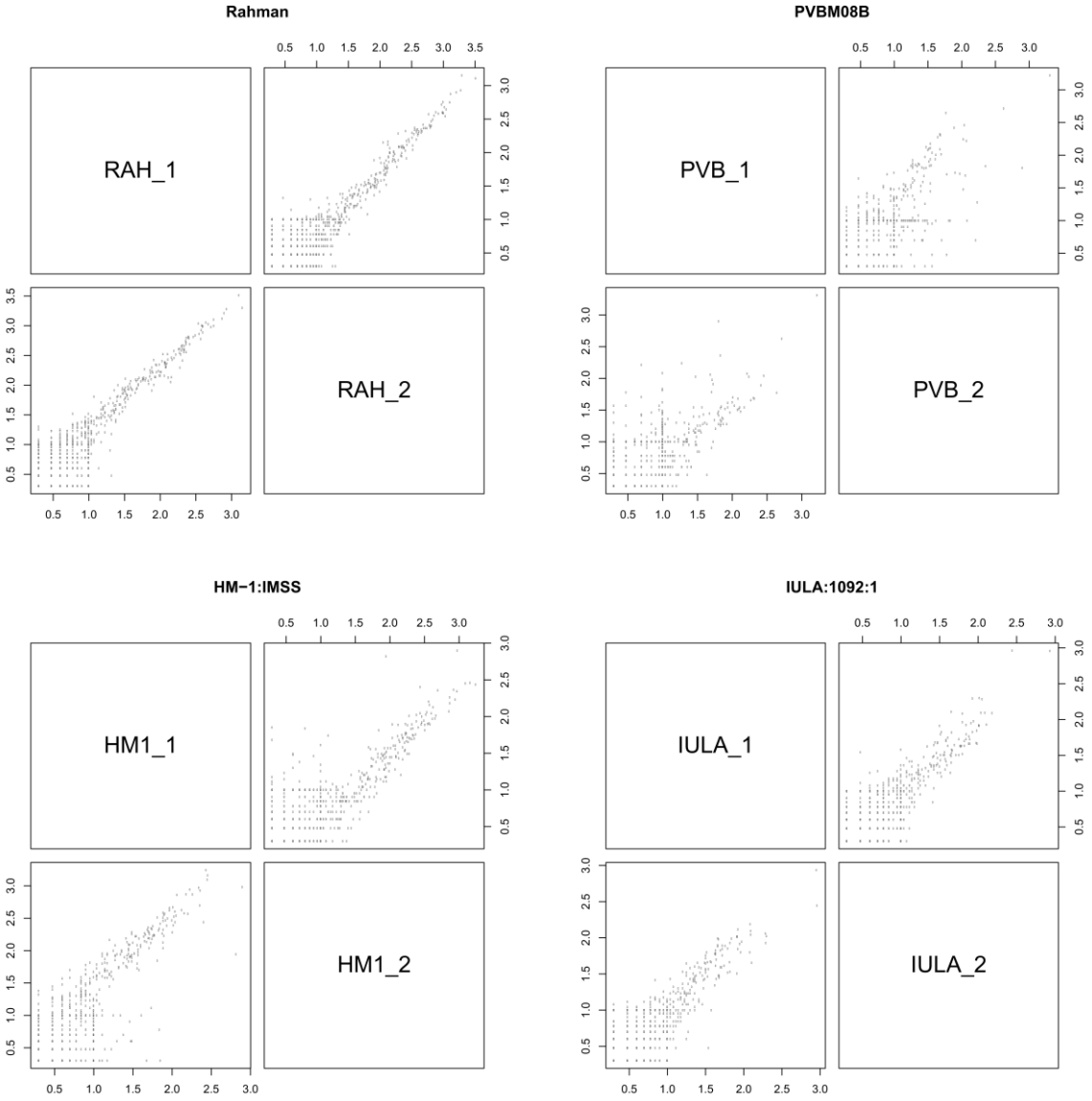


Figure 5.11: **'Within-group' variation of sRNA transcriptomes between two biological replicates in each *E. histolytica* strain.** Both X and Y graph axes represent the logarithm (base 10) of raw antisense sRNA read count per gene in each replicate.

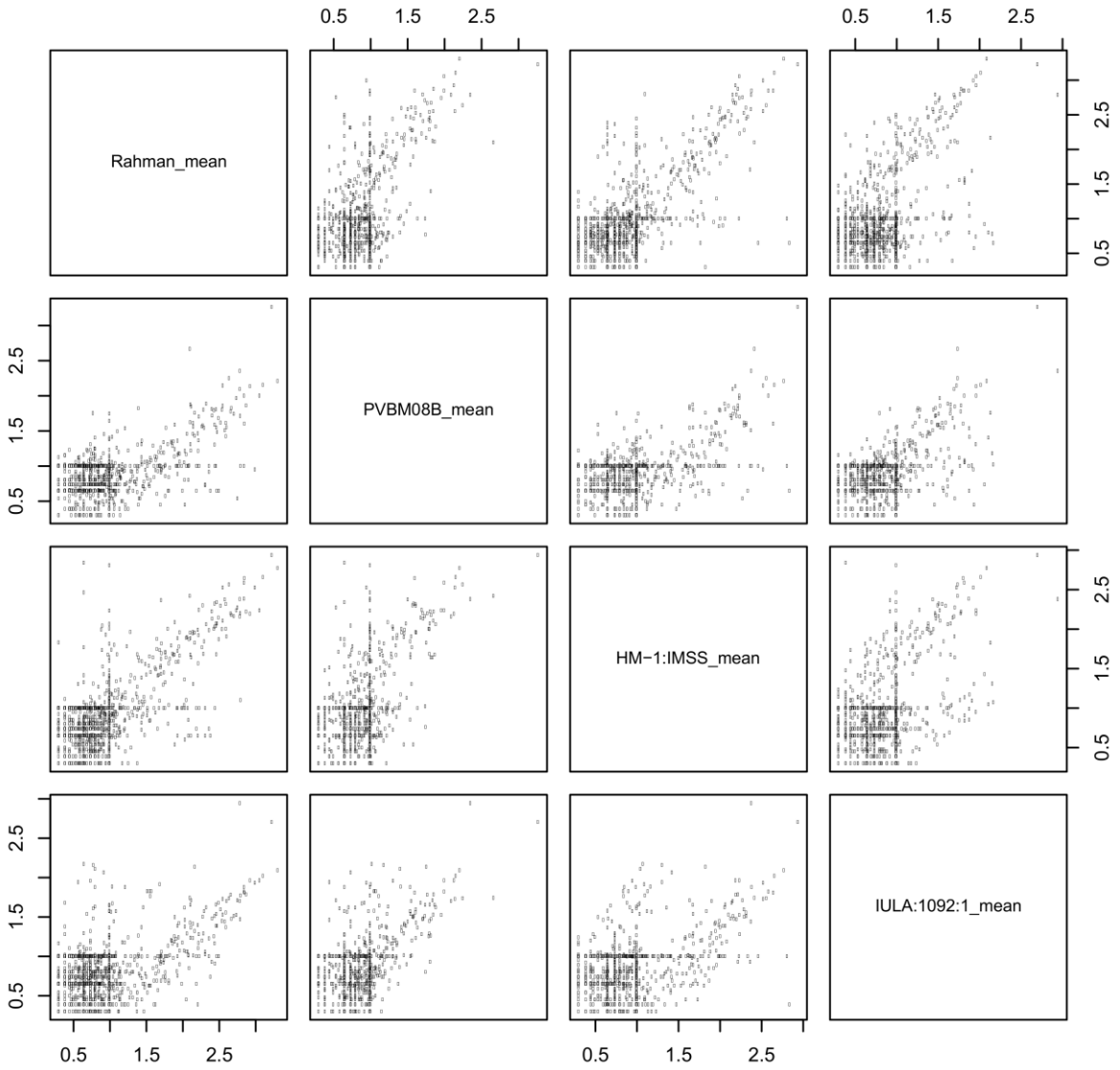


Figure 5.12: **'Between-group' variation of sRNA transcriptomes among the four *E. histolytica* strains.** Both X and Y graph axes represent the logarithm (base 10) of average antisense sRNA read count per mapped gene in each group. In overall, sRNA transcriptomic variations between groups of samples are more obvious than those within the same group previously illustrated in Figure 5.11.

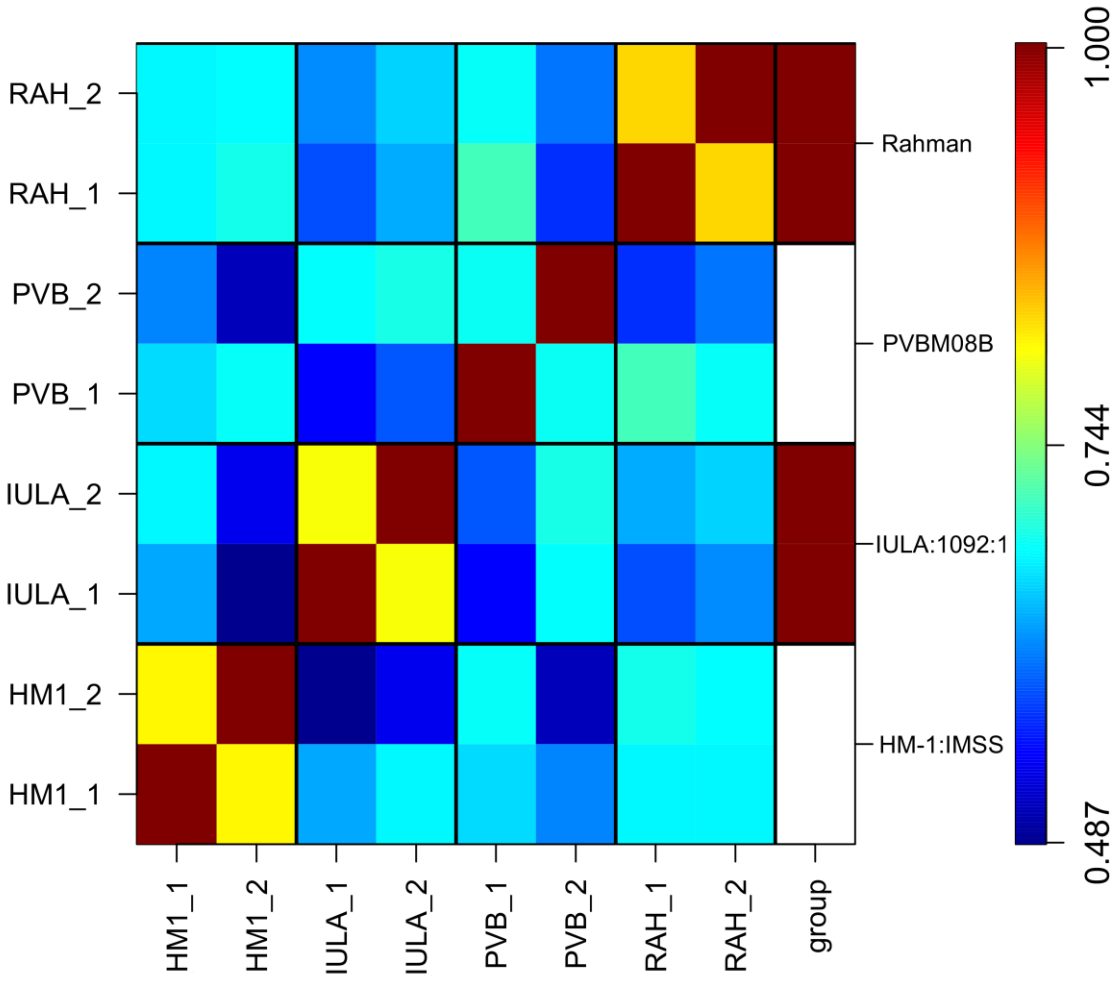


Figure 5.13: **Agglomerative hierarchical clustering of sRNA expression profiles within and among the four strain groups.** The pairwise correlation patterns are shown in 16 clusters between strains and in 4 sub-clusters between two biological replicates. The colour spectrum represents the Pearson's correlation coefficients (*r*) scoring from 0.487 to 1.000. In overall, sRNA transcriptomes within the same strain are less variable than those between different strains. Compared to the inter-library variation of RNA-Seq data in Chapter 2, a wider range of correlation scores reflects larger differences in antisense sRNA transcriptomic profiles among the strains, suggesting that there are different sets of genes mapped to these sRNAs among the strains, implying a regulatory function of these antisense sRNAs.

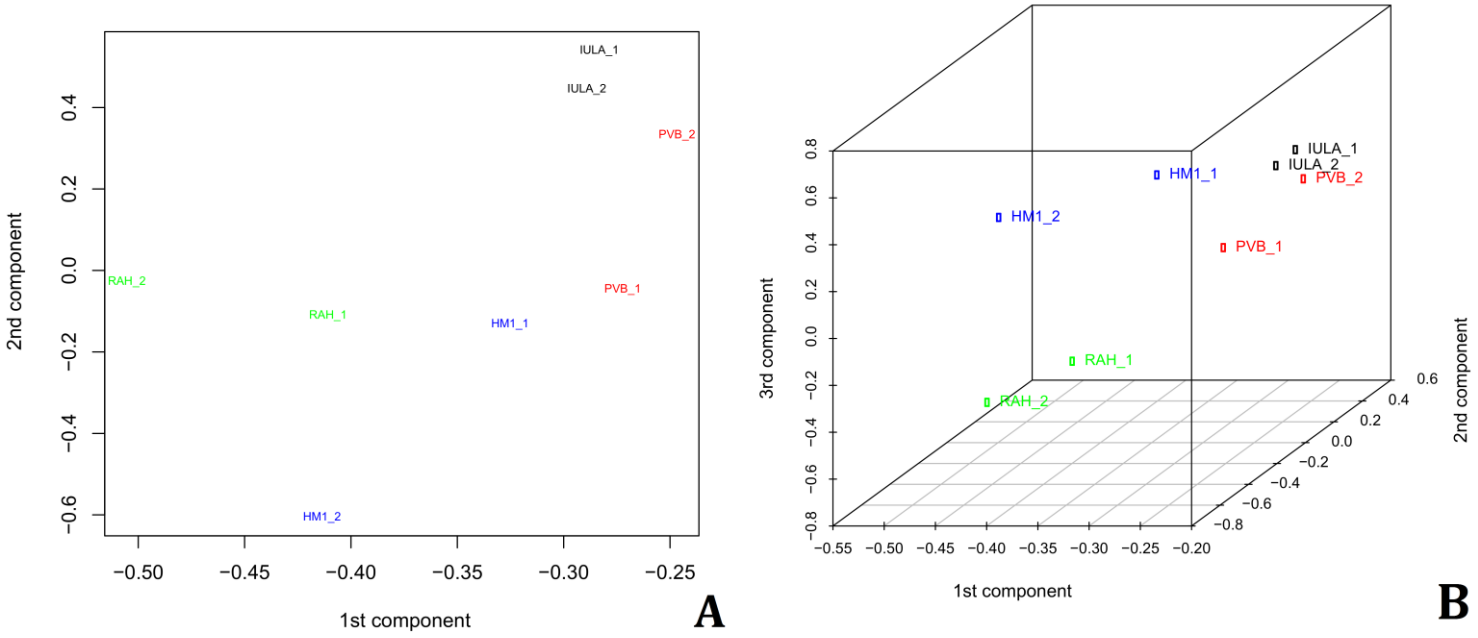


Figure 5.14: **Two and three dimensional principal component analysis of sRNA transcriptomes in the four *E. histolytica* strains.** The $\log_2$ mapped antisense sRNA HTSeq-count data for all 8,333 genes were employed to plot each sRNA library in comparison with all others. The plots show a clear discrimination among the four strains. The 1st component (X %) is dominated by the difference in total library size among replicates. Rahman and HM-1:IMSS have more variable library sizes among replicates and this is seen in the 1st component of variation. The second component shows more separation between HM-1:IMSS and PVBM08B replicates than for the other two strains and reflects the poorer correlation seen in the pairwise scatterplots in Figure 5.11.

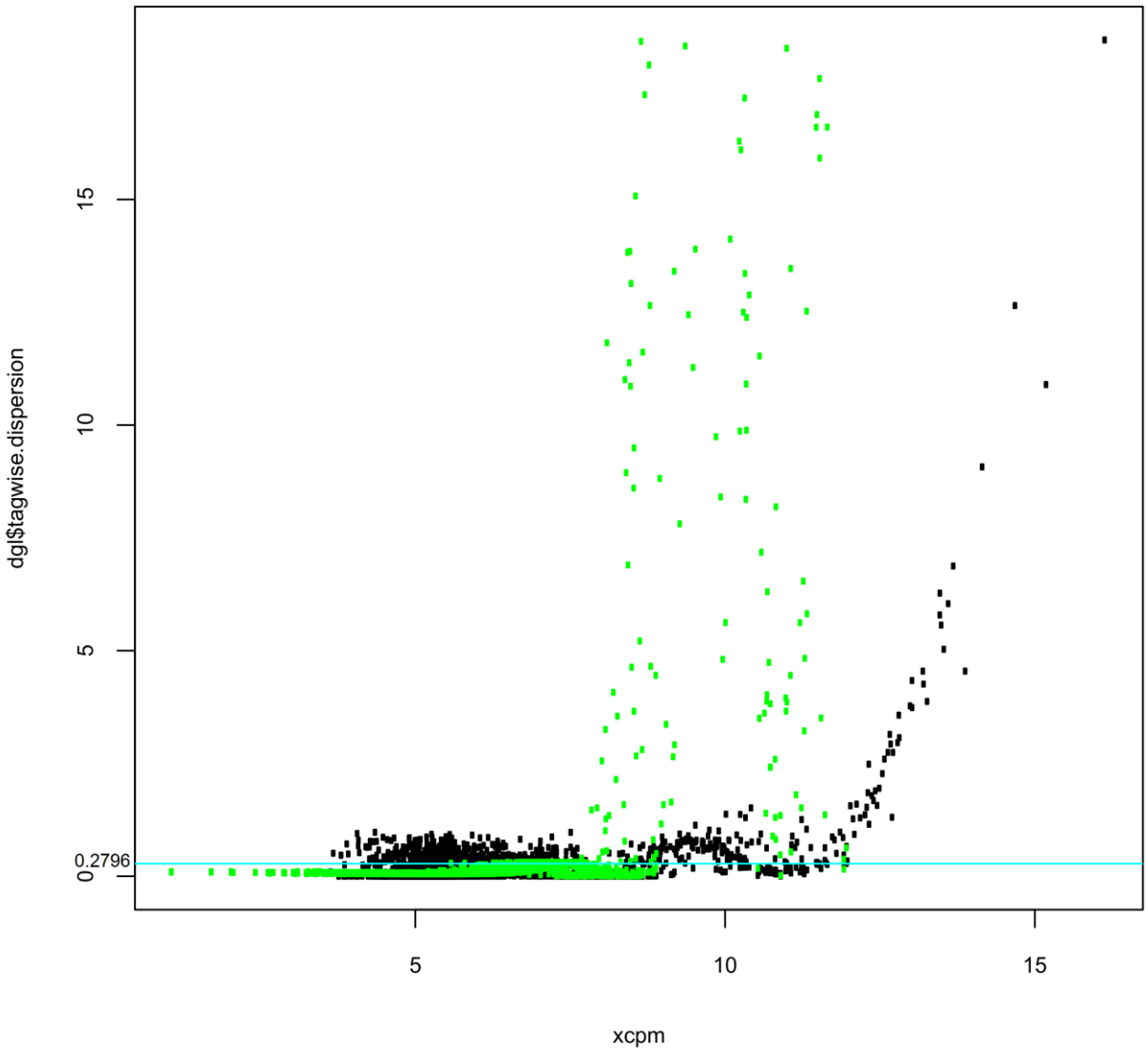


Figure 5.15: **Relationship of inter-library variation for each sRNA target gene and its corresponding abundance (log_2CPM).** The aqua blue horizontal line represents the common dispersion, equal to 0.2796 across all 8 library samples, regardless of gene. The green curve line is the trended dispersion varied by transcript abundance. The black spots show the gene-by-gene (tagwise) dispersions. Interestingly, higher dispersions could be seen in genes with average high levels of antisense sRNA transcripts, indicating a high variability of sRNA transcript levels among samples could be found within such genes.

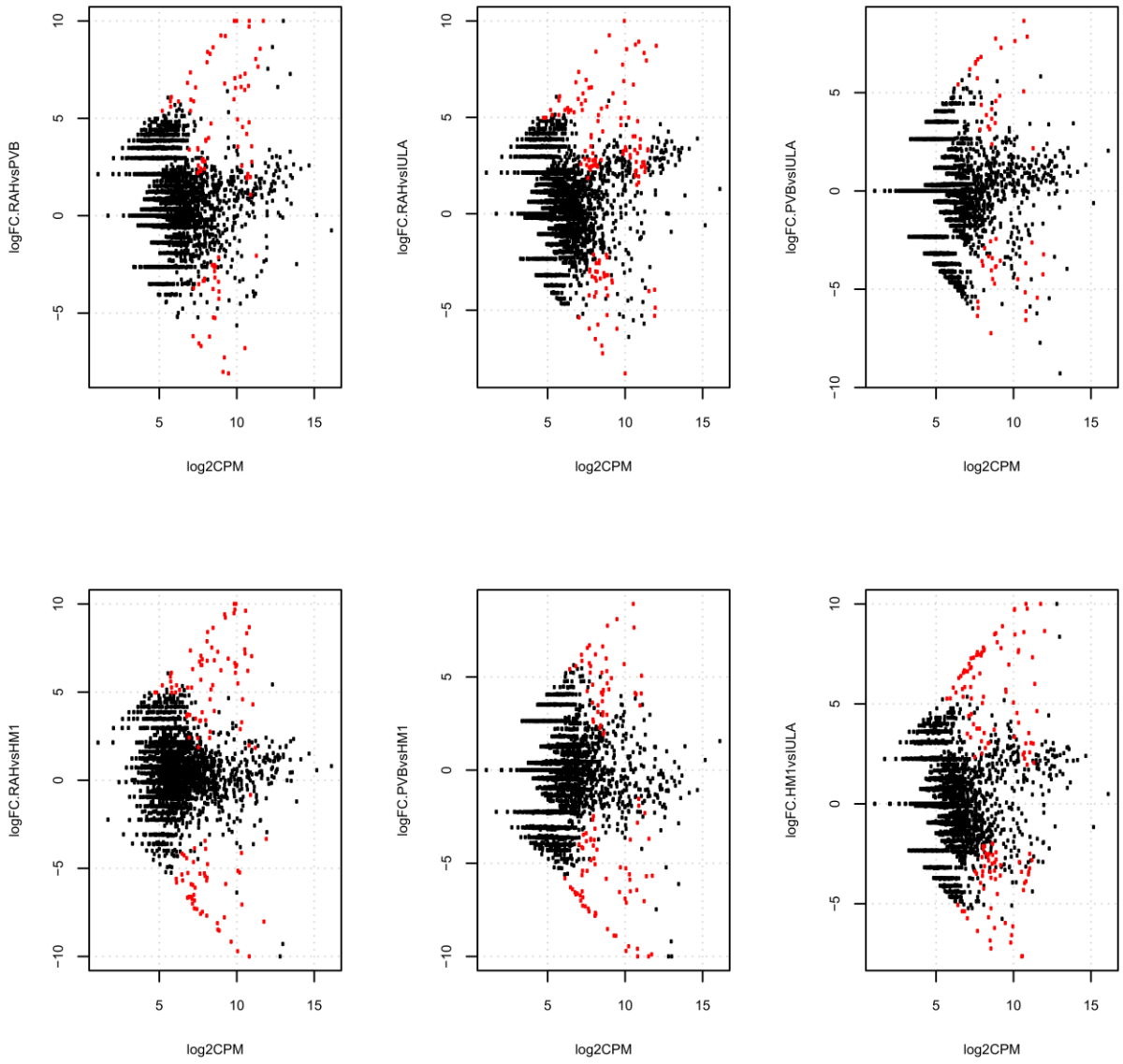


Figure 5.16: **Relationship of the fold change (log_2FC) and the average level of antisense sRNAs, i.e. counts per million mapped reads (log_2CPM), for each contrast pair.** Significant DE genes with FDR-adjusted *P*-value < 0.05 were highlighted in red. Black spots represent no significantly differential expression.

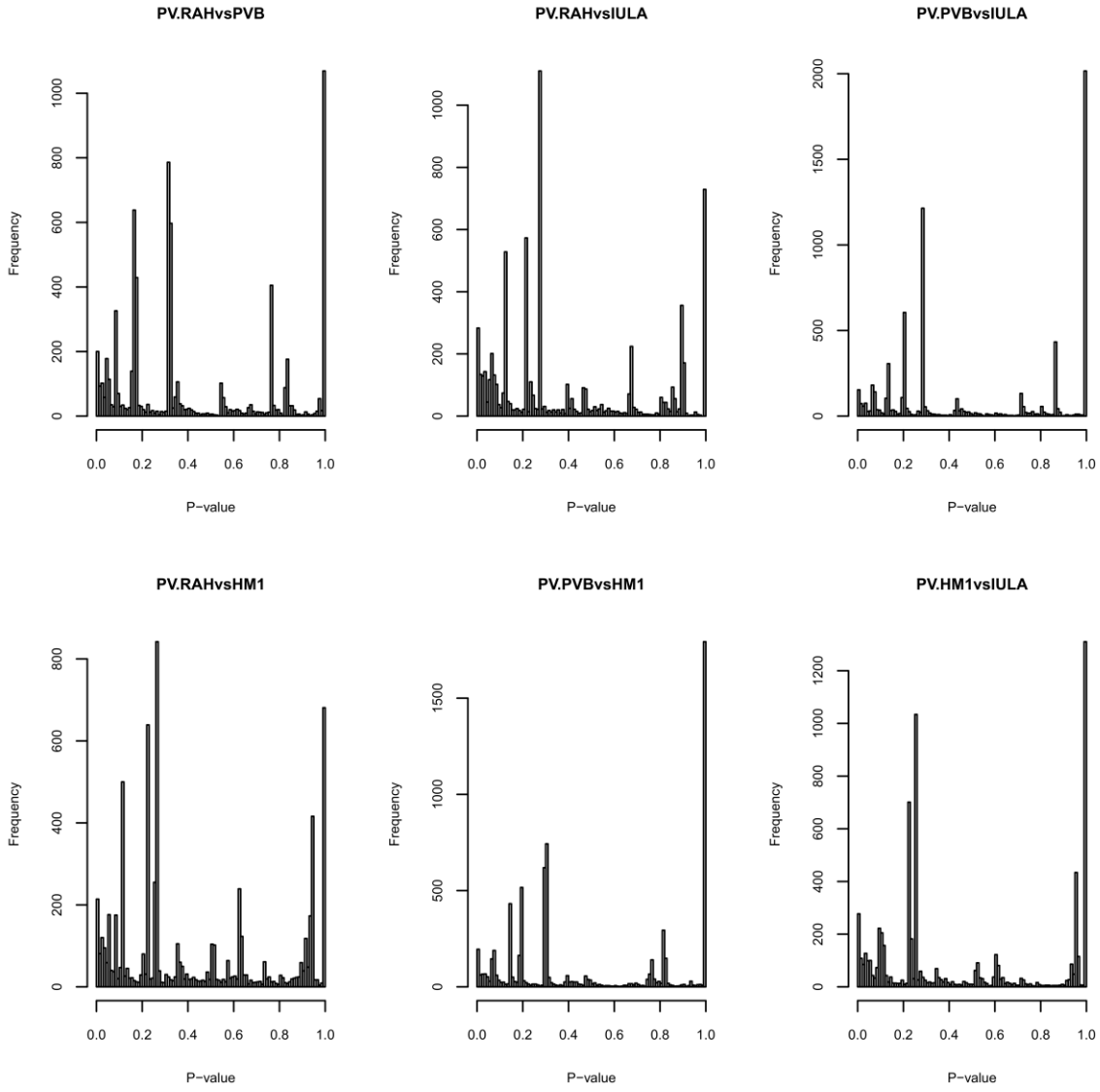


Figure 5.17: Distribution of the *P*-values for each contrast pair. Surprisingly, strong spikes of *P*-values are shown ranging from 0.2 to 0.4 in all histograms, indicating that the majority of genes between two strains have no significant difference in transcript levels of sRNAs which map to them. In other words, the number of genes with small *P*-values towards zero is less than 300 in all histograms, implying that only some genes in each strain would show the significantly higher antisense sRNA levels than another strain in a contrasting pair.

Table 5.5: The number of target genes showing significant difference (SD) in mapped antisense sRNA levels between two contrasting strains. Statistical significance was indicated if an FDR-adjusted *P*-value less than 0.05. Categories of 'SD, higher in 1st' and 'SD, higher in 2nd' mean the number of target genes with significantly higher antisense sRNA levels in 1st strain of contrast and in 2nd strain of contrast, respectively. Lower two rows of the table represent the number of target genes with significantly more than or equal to 4-fold higher antisense sRNA levels in 1st strain and 2nd strain of contrast pair, respectively.

Category	Rahman vs PVBM08B	Rahman vs HM-1:IMSS	Rahman vs IULA:1092:1	PVBM08B vs HM-1:IMSS	PVBM08B vs IULA:1092:1	HM-1:IMSS vs IULA:1092:1
SD	94	120	137	112	52	138
SD, higher in 1st	66	72	99	47	26	82
SD, higher in 2nd	28	48	38	65	26	56
SD, higher in 1st $\log_2FC \geq 2$	62	69	94	46	26	81
SD, higher in 2nd $\log_2FC \geq 2$	28	47	38	63	26	55

5.3.2 Significant negative correlation between mRNA expression and antisense sRNA transcript levels suggests a regulatory function of sRNAs

To investigate the relationship between antisense sRNAs and target gene expression, mRNA transcript levels of target genes having mapped antisense sRNAs greater than 50 reads, represented by $\log_2$(normalised transcript HTSeq-counts) on the x-axis were plotted against their mapped antisense sRNA abundance, represented by $\log_2$(normalised antisense sRNA HTSeq-counts) on the y-axis as demonstrated in Figures 5.18A, 5.19A, 5.20A and 5.21A for Rahman, PVBM08B, HM-1:IMSS and IULA:1092:1, respectively. Scatterplot analyses show significant inverse correlation between gene expression and antisense sRNA abundance in all the four strains. It is interesting that the Rahman strain exhibits the strongest correlation (r = -0.5018, P-value < $2.2e^{-16}$) compared to the others, indicating that antisense sRNAs potentially mediate regulation of gene expression.

In a similar manner, mRNA transcript levels of target genes previously used, $\log_2$(normalised transcript HTSeq-counts), were plotted against the levels of sRNA transcripts mapped sense to each gene, represented by $\log_2$(normalised sense sRNA HTSeq-counts) for all four strains as shown in Figures 5.18B, 5.19B, 5.20B and 5.21B. The scatterplots reveal the low to moderate degree of significant positive correlation (r = 0.1004-0.3318) between gene expression values and sense sRNA abundance in the four strains. One plausible interpretation is that these sense sRNAs are likely to be degradation products as highly expressed genes remarkably exhibit high levels of undegraded mRNA and corresponding short sense RNA transcripts.

Also, the 20 most prevalent functionally annotated genes targeted by antisense sRNAs were individually listed for the four strains as shown in Tables 5.6-5.9. Bases on their functional gene annotations, it could be found that functional annotated target genes encoding BspA-like LRRPs, PKs, PK domain-containing proteins, Rab family GTPases and RhoGAP domain-containing proteins are ranked within the top five orders in all four strains. Moreover, these top five functional annotations constitute a large fraction of total number of target genes. It is interesting that these target genes are members of the multigene families. As mentioned in Chapter 2, a total of 114 genes encoding BspA-like LRRPs and 307 PKs were reported in the *E. histolytica* genome [25,197]. Hence, these findings suggest that *E. histolytica* potentially regulates the expression of gene members in the multigene families using sRNA-associated mechanisms.

Intriguingly, five prevalent functional annotations, i.e. serine-threonine-isoleucine rich protein (EhSTIRP), adaptor protein family protein, C2 domain-containing protein,

proteasome regulatory subunit and Ras family GTPase, were found only in the nonvirulent Rahman as listed in Table 5.6. Of these, EhSTIRPs and C2 domain-containing proteins have been previously proven for their functional roles in association with virulence and known to be highly expressed in the virulent strains [11,168]. Altogether, the strongest inverse correlation and the unique set of virulence-associated target genes in the nonvirulent Rahman strain indicate the possible functional role of antisense sRNAs in downregulating the expression of virulence-associated genes in the nonvirulent strains.

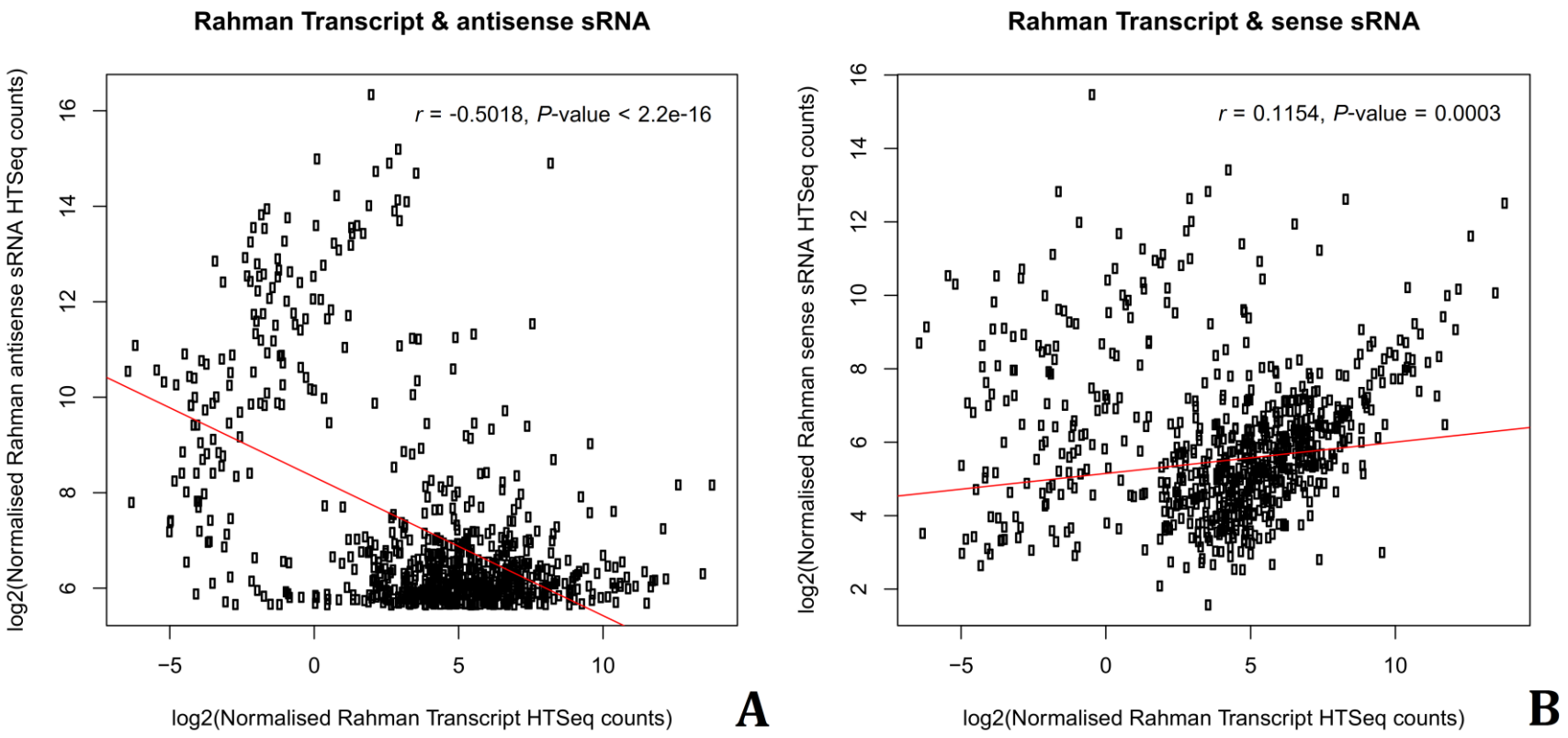


Figure 5.18: Correlation between mRNA expression levels and abundance of sRNAs mapped to a particular gene in Rahman strain. In Part A, the strong inverse correlation (r = -0.5018, P-value < $2.2e^{-16}$) is demonstrated in a set of 946 genes with the number of antisense sRNA reads > 50. The slightly positive correlation (r = 0.1154, P-value = 0.0003) is observed between mRNA abundance and sense sRNA levels, possibly due to the partial mRNA degradation in this gene set, as demonstrated in Part B. The most striking difference in correlation coefficients between above two plots strongly suggests the putative role of antisense sRNAs in regulation of gene expression in Rahman.

Table 5.6: The 20 most frequent functionally annotated genes having antisense sRNA transcript levels greater than 50 reads per kilobase of exon per million of total mapped reads in Rahman strain.

group	Functional gene annotation	Number of genes
1.	leucine-rich repeat protein, BspA family	26
2.	protein kinase domain-containing protein	10
3.	protein kinase, putative	10
4.	Rab family GTPase	8
5.	RhoGAP domain-containing protein	8
6.	DEAD/DEAH box helicase, putative	7
7.	DNA polymerase, putative	5
8.	heat shock protein 70, putative	5
9.	leucine-rich repeat-containing protein	5
10.	ubiquitin-conjugating enzyme family protein	5
11.	WD domain-containing protein	5
12.	deoxyuridine 5'-triphosphate nucleotidohydrolase domain-containing protein	4
13.	serine-threonine-isoleucine rich protein, putative	4
14.	adaptor protein (AP) family protein	3
15.	C2 domain-containing protein	3
16.	myb-like DNA-binding domain-containing protein	3
17.	myotubularin, putative	3
18.	proteasome regulatory subunit, putative	3
19.	Ras family GTPase	3
20.	zinc finger protein, putative	3

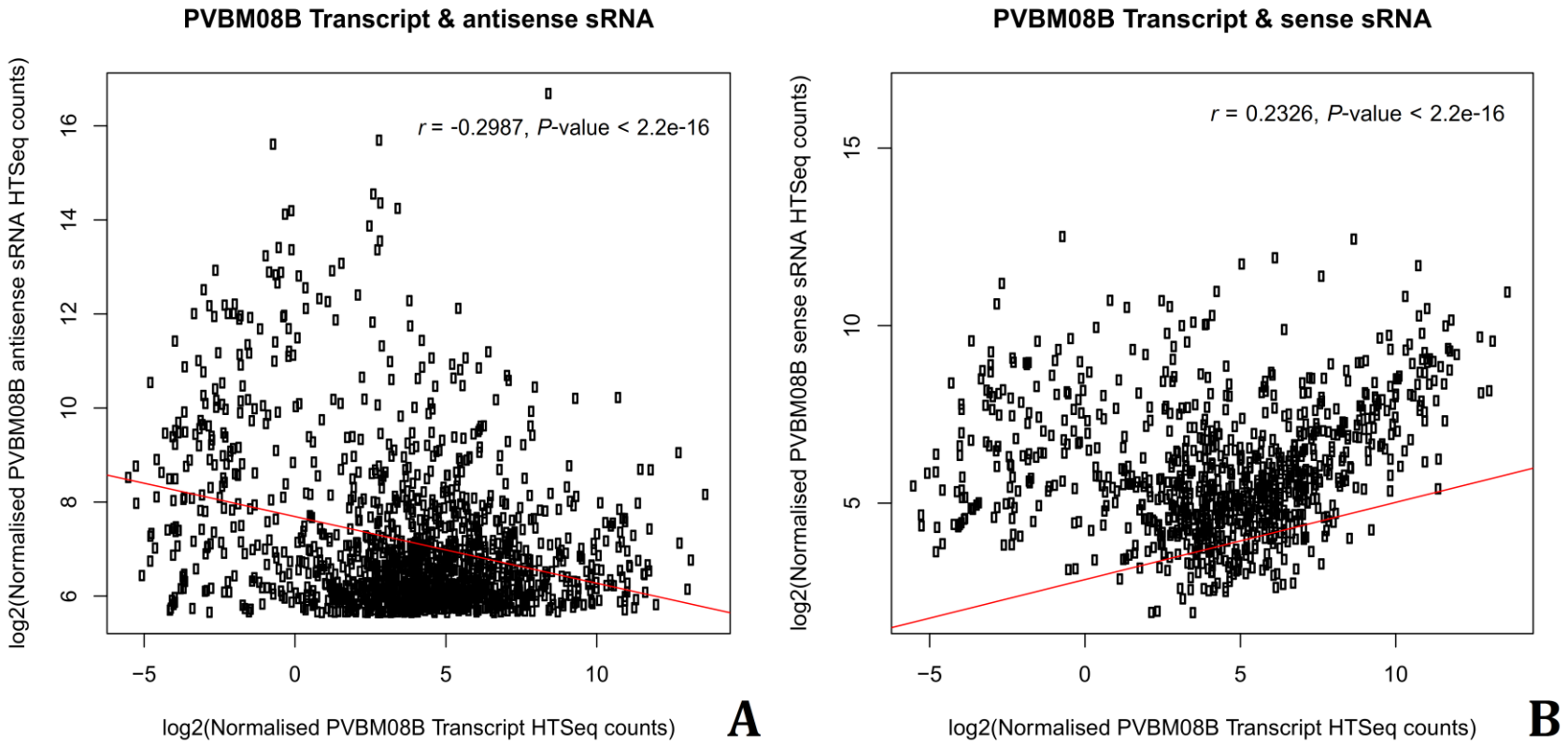


Figure 5.19: Correlation between mRNA expression levels and abundance of sRNAs mapped to a particular gene in PVBM08B strain. In Part A, the significantly inverse correlation (r = -0.2987, P-value < $2.2e^{-16}$) is demonstrated in a set of 1,593 genes with the number of antisense sRNA reads > 50. The positive correlation coefficient (r = 0.2326, P-value < $2.2e^{-16}$) between mRNA levels and sense sRNA abundance suggests the partial mRNA degradation in this gene set as shown in Part B.

Table 5.7: The 20 most frequent functionally annotated genes having antisense sRNA transcript levels greater than 50 reads per kilobase of exon per million of total mapped reads in PVBM08B strain.

group	Functional gene annotation	Number of genes
1.	Rab family GTPase	20
2.	protein kinase, putative	14
3.	RhoGAP domain-containing protein	13
4.	protein kinase domain-containing protein	12
5.	leucine-rich repeat protein, BspA family	9
6.	WD domain-containing protein	8
7.	zinc finger domain-containing protein	8
8.	protein tyrosine kinase domain-containing protein	7
9.	Rho guanine nucleotide exchange factor, putative	6
10.	acetyltransferase, GNAT family	5
11.	DNA polymerase, putative	5
12.	heat shock protein 70, putative	5
13.	ubiquitin-conjugating enzyme family protein	5
14.	zinc finger protein, putative	5
15.	acetyltransferase, putative	4
16.	ankyrin repeat protein, putative	4
17.	dual specificity protein phosphatase, putative	4
18.	helicase, putative	4
19.	LIM zinc finger domain-containing protein	4
20.	lipid phosphate phosphatase, putative	4

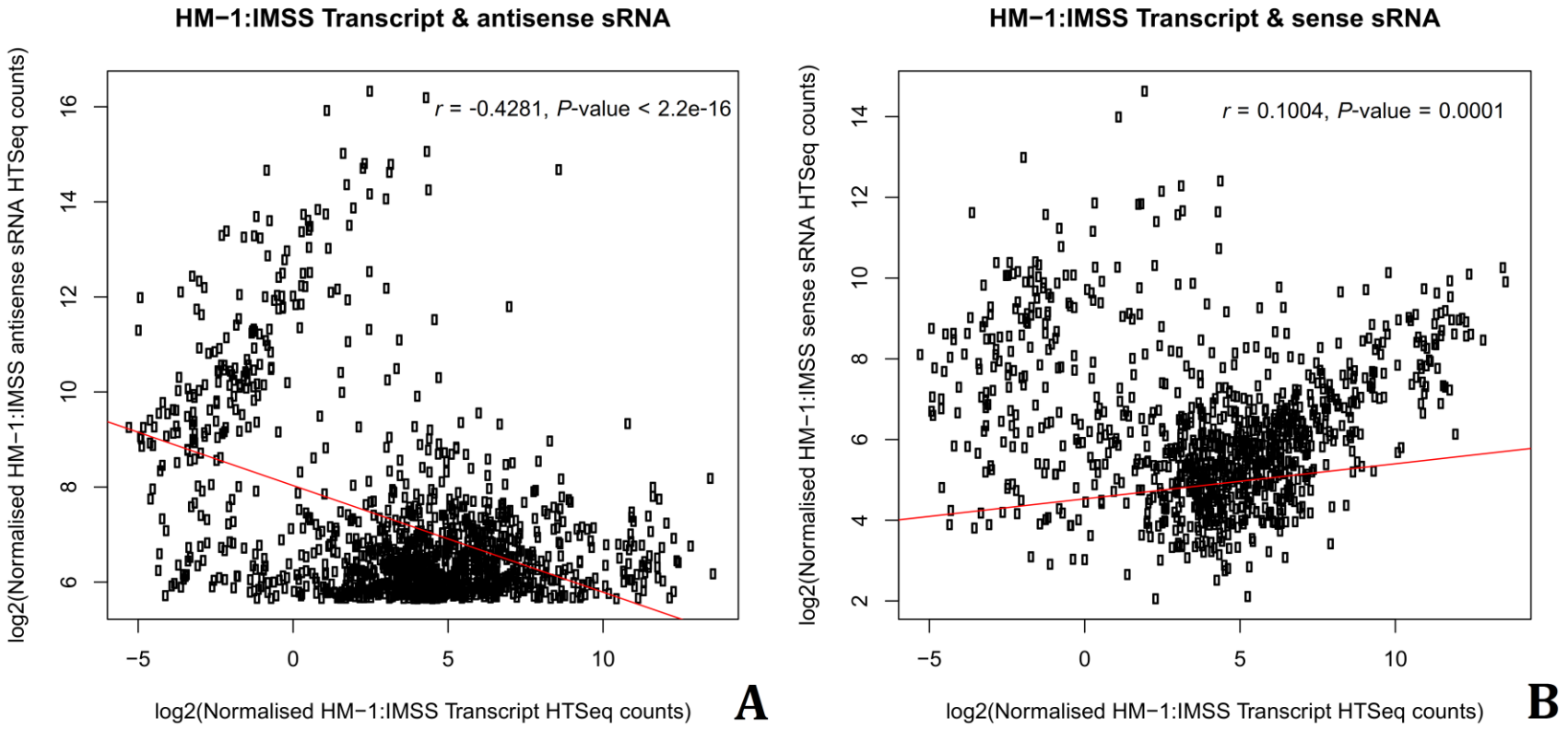


Figure 5.20: Correlation between mRNA expression levels and abundance of sRNAs mapped to a particular gene in HM-1:IMSS strain. In Part A, the significantly inverse correlation (r = -0.4281, P-value < $2.2e^{-16}$) is demonstrated in a set of 1,405 genes with the number of antisense sRNA reads > 50. The small agreement between mRNA and sense sRNA levels (r = 0.1004, P-value = 0.0001) is observed possibly due to the partial mRNA degradation in this gene set as seen in Part B.

Table 5.8: The 20 most frequent functionally annotated genes having antisense sRNA transcript levels greater than 50 reads per kilobase of exon per million of total mapped reads in HM-1:IMSS strain.

group	Functional gene annotation	Number of genes
1.	Rab family GTPase	24
2.	leucine-rich repeat protein, BspA family	12
3.	protein kinase domain-containing protein	12
4.	protein kinase, putative	11
5.	RhoGAP domain-containing protein	10
6.	myb-like DNA-binding domain-containing protein	7
7.	Rho family GTPase	7
8.	RNA recognition motif domain-containing protein	6
9.	ubiquitin-conjugating enzyme family protein	6
10.	zinc finger protein, putative	6
11.	DNA polymerase, putative	5
12.	WD domain-containing protein	5
13.	acetyltransferase, GNAT family	4
14.	calmodulin, putative	4
15.	deoxyuridine 5'-triphosphate nucleotidohydrolase domain-containing protein	4
16.	EF-hand calcium-binding domain-containing protein	4
17.	endonuclease/exonuclease/phosphatase family protein	4
18.	leucine-rich repeat-containing protein	4
19.	LSM domain-containing protein	4
20.	Rab GTPase-activating protein, putative	4

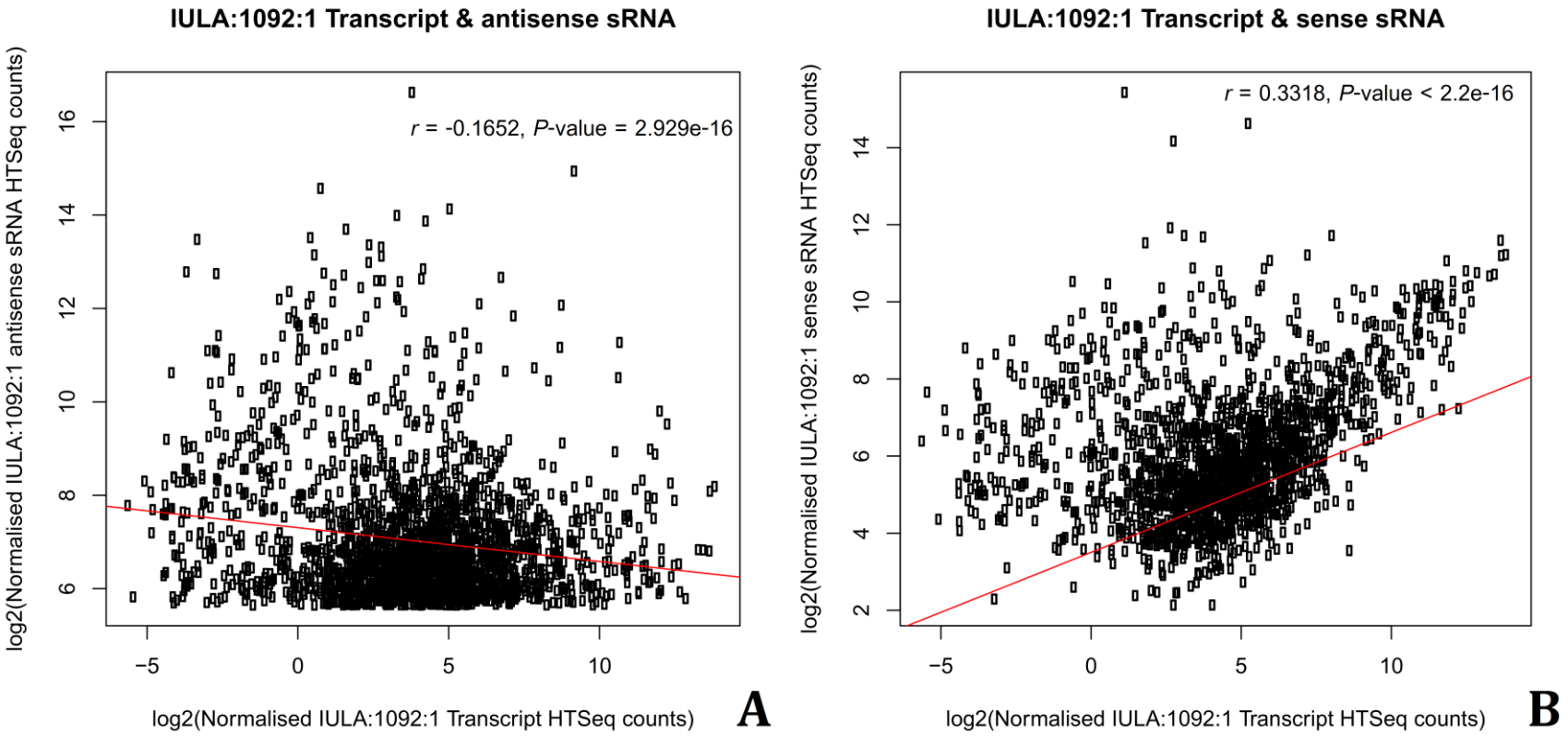


Figure 5.21: Correlation between mRNA expression levels and abundance of sRNAs mapped to a particular gene in IULA:1092:1 strain. In Part A, the significantly inverse correlation (r = -0.1652, P-value = 2.929e^{-16}) is demonstrated in a set of 2,416 genes with the number of antisense sRNA reads > 50 (A). The correlation coefficient between mRNA and sense sRNA levels is rather highly positive (r = 0.3318, P-value < 2.2e^{-16}), suggesting the remarkable mRNA degradation in this gene set as depicted in Part B.

Table 5.9: The 20 most frequent functionally annotated genes having antisense sRNA transcript levels greater than 50 reads per kilobase of exon per million of total mapped reads in IULA:1092:1 strain.

group	Functional gene annotation	Number of genes
1.	protein kinase domain-containing protein	29
2.	protein kinase, putative	27
3.	Rab family GTPase	22
4.	leucine-rich repeat protein, BspA family	20
5.	RhoGAP domain-containing protein	16
6.	WD domain-containing protein	14
7.	tyrosine kinase, putative	13
8.	DEAD/DEAH box helicase, putative	10
9.	RNA recognition motif domain-containing protein	10
10.	Rho guanine nucleotide exchange factor, putative	9
11.	heat shock protein 70, putative	8
12.	Rap/Ran GTPase-activating protein, putative	8
13.	Ras guanine nucleotide exchange factor, putative	7
14.	thioredoxin, putative	7
15.	zinc finger domain-containing protein	7
16.	leucine-rich repeat-containing protein	6
17.	LIM zinc finger domain-containing protein	6
18.	myotubularin, putative	6
19.	protein tyrosine kinase domain-containing protein	6
20.	pumilio family RNA-binding protein	6

5.3.3 High abundance of antisense sRNAs in the nonvirulent Rahman strain is associated with the downregulation of virulence-associated gene expression

As shown in Figure 5.22, Venn diagrams show that a total of 31 target genes exhibit significantly higher levels of mapped antisense sRNAs in Rahman than the other three strains. These 31 target genes could be categorised into 7 functional groups (i.e. host cell killing and mucosal invasion, calcium binding, nucleic acid interaction, protein folding, signaling, others and hypothetical) based on their functional annotations as listed in Table 5.10. It is interesting that the majority of these functional target genes have been characterised for their functional roles such as host cell killing and mucosal invasion, calcium binding, nucleic acid interaction and signaling as discussed in Chapter 2. Also, 29 of these 31 target genes show more than 4-fold higher antisense sRNA levels in Rahman, relative to the others as demonstrated in Figure 5.23. This indicates that the majority of sRNAs are remarkably different in their expression levels between nonvirulent and virulent strains.

Most members of these target genes such as genes encoding BspA-like LRRPs, DEAD/DEAH box helicases, Hsp70 chaperones, EhSTIRPs and C2 domain-containing proteins were found to be prevalent in the 20 most frequent functionally annotated genes that were targeted by greater than 50 reads of antisense sRNAs as listed in Table 5.6. It is striking that 15 of these 31 sRNA target genes as listed in Table 5.11 show significant downregulation in Rahman when compared to the other three strains, indicating that these antisense sRNAs play an important role in post-translational gene silencing. Also, this finding strongly supports that antisense sRNAs in the nonvirulent Rahman are most likely to play a key role in regulation of mRNA transcript levels, especially of virulence-associated genes.

In contrast, no common gene in all three virulent strains shows significantly higher antisense sRNA level than Rahman as demonstrated in Figures 5.24 and 5.25, suggesting that sRNA-mediated regulation is less pronounced in these three virulent strains. As such, these findings are consistent with the 454 sequencing result of the previous study showing the large difference in the number of antisense sRNAs mapped to *EhSTIRP1* gene (EHI_025700) between Rahman and HM-1:IMSS [92]. In other words, less stringency in sRNA-mediated regulation of virulence-associated gene expression in the virulent strains potentially results in gene overexpression and consequent pathogenic behaviours such as host tissue destruction and mucosal invasion. In agreement with the GO enrichment analysis in Chapter 2, these experimental findings strongly support that less regulatory stringency in

both signaling cascades and sRNA-mediated silencing potentially contribute to the virulence in this parasite.

Screenshots from the Integrative Genomics Viewer (IGV) show sequenced sRNA reads and mRNA reads mapped to *EhSTIRP* gene EHI_004340 responsible for host cell adhesion and cytotoxicity as illustrated in Figures 5.26-5.29 for Rahman, PVBM08B, HM-1:IMSS and IULA:1092:1, respectively. The inverse correlation was observed between gene expression and abundance of small RNAs in all four axenic *E. histolytica* strains. For Rahman in Figure 5.26, it shows a high level of sRNA reads (~290 antisense reads) mapped to the *EhSTIRP* gene EHI_004340 but exhibits very low *EhSTIRP* mRNA expression. Conversely, the two strains associated with virulence (i.e. PVBM08B and HM-1:IMSS) have no mapping of any small RNA to the *EhSTIRP* gene but show very high *EhSTIRP* gene expression. Surprisingly, the less virulent strain IULA:1092:1 has very few sRNA reads (~8.5 antisense reads) mapped to the *EhSTIRP* gene but shows the moderate gene expression with marked reduction in *EhSTIRP* mRNA transcripts relative to PVBM08B and HM-1:IMSS, implying that antisense sRNAs are very effective in gene silencing.

In addition, the majority of sRNA transcripts are oriented in the antisense direction and predominantly map to the 5' end of the *EhSTIRP* gene as shown in blue in the IGV alignment. These features have been previously reported in an abundant population of 27 nt antisense sRNAs with 5'-polyphosphate termini in *E. histolytica* and found to be enriched in the Argonaute immunoprecipitated sample, first described by Zhang *et al*., 2008 [85]. Furthermore, these 27 nt 5'-polyphosphorylated sRNAs also constitute a large fraction of the sRNA transcriptome and possess a biased 5'-G sequence [85]. Zhang *et al*., 2011 demonstrated that such sRNAs play a crucial role in long term transcriptional gene silencing through a siRNA-mediated pathway in the genetically engineered *E. histolytica* G3 strain [86].

In this study, the experiment of size selection by 3% Agarose Pippin Prep was designed based on the expected sequence length of miRNA molecules (21-23 nt). However, the majority of the size-fractionated sRNAs have sequence length distributions at approximately 23-26 nt, slightly less than the 27 nt antisense sRNAs previously reported [85,86,92]. Taken together, these findings strongly indicate that the 23-26 nt antisense sRNAs in this study are not miRNAs as firstly hypothesised, and these antisense sRNAs are likely to play a key role in regulating virulence-associated gene expression via the siRNA pathway as previously mentioned.

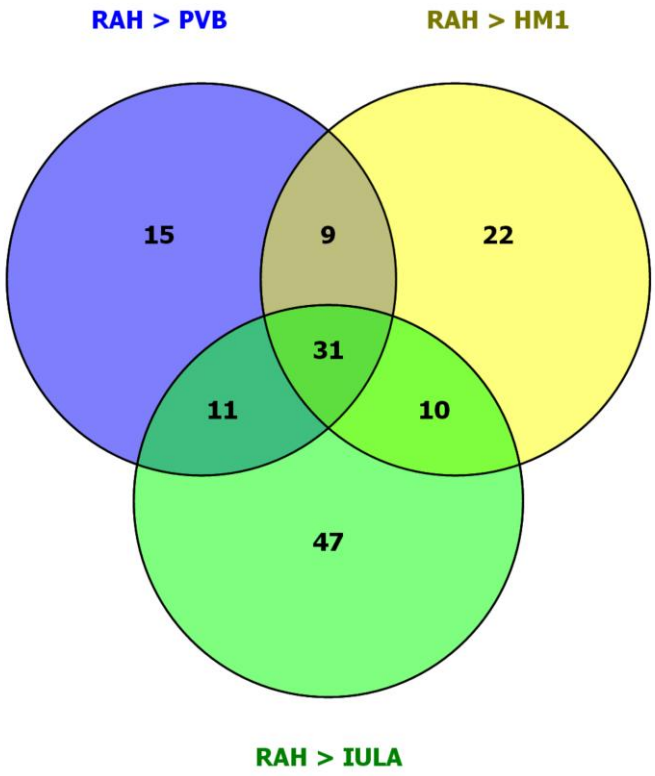


Figure 5.22: The number of target genes with significantly higher antisense sRNA transcript levels (FDR-adjusted *P*-value < 0.05) in Rahman than thother three virulent strains (i.e. PVBM08B, HM-1:IMSS and IULA:1092:1). A total of 31 target genes regardless of their log_2FC show higher antisense sRNA levels in Rahman than all the others.

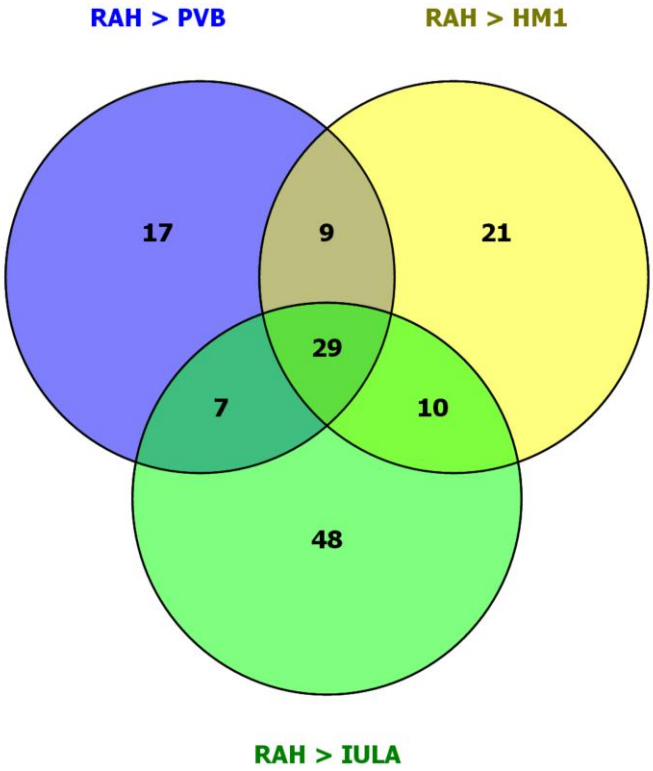


Figure 5.23: The number of target genes with significantly more than or equal to 4-fold higher antisense sRNA transcript levels ($log_2FC \geq 2$) in Rahman than the others. A total of 29 target genes show markedly high levels of antisense sRNA transcripts in Rahman with $log_2FC \geq 2$.

Table 5.10: Summary of 31 target genes showing markedly high antisense sRNA levels in Rahman (29 members with $log_2FC \geq 2$ and 2 members (*) with $log_2FC < 2$), assigned to 7 functional categories with their functional gene annotations and AmoebaDB_IDs.

Gene Category	Functional gene annotation	Number of genes	AmoebaDB_ID
Host cell killing and mucosal invasion	serine-threonine-isoleucine rich protein	2	EHI_004340, EHI_025700
	leucine-rich repeat protein, BspA family	5	EHI_015120, EHI_095060, EHI_100700, EHI_127710, EHI_194290
Calcium binding	C2 domain-containing protein	2	EHI_059860, EHI_069320
Nucleic acid interaction	DEAD/DEAH box helicase, putative	1	EHI_119620
	RNA-binding protein, putative	1	EHI_053170
Protein folding	heat shock protein 70, putative	1	EHI_133950, EHI_150770
Signaling	Rap/Ran GTPase-activating protein, putative	1	EHI_108750
	dedicator of cytokinesis domain-containing protein	1	EHI_185270
Others	acetate kinase	1	EHI_170010(*)
	CXXC-rich protein	1	EHI_050970
Hypothetical	N/A	14	EHI_004560, EHI_010280, EHI_012080, EHI_020890, EHI_021580, EHI_074080, EHI_098720, EHI_113790, EHI_119790, EHI_165190, EHI_168830(*), EHI_174500, EHI_180410, EHI_188910
	Total	**31**	

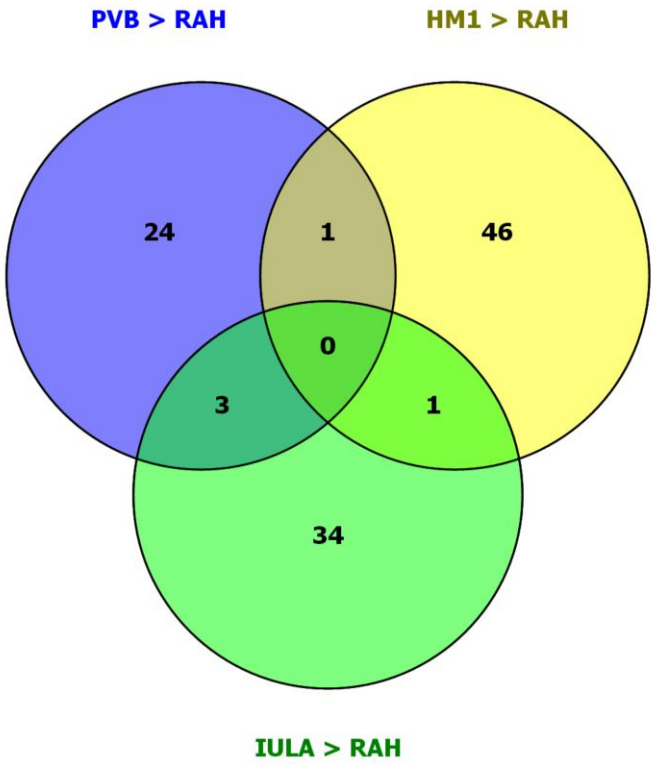


Figure 5.24: The number of target genes known to have significantly higher antisense sRNA transcript levels (FDR-adjusted *P*-value < 0.05) in the three virulent strains than the Rahman. In these three virulent strains, no common target gene shows higher antisense sRNA level, compared to Rahman.

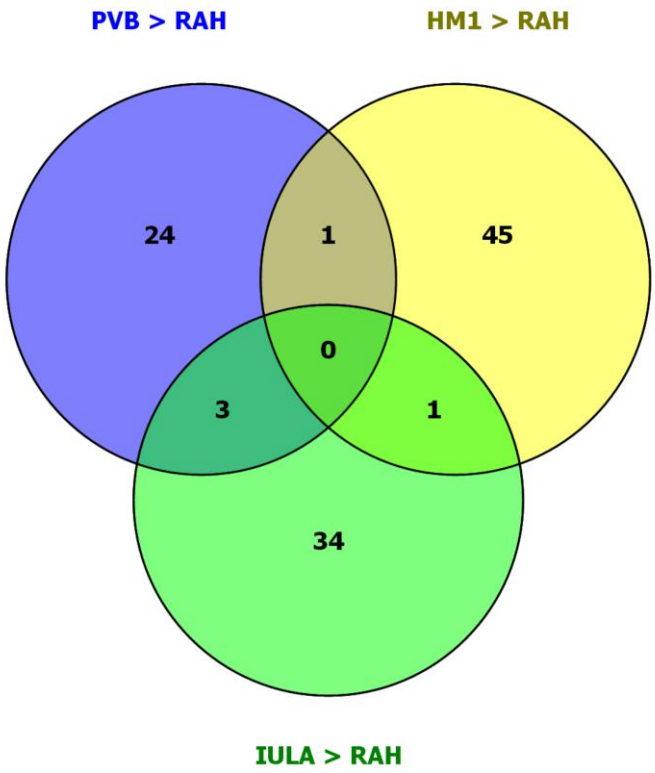


Figure 5.25: The number of target genes with significantly more than or eqaul to 4-fold higher antisense sRNA transcript levels ($log_2FC \geq 2$) in the three virulent strains than Rahman.

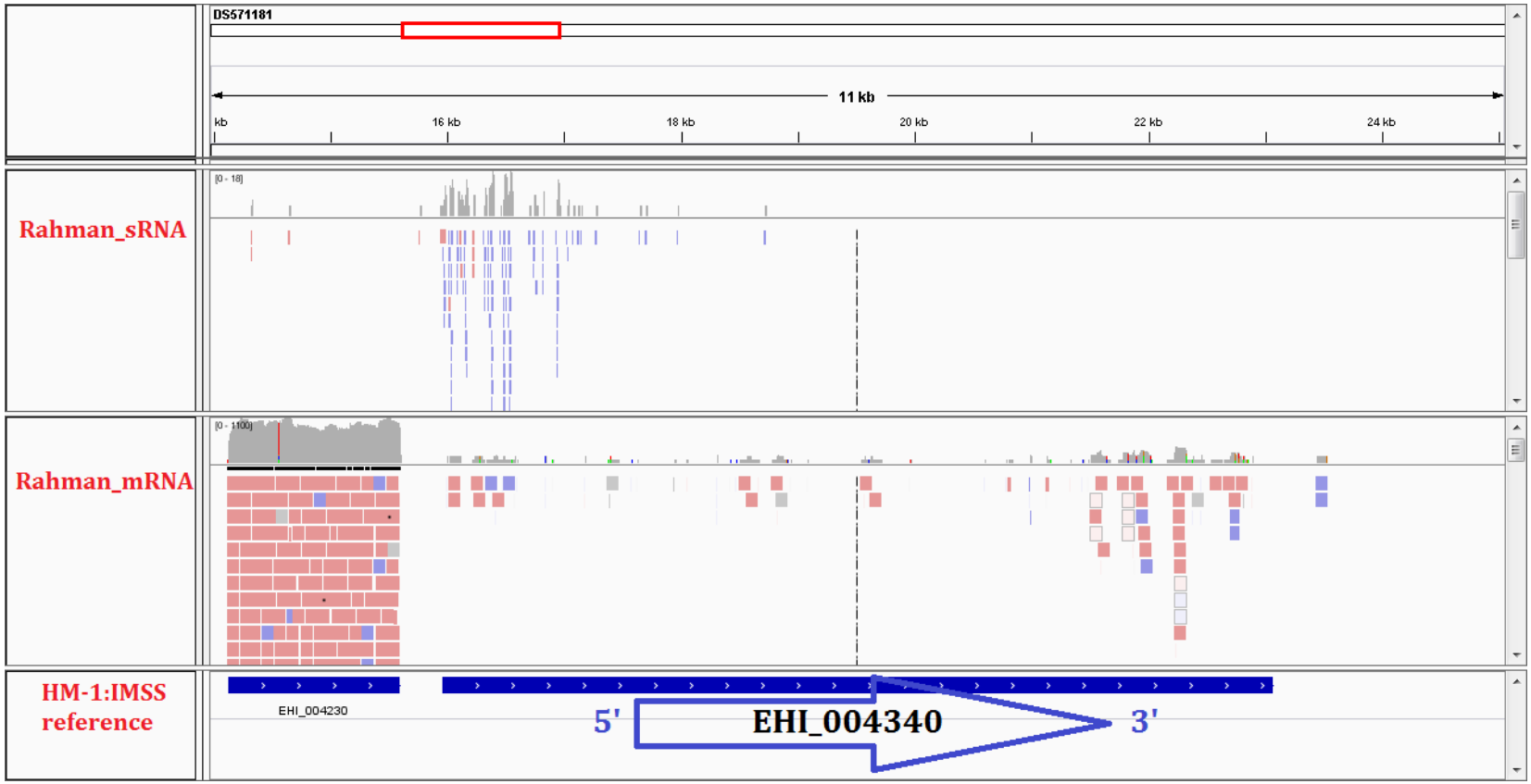


Figure 5.26: The Integrative Genomics Viewer (IGV) showing the population of small RNA transcripts mapped to the very lowly expressed *EhSTIRP* gene EHI_004340 in the nonvirulent Rahman strain. Strikingly, the majority of sRNA transcripts are oriented in the antisense direction (blue colour) and predominantly map to the 5' end of gene. The adjacent gene (EHI_004230) encoding guanine nucleotide regulatory protein shows high mRNA expression with very few sense sRNAs, probably degraded mRNA fragments. The high level of antisense sRNAs mapped to the *EhSTIRP* gene EHI_004340 is associated with downregulation of the EHI_004340 mRNA transcripts, implying their possible role in gene silencing [92].

Figure 5.27: The Integrative Genomics Viewer (IGV) showing no sRNA mapping to the highly expressed *EhSTIRP* gene EHI_004340 in the virulent PVBM08B strain.

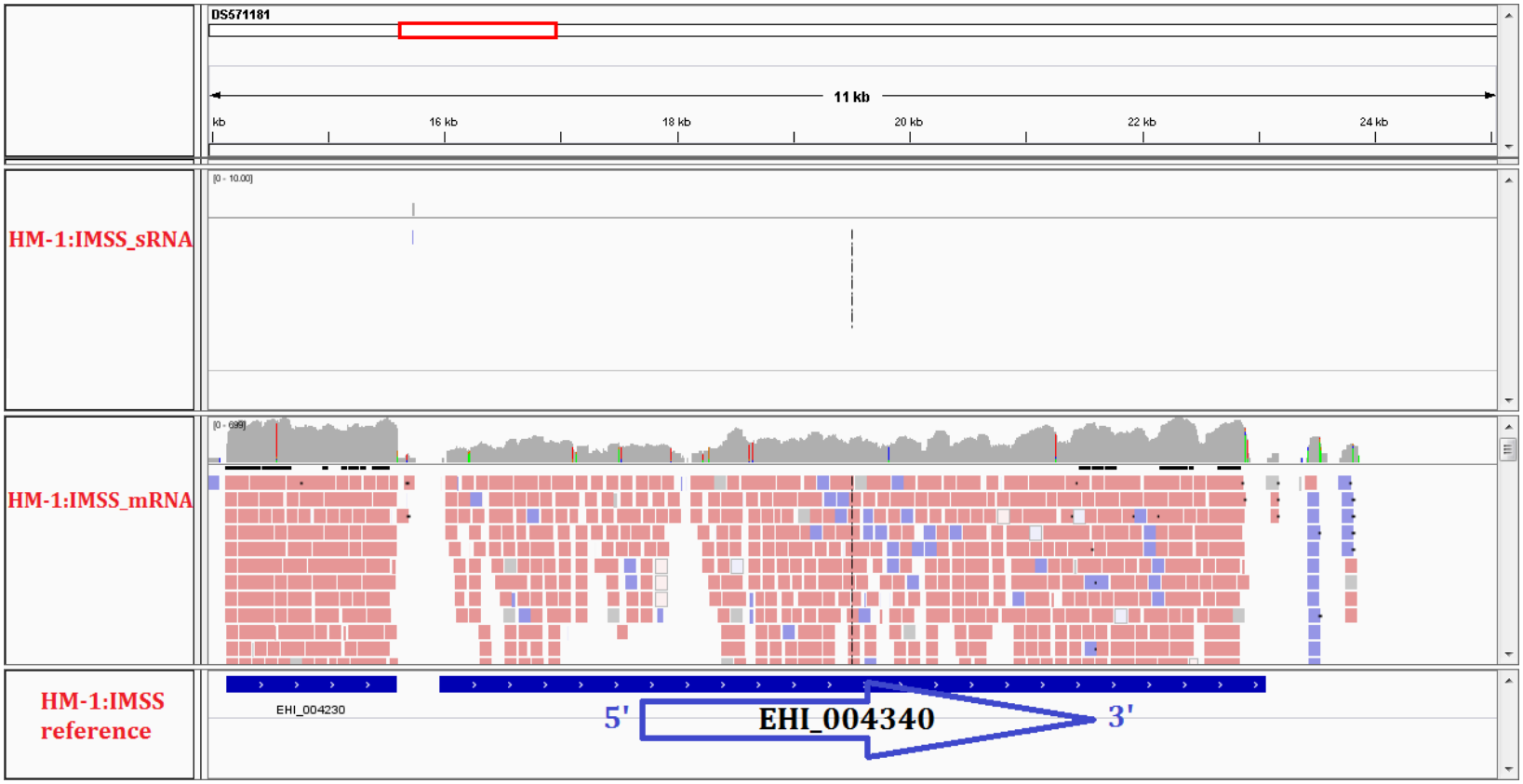


Figure 5.28: The Integrative Genomics Viewer (IGV) showing no sRNA mapping to the highly expressed *EhSTIRP* gene EHI_004340 in the virulent HM-1:IMSS strain.

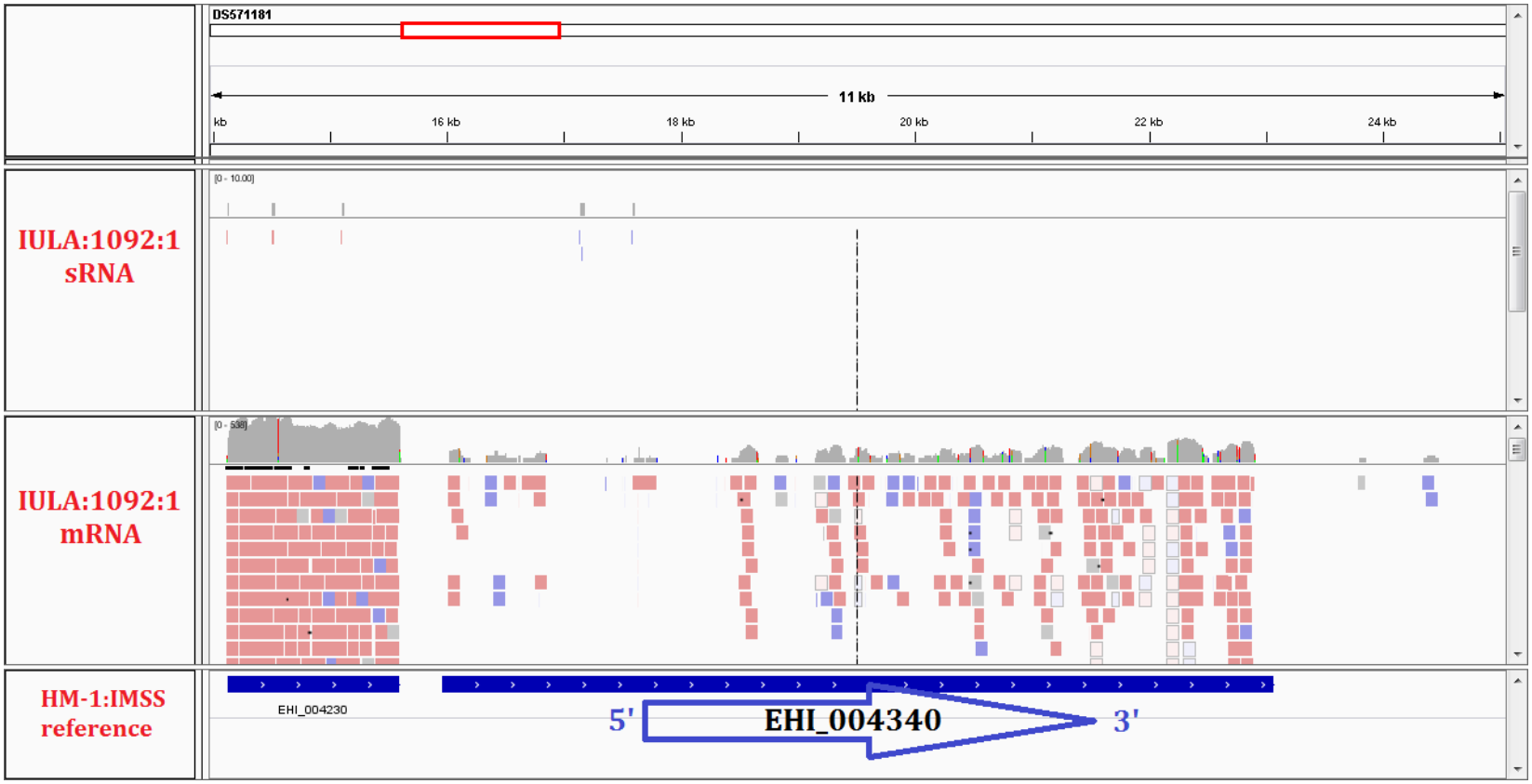


Figure 5.29: The Integrative Genomics Viewer (IGV) showing very few antisense sRNA transcripts mapped to the moderately expressed *EhSTIRP* gene EHI_004340 in the virulent IULA:1092:1 strain.

5.3.4 Small RNAs partially contribute to genome-wide transcriptomic variation between nonvirulent and virulent *E. histolytica* strains

Comparing with the whole transcriptome results in Chapter 2, it was found that only 31 target genes show significant differences in the number of antisense sRNAs mapped to a particular gene between nonvirulent Rahman and the other three virulent strains as listed in Table 5.10, whilst a total of 2,159 genes, that comprise 1,162 and 997 genes with upregulation and dowregulation in the three virulent strains, exhibit significant modulation in their gene expression levels between nonvirulent Rahman and the other three virulent strains.

As shown in Figure 5.30, only 15 genes show significantly higher mRNA levels but have lower levels of mapped antisense sRNAs in all three virulent strains relative to Rahman whereas the other 16 genes with no significant difference in mRNA expression among the four *E. histolytica* strains exhibit markedly higher antisense sRNA levels in Rahman. Interestingly, only approximately 1.29% (15/1,162) of total genes showing lower mRNA expression in Rahman have remarkably high antisense sRNA levels, implying antisense sRNA molecules partially contribute to differential expression among the four *E. histolytica* strains.

In summary, these results show that transcriptomic variations among *E. histolytica* strains are affected by diverse gene regulatory elements such as the sRNA-mediated RNAi pathway and other genomic factors including copy number variation, segmental genome duplication, gene gain or gene loss and even single nucleotide polymorphisms. Besides the siRNA-associated silencing, other epigenetic mechanisms, e.g. DNA methylation and histone modification, have also been reported to be involved in transcriptional gene silencing in *E. histolytica* [144-146]. Conclusively, my experimental findings with the previously published evidences enable us to understand that the parasite exploits many cellular tools for regulating their transcriptomes in a synergistic manner in response to the host environmental stress and in the long term adaptation [85,86,92,93,144-146].

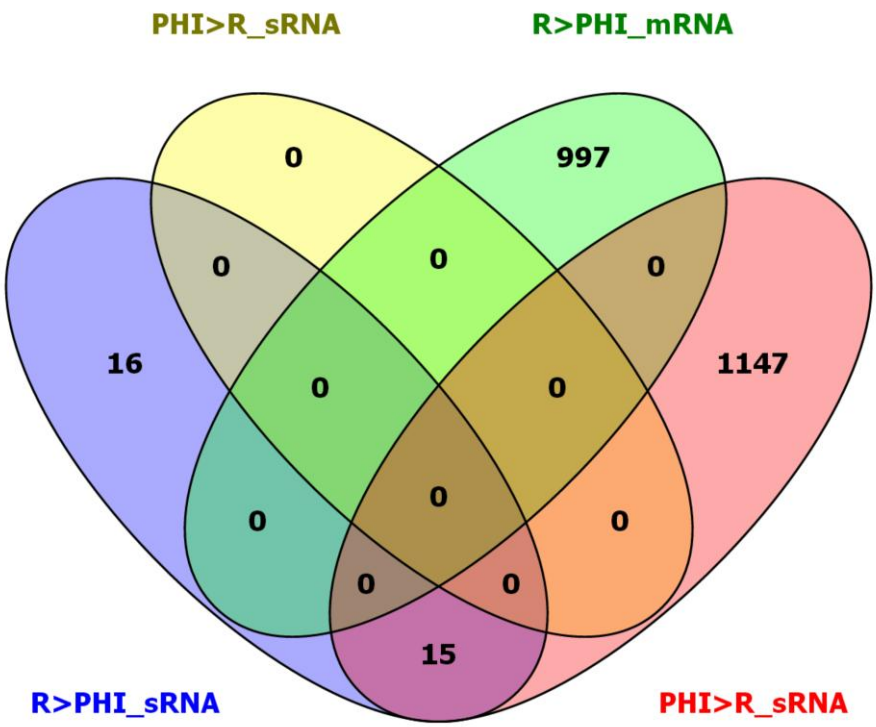


Figure 5.30: The number of genes having significantly higher mRNA levels and lower antisense sRNA levels in all three virulent strains relative to Rahman (n=15) and the number of genes having no difference in expression among the four *E. histolytica* strains but showing markedly higher levels of antisense sRNAs in Rahman (n=16). The intersection of gene members in each coloured area is based on AmoebaDB_IDs.

Abbreviations are as follows: R>PHI_sRNA = the number of genes with higher antisense sRNA transcripts in Rahman than the other three virulent strains (i.e. PVBM08B, HM-1:IMSS and IULA:1092:1); PHI>R_sRNA = the number of genes with higher antisense sRNA transcripts in the three virulent strains than Rahman; R>PHI_mRNA = the number of genes with higher mRNA expression in Rahman than the others; PHI>R_mRNA = the number of genes with higher mRNA expression in the three virulent strains than Rahman.

The details of the two unique gene sets mentioned above are summarised in Tables 5.11 and 5.12.

Table 5.11: Summary of 15 target genes having higher mRNA expression in all three virulent strains and showing markedly high antisense sRNA levels in Rahman with $log_2FC \geq 2$, assigned to 7 functional categories with their functional gene annotations and AmoebaDB_IDs.

Gene Category	Functional gene annotation	Number of genes	AmoebaDB_ID
Host cell killing and mucosal invasion	serine-threonine-isoleucine rich protein	2	EHI_004340, EHI_025700
	leucine-rich repeat protein, BspA family	2	EHI_015120, EHI_095060
Calcium binding	C2 domain-containing protein	1	EHI_069320
Nucleic acid interaction	RNA-binding protein, putative	1	EHI_053170
Protein folding	heat shock protein 70, putative	2	EHI_133950, EHI_150770
Signaling	Rap/Ran GTPase-activating protein, putative	1	EHI_108750
	dedicator of cytokinesis domain-containing protein	1	EHI_185270
Others	CXXC-rich protein	1	EHI_050970
Hypothetical	N/A	4	EHI_010280, EHI_074080, EHI_180410, EHI_188910
	Total	**15**	

Table 5.12: Summary of 16 target genes having no differential expression among the four *E. histolytica* strains but showing markedly high antisense sRNA levels in Rahman (14 members with $log_2FC \geq 2$ and 2 members (*) with $log_2FC < 2$) with $log_2FC \geq 2$, assigned to 5 functional categories with their functional gene annotations and AmoebaDB_IDs.

Gene Category	Functional gene annotation	Number of genes	AmoebaDB_ID
Host cell killing and mucosal invasion	leucine-rich repeat protein, BspA family	3	EHI_100700, EHI_127710, EHI_194290
Calcium binding	C2 domain-containing protein	1	EHI_059860
Nucleic acid interaction	DEAD/DEAH box helicase, putative	1	EHI_119620
Others	acetate kinase	1	EHI_170010(*)
Hypothetical	N/A	10	EHI_004560, EHI_010280, EHI_020890, EHI_021580, EHI_098720, EHI_113790, EHI_119790, EHI_165190, EHI_168830(*), EHI_174500
	Total	**16**	

5.3.5 Discovery of novel miRNA candidates by miRDeep2 software suggests the existence of regulatory miRNA in *E. histolytica*

Based on its biogenesis, miRNAs have remarkably distinctive characteristics in structure [290,291,293]. Their precursors called 'pri-miRNA' are transcribed by RNA polymerase II in the nucleus and fold back into imperfect stem-loop hairpin structures. The pri-miRNAs are further processed into resulting pre-miRNAs with approximately 70 nt in length by the microprocessor complex mainly consisting of RNase III and Drosha. Then, the pre-miRNAs are exported into the cytoplasm through the nuclear pore complex. In the cytoplasm, the loop regions are cleaved from the stem portions of the pre-miRNAs by the RNase III enzyme Dicer. After Dicer processing, the remaining stem of the pre-miRNAs is composed of one strand of mature miRNA with approximately 19-23 nt and the other as a star sequence, miRNA*.

De *et al*., 2006 previously reported seventeen putative miRNA candidates in *E. histolytica* using a computational method [95]. Briefly, hairpin repeats were identified throughout the *E. histolytica* genome (1,819 contigs). Approximately 15,000 repeats were further determined for folding energy, structural filters, i.e. length and gap. Using the nucleotide BLAST analysis, the obtained repeats were then eliminated if having homology over 97% over the length of 45 nt or more since they would be considered as a part of coding DNA sequence. Also, the repeats with less than 60% similarity with the coding mRNA were filtered out as they seem unlikely to anneal with the target mRNAs. Finally, 17 distinctive repeats which only one strand of them aligned with the coding sequence were identified as putative miRNA candidates (miR-1 to miR-17) [95].

Using these seventeen candidates as novel predicted miRNAs, the nucleotide BLAST analysis can identify 32 targets allowing for no more than 2 gap mismatches. The majority of target genes are identified as 'hypothetical' while some are involved in signaling pathways and encystation process [95]. Also, several machinery proteins involved in miRNA- and siRNA-mediated gene silencing such as AGO, DEAD/DEAH box helicase, RdRp as well as RNase III Dicer have been previously identified [95,295]. So, it is highly promising that a miRNA machinery might be found in this parasite.

Mar-Aguilar *et al*., 2013 reported the potential 199 miRNA candidates that were identified from hairpin-forming precursors in the sequenced small RNA dataset and validated by microarray analysis in the HM-1:IMSS strain [94]. Also, 9 of 10 selected miRNA candidates were amplified by real-time PCR, indicating the reliability of novel miRNA prediction [94]. In Mar-Aguilar *et al.*'s study, a total of 66 miRNA target genes were

predicted by the miRanda algorithm and found to be involved in many cellular processes, for example: transcriptional regulation, e.g. zinc finger protein; signal transduction, e.g. Ras family GTPase, Rap/Ran GTPase-activating protein and protein kinase; calcium-dependent regulation, e.g. C2 domain-containing protein; receptor-mediated endocytosis, e.g. clathrin adaptor complex small chain [94]. Although this study was the first experimental identification using the sRNA sequencing data with filtering reads matched with other types of non-coding sRNAs, such 199 novel miRNAs were predicted mainly based on the ability of their putative precursor to form the hairpin secondary structure, prone to be contaminated with other background hairpins. Therefore, the novel miRNA prediction in this parasite needs to be further elucidated.

Recently, the advance of deep sequencing technology has allowed researchers to predict novel miRNAs from small RNA transcriptomes based on the unique characteristics of miRNA biogenesis as explained in Figure 5.9. The miRDeep2 package was designed to check the compatibility between raw sequenced small RNA reads and predicted pre-miRNA precursors using the probabilistic algorithm based on positions, frequencies of sequenced reads matched with Dicer cleavage and thermodynamic stability [297]. Therefore, miRDeep2 allows users to identify both known and novel predicted miRNAs and also provides the false-positive rate and statistical significances of energetic stability.

In this study, miRDeep2 was applied to predict novel miRNA candidates from the sRNA sequencing data obtained from each sRNA library. The processed read FASTA files were aligned against the HM-1:IMSS genomic reference in comparison to the known miRNAs and precursors of sibling species *D. discoideum* to identify the novel miRNA precursors with probabilistic scoring, secondary structure and minimal free energy as detailed in Table 5.13. Only potential precursors that could form a stem-loop hairpin and had mapped to short sequence reads in a manner compatible with Dicer cleavage were analysed. A total of three different potential miRNA candidates: miR-Rah1, miR-Rah2 and miR-PVB2, as shown in Figures 5.31-5.34 were identified separately from three sRNA libraries: Rahman_01, Rahman_02 and PVBM08B_2, respectively.

Comparing with the previous studies, these three novel miRNA candidates in this study do not match with those miRNA candidates previously published [94,95]. This could be due to the different algorithms used in the former studies which investigated solely the presence of hairpin forming repeats with appropriate folding energy in the genomic DNA sequence. Moreover, purely computational analysis might encounter a large number of false positive candidates due to background hairpins and needs experimental validations which are complicated for rare miRNA transcripts.

Differently, the novel miRNA candidates in this study were predicted from the deep sequencing data using the probabilistic scoring algorithm as explained before. Essentially, the miRDeep software package was designed to use the sequenced reads as a guideline to excise the plausible miRNA precursors from the genomic sequences with statistical confidence according to a highly characteristic model of miRNA biogenesis. Moreover, deep-sequencing data can detect the transcripts with a vast dynamic range and also illustrate the number of reads assigned to each particular genomic position, enabling us to re-evaluate for the appropriate mature and star sequences. Therefore, this expression-based identification could provide us much more promising miRNA signatures in *E. histolytica*, compared to the previous reports.

Using the same genomic HM-1:IMSS reference, however, the prediction data exhibit the novel miRNA candidates only for Rahman and PVBM08B libraries, suggesting that such particular candidates are variable in expression among the strains. Based on the important criteria for miRNA annotation previously published, reliability of miRNA candidates are mainly based on the presence of multiple sequenced reads with homogeneity at the 5' end and a two nucleotide overhang at the 3' end of the pre-miRNA precursor [298,300]. Also, the mature sequence reads must be consistent with 5' processing by starting at the same nucleotide position as explained in Figure 5.9. Differently, 16 of the first 22 nt positions in the animal miRNAs typically exhibit complementarity with the star sequence whilst the plant miRNAs possess more stringent complementary base pairing between miRNA and star arms with no more than 4 mismatches. Herein, the miR-Rah1 candidate shows all qualified attributes mentioned above as illustrated in Figure 5.32, therefore this predicted miRNA candidate was chosen for the validation. The qPCR amplication curve reveals the expression of this predicted miRNA in *E. histolytica* strains, as shown in Figure 5.35 and Table 5.14

De *et al*., 2006 reported thirty-two targets of putative predicted miRNAs using the nucleotide BLAST analysis, showing base pairing greater than or equal to 21 nt and allowing one or two mismatches [95]. In this study, I have tried to identify the potential miRNA targets by applying the miR-Rah1 sequence as a query in the nucleotide BLAST (available at http://www.ncbi.nlm.nih.gov/BLAST/) and found that a DNA polymerase gene (EHI_164190) shows a perfect sense-antisense match with this query, implying the possible regulatory role of this putative miRNA. Against the Pfam database, this DNA polymerase gene possesses the DNA polymerase type B domain which has significant sequence similarity with a known viral polymerase domain [187]. So, this domain sequence hit suggests that this putative miRNA potentially plays an adaptive role in inhibiting overexpression of virus-derived genes. However, in animals, one miRNA molecule can have

a vast variety of different mRNA targets by partial complementary base pairing between the seed region (6-8 nt) of the 5' end of miRNA and the target sequence. Moreover, one particular target might be regulated by different species of miRNAs [294,301,302]. Therefore, this partial complementarity and combinatorial nature of miRNA regulation make the miRNA target prediction more complicated than previously thought.

Hence, in this study I have identified novel miRNA candidates from the small RNA-Seq data and demonstrated the presence of one putative miRNA named 'miR-Rah1' in the transcriptomes of *E. histolytica* strains using the qPCR analysis. Taken together with the evidence of RNAi-associated machinery proteins, these experimental findings suggest that miRNA-based regulation exists in *E. histolytica* and potentially play a role in modulating parasite gene expression.

Table 5.13: Details of miRNA candidates predicted by the miRDeep2 software. The loop nucleotide sequences are highlighted in the grey box.

miRNA candidate	miRDeep2 score	estimated probability that the miRNA candidate is a true positive	total read count	mature read count	loop read count	star read count	significant randfold *P*-value	precursor coordinate	Sequences (consensus mature sequence, consensus star sequence and consensus precursor sequence)
Rahman_1 library									
miR-Rah1	1.4	0.94 ± 0.14	30	30	0	0	**yes**	DS571214: 4948-5003 (+) and DS571763: 3032-3087 (-)	**Consensus mature sequence** (24 nt): 5’ agauggauuagaaaagacgguugu 3’
									Consensus star sequence (25 nt): 5’ uuccaucuuuucauaauccuucuua 3’
									Consensus precursor sequence (55 nt): **MFE = -73.05 kJ·mol^{-1}** 5’agauggauuagaaaagacgguuguuguuuuuuccaucuuuucauaauccuucuua 3’
Rahman_2 library									
miR-Rah2	1.6	0.51 ± 0.50	2	1	0	1	**yes**	DS571259: 13503-13548 (-)	**Consensus mature sequence** (20 nt): 5’ gggcuguaggacuauugacu 3’
									Consensus star sequence (20 nt): 5’ uauauugcugguccuacauc 3’
									Consensus precursor sequence (45 nt): **MFE = -61.17 kJ·mol^{-1}** 5’ uauauugcugguccuacaucaagugggcuguaggacuauugacu 3’
PVBM08B_2 library									
miR-PVB2	0.7	0.63± 0.49	2	1	0	1	**yes**	DS571345: 26315-26363 (+)	**Consensus mature sequence** (20 nt): 5’ ugauagucguaaauguuaua 3’
									Consensus star sequence (23 nt): 5’ caauguuuauggcaugucugaua 3’
									Consensus precursor sequence (48 nt): **MFE = -49.33 kJ·mol^{-1}** 5’ugauagucguaaauguuauacaaaacaauguuuauggcaugucugaua 3’

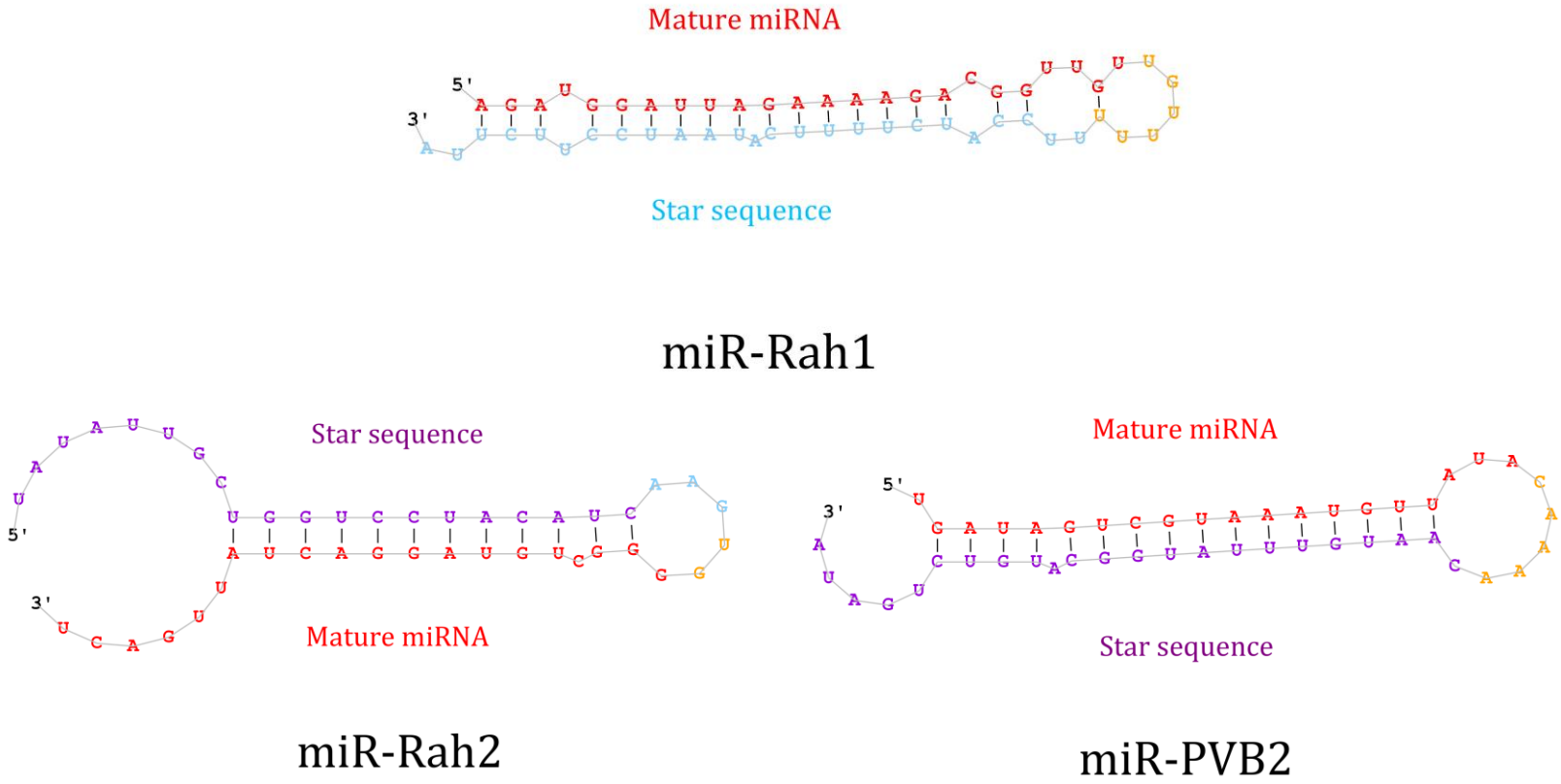


Figure 5.31: Predicted secondary structures of the novel miRNA candidate precursors obtained by miRDeep2 analysis. The relative nucleotide positions of the mature miRNA strand, star sequence and loop portion in the pre-miRNA hairpin structure are represented by different colours as illustrated in the figure. The mimimal free energy values are -73.05, -61.17 and -49.33 kJ·mol^{-1} for pre-miRNA hairpins of miR-Rah1, miR-Rah2 and miR-PVB2, respectively.

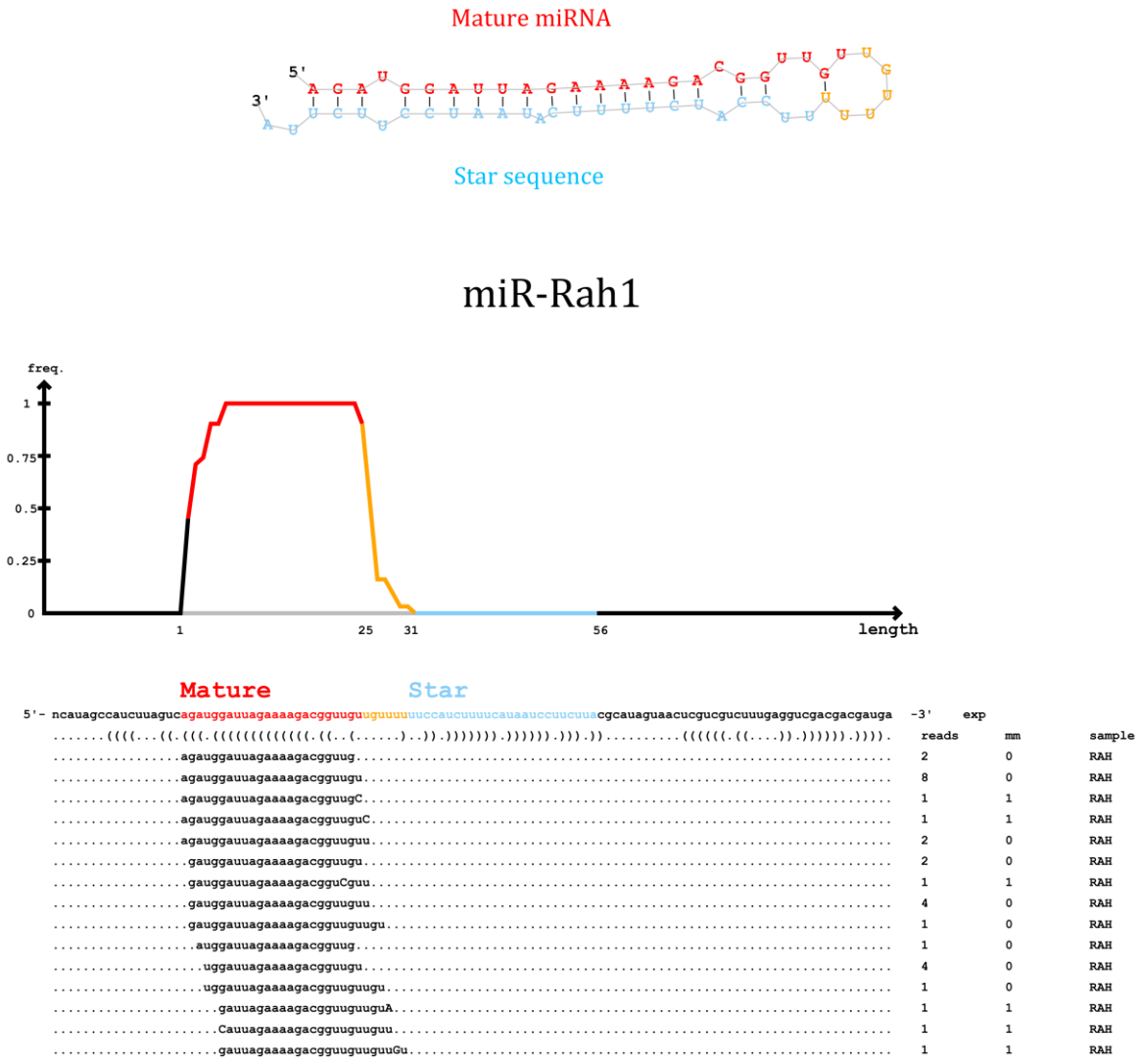


Figure 5.32: The 1st predicted pre-miRNA precursor structure, miR-Rah1, with its mature miRNA (red) and star sequences (sky blue). MiR-Rah1 is 24 nt in length (5' AGAUGGAUUAGAAAAGACGGUUGU 3') and located on two scaffolds: DS571214: 4948-5003 (+) and DS571763: 3032-3087 (-). The number of sequencing reads aligned to the particular genomic position is reported as above.

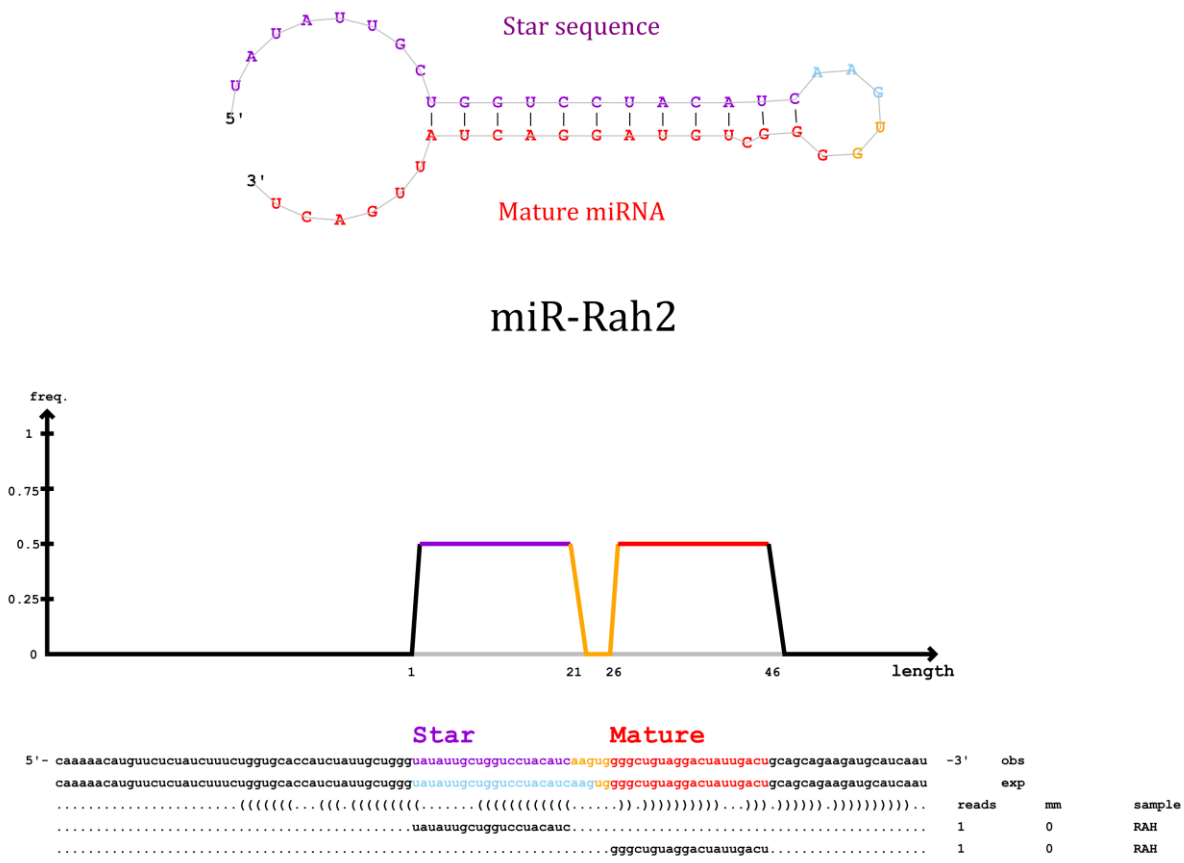


Figure 5.33: The 2nd predicted pre-miRNA precursor structure, miR-Rah2, with its mature miRNA (red) and star sequences (violet). MiR-Rah2 is 20 nt in length (5' GGGCUGUAGGACUAUUGACU 3') and located on scaffold DS571259: 13503-13548 (-). The number of sequencing reads aligned to the particular genomic position is reported as above.

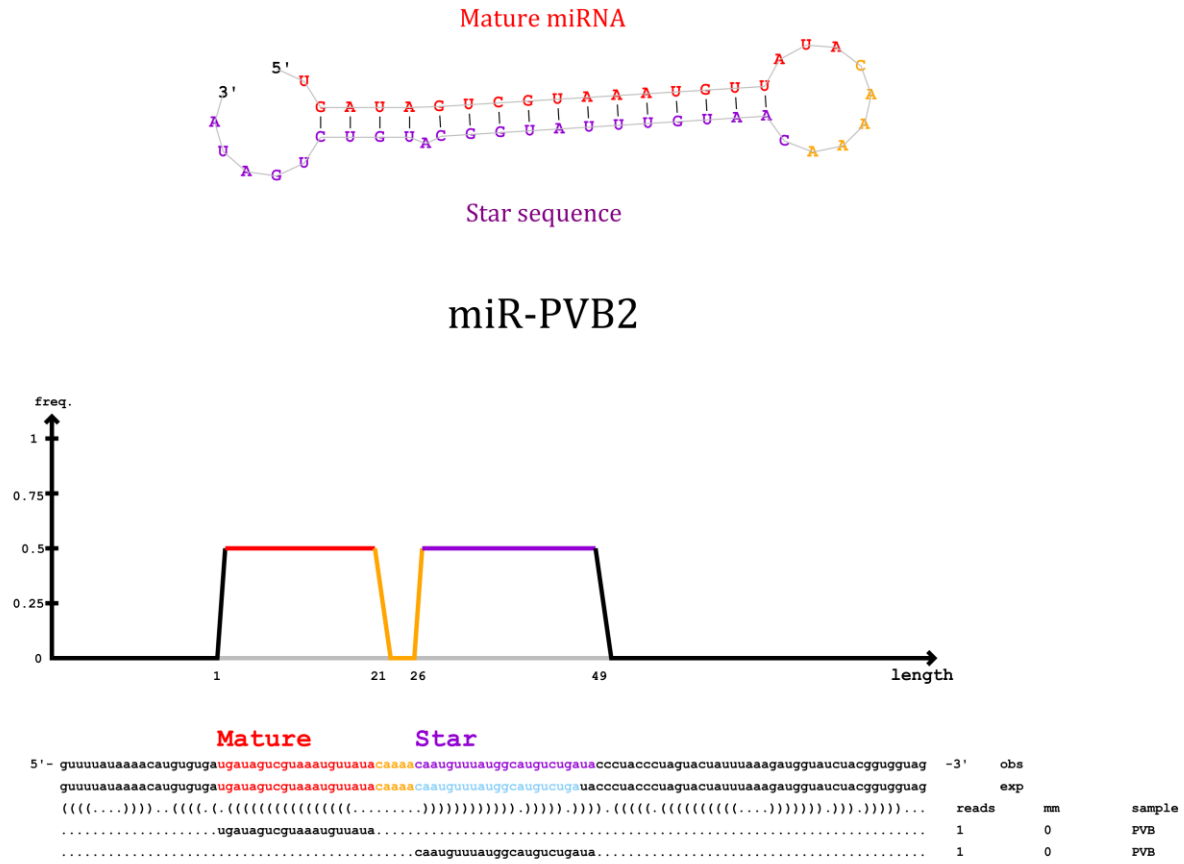


Figure 5.34: The 3rd predicted pre-miRNA precursor structure, miR-PVB2, with its mature miRNA (red) and star sequences (violet). MiR-PVB2 is 20 nt in length (5' UGAUAGUCGUAAAUGUUAUA 3') and located on scaffold: DS571345: 26315-26363 (+). The number of sequencing reads aligned to the particular genomic position is reported as above.

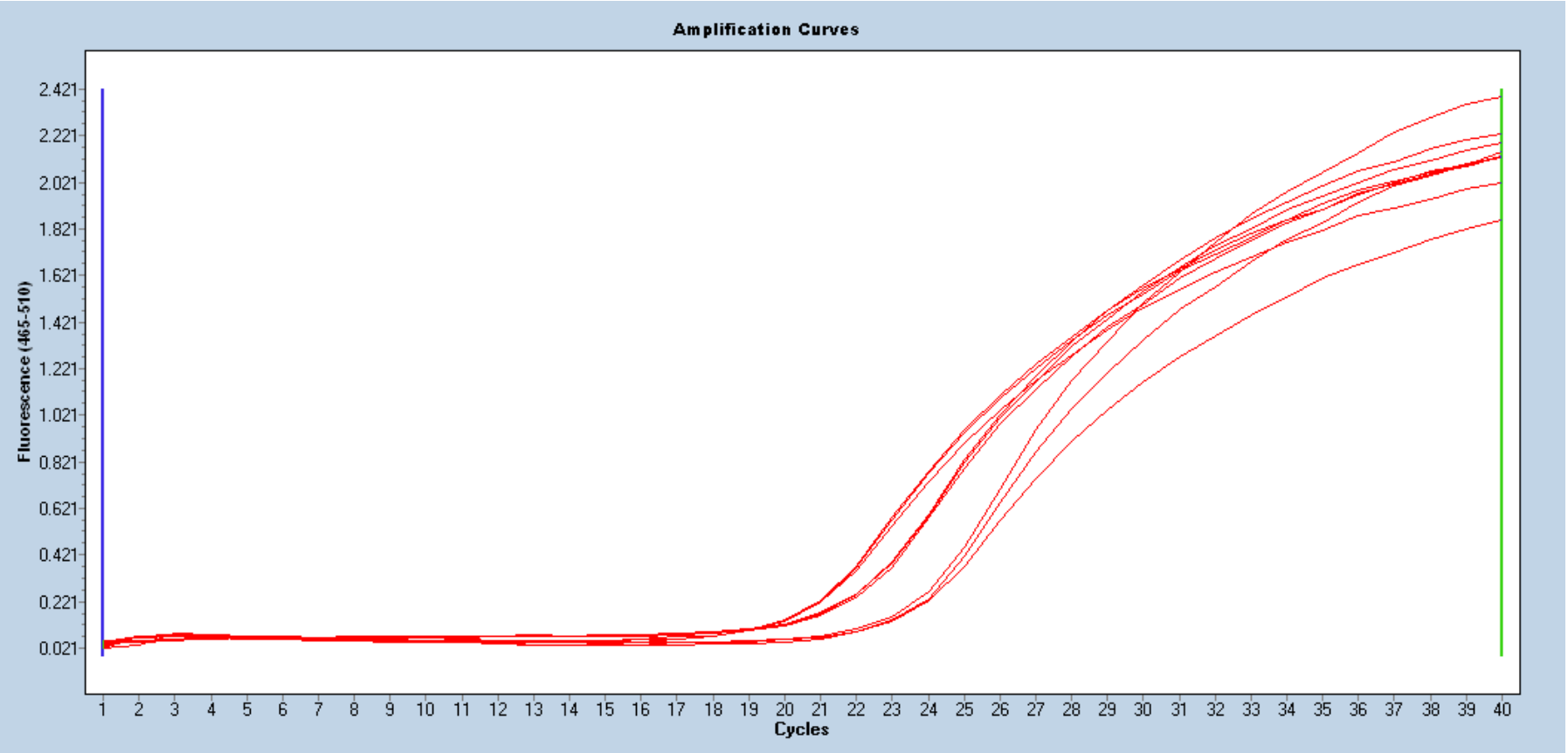


Figure 5.35: qPCR amplification curve for validation of miR-Rah1 candidate expression in three *E. histolytica* strains (i.e. Rahman, PVBM08B and HM-1:IMSS).

Table 5.14: **Details of crossing point (Cp) and standard deviation (SD Cp) in each RNA sample group.**

Total RNA sample	Crossing point (Cp)	Mean Cp ± SD Cp
Rahman		
Rahman_1	20.54	
Rahman_2	20.75	20.66 ± 0.11
Rahman_2	20.69	
PVBM08B		
PVBM08B_1	20.2	
PVBM08B_2	20.19	20.17 ± 0.04
PVBM08B_3	20.13	
HM-1:IMSS		
HM-1:IMSS_1	23.2	
HM-1:IMSS_2	22.9	23.09 ± 0.17
HM-1:IMSS_3	23.17	
IULA:1092:1		
IULA:1092:1_1	N/A	
IULA:1092:1_2	N/A	N/A
IULA:1092:1_3	N/A	

N.B. IULA samples were not sufficient for this qPCR experiment due to their very low concentration.

5.4 Concluding remarks

This present study demonstrates the comparative analysis of small RNA expression in the transcriptomes of four laboratory cultured *E. histolytica* strains using the Illumina MiSeq technology. The size-fractionated sRNA sequencing data unveil the contrary relationship between target gene expression levels and antisense sRNA abundance, strongly suggesting the regulatory function of antisense sRNAs. It is intriguing that members of the multigene families, encoding BspA-like LRRPs, PKs, Rab family GTPases and RhoGAP domain-containing proteins, constitute a large fraction of total number of sRNA target genes in all four strains, implying that antisense sRNAs potentially facilitate in switching on/off gene expression of these multigene family members in *E. histolytica*.

Furthermore, I found that the differential sRNA regulation of virulence-associated gene expression, i.e. EhSTIRPs, BspA-like LRRPs and C2 domain-containing proteins, occurs among *E. histolytica* strains, indicating that sRNA-mediated post-transcriptional regulation may be important in shaping the parasite virulence in *E. histolytica*. Consistent with previous publications, the findings of this study indicate that antisense sRNAs are likely to downregulate expression of virulence-associated genes in a strain-specific manner through the siRNA pathway [85,86,92].

Due to a limited number of sRNA target genes, it can be inferred that substantial global transcriptomic variability between *E. histolytica* strains is as a result of combinatorial regulation including gene copy number differences, sequence polymorphisms as well as epigenetic processes such as sRNA-mediated silencing, DNA methylation and histone modification. In addition, this study identified the novel putative miRNA from the sRNA sequencing data using the biogenesis-based bioinformatic analysis and qPCR validation, implying that miRNA potentially play a regulatory role in *E. histolytica*.

6.1 Overall perspectives

Amoebiasis remains to be a challenge of the global public health issues. In 1980s, the global mortality was estimated up to 100,000 deaths per year as a third rank after malaria and schistosomiasis [3,81]. Recombinant vaccines against the parasite Gal/GalNAc lectin have been developed and found to be most promisingly protective against both intestinal and hepatic amoebiasis in animal models, however clinical trials for efficacy in humans need to be determined [81,303]. As *E. histolytica* infection occurs worldwide, especially in many developing countries and the spectrum of clinical manifestations ranges from asymptomatic carriers to extraintestinal amoebic abscess, genome-wide characterisation of virulence-associated genes and pathways in the *E. histolytica* strains may advance the understanding of molecular mechanisms of virulence modulation which could be applied to predict the disease prognosis and improve treatment strategies as well as reduce the transmission rate. As such, this project was designed to explore some of genomic and transcriptomic differences between *E. histolytica* strains as well as investigate how the trophozoites regulate levels of gene expression towards their differential virulence.

In this study, the transcriptomic profilings were done in the four axenically cultured laboratory-adapted strains as an axenic parasite culture can provide a substantial amount of parasite RNA transcripts, without any bacterial RNA contamination, suitable for downstream analysis. Moreover, different parasite strains were cultured with the same axenic culture media and conditions, therefore this *in vitro* study enables us to rule out other influencing factors, which are commonly found in the *in vivo* condition such as microbiotic interaction, non-enriched intestinal environment, host immune attack, etc.

6.2 Modulations of gene expression in *in vitro* and *in vivo* are associated with differential virulence between *E. histolytica* strains

As previously published, the substantial changes in the transcriptome were observed in trophozoites isolated from infected colon, compared to *in vitro* culture [77,82]. Capability to adapt in the host intestinal environment was accomplished by increased expression of some signaling genes such as transmembrane kinases, Ras and Rho family GTPases as well as calcium-binding proteins. Additionally, transcriptional modulations of genes involved in energy metabolism, signal transduction, bacterial killing, DNA binding as well as virulence were found in trophozoites in the *in vivo* infection [77,82].

Correspondingly, characterising transcriptomic landscapes of the four laboratory-adapted strains by RNA-Seq in this study can reveal considerable differences of expression of gene members involved in signal transduction, actin cytoskeleton dynamics, proteasomal degradation, DNA binding process as well as response to stress between nonvirulent and virulent strains and also provide the evolutionary explanation for the transcriptomic diversity in relation to their genetic differences (i.e. SNPs, CNVs and gene gain or gene loss) as previously discussed in Chapter 2 and 4.

Based on functional characterisation, the majority of the upregulated and downregulated protein domains in the three virulent strains are considerably different. Most members of prevalent upregulated domains are implicated in various functions including signal transduction, actin dynamics, protein degradation, protein-protein interaction, transcriptional control and phagocytosis whilst downregulated domains are mainly involved in signal transduction, protein-protein interaction and transcriptional control. As such, the increased virulence phenotype of parasites requires a vast variety of cellular functions as well as selectively transcriptional control of genes involved in virulence. Also, this implies that different fates of signal transduction as well as protein-protein interaction seem to be a key determining factor for different networks of cellular functions among the strains, resulting in difference in virulence.

Essentially, the genome-wide analyses in this study can point out that differences in constitutive expression profiles among the *E. histolytica* strains are associated with their virulence variability. Therefore, this promising transcriptomic data would enable researchers to precisely target virulence-associated genes and associated pathways as well as to study the effect of host environmental stimuli in modulating expression of such genes by comparing with their constitutive expression levels.

6.3 Genomic plasticity and sRNA-mediated regulation are important mechanisms of virulence modulation in *E. histolytica*

As demonstrated in Chapter 2, it is noteworthy that sequence polymorphisms of genes involved in the host-parasite interaction is significantly correlated with the variation in expression levels among strains, reflecting that the nucleotide changes under positive selection would contribute to the transcriptional variability due to the possible alterations in the binding of transcriptional factors and associated regulatory elements. However, it seems to be that transcriptional variation due to sequence divergence is limited as overall single nucleotide diversity is rather low throughout the parasite genome [70].

Contrastedly, complete genome sequencing of *E. histolytica* strains unveiled large differences in gene copy numbers among the genomes, suggesting that a high degree of genomic plasticity and variation in the number of gene family members potentially result in transcriptomic variation across the strains [70]. Expectedly, the genome-wide correlation study in Chapter 4 of this study reveals the positive relationship between gene copy number variation and transcriptomic variability, strongly suggesting that variation in gene copy number is likely to be a key regulation of gene expression levels among the parasite strains. Consistently, gene copy number variation is common in other human protozoan parasites such as *Trypanosomes* and *Leishmania*, and found to be associated with different biological attributes among parasite strains, suggesting that copy number variation is a potentially important mechanism in generating genomic diversity and transcriptional modulation of gene expression in almost exclusively asexual parasite group [269-271,280,281] .

Besides the effect of genomic diversity, transcriptomic changes are potentially determined in part by the host environmental stimuli. As previously *in vivo* studied by Gilchrist *et al.*, 2006, genes accounting for ~5.2% of the genome were found to be modulated in the transcriptomes of HM-1:IMSS trophozoites isolated from the mice colon on Days 1 and 29 after inoculation, implying that trophozoites have regulations of gene expression in both short-term and long-term responses to host stimuli [77]. Taken together, it can be inferred that global transcriptomic variability in *E. histolytica* strains are mainly influenced in a combinatorial manner by both the genomic variation and the external host stimuli as shown in Figure 6.

Moreover, the 5th chapter in this study demonstrates the sRNA-mediated gene regulation towards differential virulence in *E. histolytica* strains, indicating that virulence is determined in part at the post-transcriptional level. Using the biogenesis-based bioinformatic analysis, the novel putative miRNA candidates (miR-Rah1, miR-Rah2 and miR-PVB2) were also predicted, suggesting the possible role in regulating parasite gene expression. Ultimately, the experimental findings in this present study strongly indicate that genomic plasticity and sRNA-mediated regulation are important cellular mechanisms of virulence modulation in *E. histolytica* parasite.

6.4 Future plan

As previously published, addition of polyadenylated sequence to the 3' end of siRNAs by *in vitro* transcription can apparently increase their gene silencing activity in MCF-7 breast cancer cell line, suggesting a possible role of 3'-polyadenylated sequence in the RNAi pathway [304]. Recently, 3'-polyadenylated antisense sRNAs were discovered and found to be associated with stronger silencing effect in *E. histolytica* (personal communication, Singh *et al*.). Consistent with the findings of Singh *et al*., antisense sRNA population of the four strains in this study could be divided into two distinct groups as illustrated in Figures 5.18A, 5.19A, 5.20A and 5.21A of Chapter 5. As such, it could be postulated that these two subpopulations of antisense sRNAs possibly possess different length of 3'-polyadenylated tail, resulting in their different silencing efficiency. However, the sRNA libraries in this present study were size-fractionated prior to sequencing, providing a narrow peak size of sequenced reads (23 to 28 nt). Thus, the next research project will explore the length distribution of 3'-polyadenylated tail in a larger-sized sRNA population and determine whether difference in the silencing activity of antisense sRNAs is directly related to their 3' tail length. This investigation will help us understand post-transcriptional gene regulatory mechanisms in *E. histolytica* more throughly.

This study also identified one novel putative miRNA, i.e. miR-Rah1, expressed in *E. histolytica* strains, suggesting that miRNA-based regulation potentially facilitate transcriptomic modulation in this parasite. Interestingly, miR-Rah1 shows perfect complementarity with a gene encoding viral-type DNA polymerase (EHI_164190), suggesting that *E. histolytica* potentially has molecular defence mechanisms for inhibiting viral replication. However, miRNA target prediction requires a specific tool with specialised algorithm due to the complexity of miRNA combinatorial regulation as mentioned before [294,301,302]. Therefore, accurate miRNA target prediction will be another future step that can provide the advanced knowledge of post-transcriptional regulation towards the pathogenesis and virulence in this parasite.

Figure 6: Interrelationship between genome diversity and transcriptomic difference and host environmental stimuli.

References

[1] Espinosa-Cantellano M, Martinez-Palomo A: **Pathogenesis of intestinal amoebiasis: from molecules to disease.** *Clinical Microbiology Reviews* 2000, **13**: 318-331.

[2] Weinke T, Friedrich-Jänicke B, Hopp P, Janitschke K: **Prevalence and clinical importance of *Entamoeba histolytica* in two high-risk groups: travelers returning from the tropics and male homosexuals.** *J Infect Dis* 1990, **161**: 1029-1031.

[3] Walsh JA: **Problems in recognition and diagnosis of amebiasis: estimation of the global magnitude of morbidity and mortality.** *Rev Infect Dis* 1986, **8**: 228–238.

[4] Mitra BN, Pradel G, Frevert U, Eichinger D: **Compounds of the upper gastrointestinal tract induce rapid and efficient excystation of *Entamoeba invadens*.** *Int J Parasitol* 2010, **40**: 751-760.

[5] The Centers for Disease Control and Prevention (CDC; 2010): **www.cdc.gov/parasites/amebiasis**

[6] Tilak KVGK: **Lecture on Amoebiasis**, 2nd February, APICON2013 Conference, Coimbatore, Tamil Nadu, India, 2013.

[7] Moncada D, Keller K, Chadee K: ***Entamoeba histolytica*-secreted products degrade colonic mucin oligosaccharides.** *Infect Immun* 2005, **73**: 3790-3793.

[8] Lejeune M, Rybicka JM, Chadee K: **Recent discoveries in the pathogenesis and immune response toward *Entamoeba histolytica*.** *Future Microbiol* 2009, **4**: 105-118.

[9] Lidell ME, Moncada DM, Chadee K, Hansson GC: ***Entamoeba histolytica* cysteine proteinases cleave the MUC2 mucin in its C-terminal domain and dissolve the protective colonic mucus gel.** *Proc Natl Acad Sci U S A* 2006, **103**: 9298-9303.

[10] Que X, Reed SL: **Cysteine proteinases and the pathogenesis of amebiasis.** *Clin Microbiol Rev* 2000, **13**: 196–206.

[11] MacFarlane RC, Singh U: **Identification of an *Entamoeba histolytica* serine-, threonine-, and isoleucine-rich protein with roles in adhesion and cytotoxicity.** *Eukaryot Cell* 2007, **6**: 2139-2146.

[12] Leroy A, Lauwaet T, De Bruyne G, Cornelissen M, Mareel M: ***Entamoeba histolytica* disturbs the tight junction complex in human enteric T84 cell layers.** *FASEB J* 2000, **14**: 1139-1146.

[13] Sharma M, Vohra H, Bhasin D: **Enhanced pro-inflammatory chemokine/cytokine response triggered by pathogenic *Entamoeba histolytica* : basis of invasive disease.** *Parasitology* 2005, **131**: 783-796.

[14] Guo X, Houpt E, Petri WA Jr: **Crosstalk at the initial encounter: interplay between host defense and ameba survival strategies.** *Curr Opin Immunol* 2007, **19**: 376-384.

[15] Choi MH, Sajed D, Poole L, Hirata K, Herdman S, Torian BE, Reed SL: **An unusual surface peroxiredoxin protects invasive *Entamoeba histolytica* from oxidant attack.** *Mol Biochem Parasitol* 2005, **143**: 80-89.

[16] Davis PH, Zhang X, Guo J, Townsend RR, Stanley SL Jr: **Comparative proteomic analysis of two *Entamoeba histolytica* strains with different virulence phenotypes identifies peroxiredoxin as an important component of amoebic virulence.** *Mol Microbiol* 2006, **61**: 1523-1532.

[17] Bruchhaus I, Tannich E: **Induction of the iron-containing superoxide dismutase in *Entamoeba histolytica* by a superoxide anion-generating system or by iron chelation.** *Mol Biochem Parasitol* 1994, **67**: 281-288.

[18] Choi MH, Sajed D, Poole L, Hirata K, Herdman S, Torian BE, Reed SL: **An unusual surface peroxiredoxin protects invasive *Entamoeba histolytica* from oxidant attack.** *Mol Biochem Parasitol* 2005, **143**: 80-89.

[19] Lo HS, Reeves RE: **Purification and properties of NADPH:flavin oxidoreductase from *Entamoeba histolytica*.** *Mol Biochem Parasitol* 1980, **2**: 23-30.

[20] Zhang X, Zhang Z, Alexander D, Bracha R, Mirelman D, Stanley SL Jr: **Expression of amoebapores is required for full expression of *Entamoeba histolytica* virulence in amebic liver abscess but is not necessary for the induction of inflammation or tissue damage in amebic colitis.** *Infect Immun* 2004, **72**: 678-683.

[21] Tillack M, Nowak N, Lotter H, Bracha R, Mirelman D, Tannich E, Bruchhaus I: **Increased expression of the major cysteine proteinases by stable episomal transfection underlines the important role of EhCP5 for the pathogenicity of *Entamoeba histolytica*.** Mol Biochem *Parasitol* 2006, **149**: 58-64.

[22] Santi-Rocca J, Weber C, Guigon G, Sismeiro O, Coppée JY, Guillén N: **The lysine- and glutamic acid-rich protein KERP1 plays a role in *Entamoeba histolytica* liver abscess pathogenesis.** *Cell Microbiol* 2008, **10**: 202-217.

[23] Nickel R, Ott C, Dandekar T, Leippe M: **Pore-forming peptides of *Entamoeba dispar*. Similarity and divergence to amoebapores in structure, expression and activity.** *Eur J Biochem* 1999, **265:** 1002-1007.

[24] Clark CG, Alsmark UC, Tazreiter M, Saito-Nakano Y, Ali V, Marion S, *et al.*: **Structure and content of the *Entamoeba histolytica* genome.** *Adv Parasitol* 2007, **65**: 51-190.

[25] Lorenzi HA, Puiu D, Miller JR, Brinkac LM, Amedeo P, Hall N, Caler EV: **New assembly, reannotation and analysis of the *Entamoeba histolytica* genome reveal new genomic features and protein content information.** *PLoS Negl Trop Dis* 2010, **4**: e716.

[26] Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Caler EV, *et al.*: **AmoebaDB and MicrosporidiaDB: functional genomic resources for *Amoebozoa* and *Microsporidia* species.** *Nucleic Acids Res* 2011, **39**: D612–D619.

[27] Clark CG, Ali IK, Zaki M, Loftus BJ, Hall N: **Unique organisation of tRNA genes in *Entamoeba histolytica*.** *Mol Biochem Parasitol* 2006, **146**: 24-29.

[28] Tawari B, Ali IK, Scott C, Quail MA, Berriman M, Hall N, Clark CG: **Patterns of evolution in the unique tRNA gene arrays of the genus *Entamoeba*.** *Mol Biol Evol* 2008, **25**: 187–198.

[29] Das K, Ganguly S: **Evolutionary genomics and population structure of *Entamoeba histolytica.*** *Comput Struct Biotechnol J* 2014, **12**: 26–33.

[30] Zermeno V, Ximenez C, Moran P, Valadez A, Valenzuela O, Rascon E: **Worldwide genealogy of *Entamoeba histolytica*: an overview to understand haplotype distribution and infection outcome.** *Infect Genet Evol* 2013, **17**: 243–252.

[31] Ali IK, Zaki M, Clark CG: **Use of PCR amplification of tRNA gene-linked short tandem repeats for genotyping *Entamoeba histolytica*.** *J Clin Microbiol* 2005, **43**: 5842-5847.

[32] Ali IK, Mondal U, Roy S, Haque R, Petri WA Jr, Clark CG: **Evidence of a link between parasite genotype and outcome of infection with *Entamoeba histolytica*.** *J Clin Microbiol* 2007, **45**: 285-289.

[33] Ali IK, Solaymani-Mohammadi S, Akhter J, Roy S, Gorrini C: **Tissue invasion by *Entamoeba histolytica*: evidence of genetic selection and/or DNA reorganization events in organ tropism.** *PLoS Negl Trop* 2008, **2**: e219.

[34] Feng M, Cai J, Yang B, Fu Y, Min X: **Unique short tandem repeat nucleotide sequences in *Entamoeba histolytica* isolates from China.** *Parasitol Res* 2012, **111**: 1137–1142.

[35] Kumari V, Sharma R, Yadav VP, Gupta AK, Bhattacharya A, Bhattacharya S: **Differential distribution of a SINE element in the *Entamoeba histolytica* and *Entamoeba dispar* genomes: Role of the LINE-encoded endonuclease.** *BMC Genomics* 2011, **12**: 267.

[36] Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, Amedeo P, *et al.*: **The genome of the protist parasite *Entamoeba histolytica*.** *Nature* 2005, **433**: 865-868.

[37] Mukherjee C, Clark CG, Lohia A: ***Entamoeba* shows reversible variation in ploidy under different growth conditions and between life cycle phases.** *PLoS Negl Trop Dis* 2008, **2**: e281.

[38] Weedall GD, Sherrington J, Paterson S, Hall N: **Evidence of gene conversion in genes encoding the Gal/GalNac lectin complex of *Entamoeba*.** *PLoS Negl Trop Dis* 2011, **5**: e1209.

[39] Weedall GD, Hall N: **Sexual reproduction and genetic exchange in parasitic protists.** *Parasitology* 2015, **142**: S120-S127.

[40] Brumpt E: **Étude sommaire de l' *"Entamoeba dispar"* n. sp. Amibe à kystes quadrinucléés, parasite de l'homme.** *Bull Acad Méd (Paris)* 1925, **94**: 943-952.

[41] Walker EL, Sellards AW: **Experimental entamoebic dysentery.** *Philippine J Sci B Trop Med* 1913, **8**: 253–331.

[42] Sargeaunt PG, Williams JE, Grene JD: **The differentiation of invasive and non-invasive *Entamoeba histolytica* by isoenzyme electrophoresis.** *Trans R Soc Trop Med Hyg* 1978, **72**: 519-521.

[43] Strachan WD, Chiodini PL, Spice WM, Moody AH, Ackers JP: **Immunological differentiation of pathogenic and non-pathogenic isolates of *Entamoeba histolytica*.** *Lancet* 1988, **1**: 561-563.

[44] Tannich E, Horstmann RD, Knobloch J, Arnold HH: **Genomic DNA differences between pathogenic and nonpathogenic *Entamoeba histolytica*.** *Proc Natl Acad Sci U S A* 1989, **86**: 5118–5122.

[45] Levecke B, Dreesen L, Dorny P, Verweij JJ, Vercammen F, *et al.*: **Molecular identification of *Entamoeba spp.* in captive nonhuman primates.** *J Clin Microbiol* 2010, **48**: 2988-2990.

[46] Weedall GD and Hall N: **Evolutionary genomics of *Entamoeba.*** *Res Microbiol* 2011, **162**: 637–645.

[47] Bruchhaus I, Jacobs T, Leippe M, Tannich E**: *Entamoeba histolytica* and *Entamoeba dispar*: differences in numbers and expression of cysteine proteinase genes.** *Mol Microbiol* 1996, **22:** 255-263.

[48] Willhoeft U, Hamann L, Tannich E: **A DNA sequence corresponding to the gene encoding cysteine proteinase 5 in *Entamoeba histolytica* is present and positionally conserved but highly degenerated in *Entamoeba dispar*.** *Infect Immun* 1999, **67**: 5925-5929.

[49] Bruchhaus I, Loftus BJ, Hall N, Tannich E: 2003. **The intestinal protozoan parasite *Entamoeba histolytica* contains 20 cysteine proteinase genes, of which only a small subset is expressed during in vitro cultivation.** *Eukaryot Cell* 2003, **2:** 501-509.

[50] Perdomo D, Baron B, Rojo-Domı'nguez A, Raynal B, England P: **The α-helical regions of KERP1 are important in *Entamoeba histolytica* adherence to human cells.** *Nat Sci Reports* 2013, **3**: 1171.

[51] Dolabella SS, Serrano-Luna J, Navarro-García F, Cerritos R, Ximénez C: **Amoebic liver abscess production by *Entamoeba dispar*.** *Ann Hepatol* 2012, **11**: 107–117.

[52] Galvan-Moroyoqui JM, Del Carmen Dominguez-Robles M, Franco E, Meza I: **The Interplay between *Entamoeba* and Enteropathogenic Bacteria Modulates Epithelial Cell Damage.** *PLos Negl Trop Dis* 2008, **2**: e266.

[53] Mirelman D, Feingold C, Wexler A, Bracha R: **Interactions between *Entamoeba histolytica*, bacteria and intestinal cells.** *Ciba Found Symp* 1983, **99**: 2–30.

[54] Mirelman D, Bracha R, Chayen A, Aust-Kettis A, Diamond LS: ***Entamoeba histolytica*: effect of growth conditions and bacterial associates on isoenzyme patterns and virulence.** *Exp Parasitol* 1986, **62**: 142-148.

[55] Tshalaia LE: **On a species of *Entamoeba* detected in sewage effluents.** *Med Parazit (Moscow)* 1941, **10:** 244-252.

[56] Dreyer DA: **Growth of a strain of *Entamoeba histolytica* at room temperature.** *Tex Rep Biol Med* 1961, **19**: 393-396.

[57] Clark CG, Diamond LS: **The Laredo strain and other *Entamoeba histolytica*-like amoebae are *Entamoeba moskovskii*.** *Mol Biochem Parasitol* 1991, **46**: 11-18.

[58] Shimokawa C, Kabir M, Taniuchi M, Mondal D, Kobayashi S: ***Entamoeba moshkovskii* is associated with diarrhea in infants and causes diarrhea and colitis in mice.** *J Infect Dis* 2012, **206**: 744–751.

[59] Moonah SN, Jiang NM, Petri WA Jr: **Host immune response to intestinal amebiasis.** *PLoS Pathog* 2013, **9**: e1003489.

[60] Carrero JC, Cervantes-Rebolledo C, Aguilar-Díaz H, Díaz-Gallarado MY, Laclette JP, Morales-Montor J: **The role of the secretory immune response in the infection by *Entamoeba histolytica.*** *Parasite Immunol* 2007, **29**: 331-338.

[61] Haque R, Huston CD, Hughes M, Houpt E, Petri WA Jr: **Amebiasis.** *N Engl J Med* 2003, **348**: 1565-1573.

[62] Reed SL, Sargeaunt PG, Braude AI: **Resistance to lysis by human serum of pathogenic *Entamoeba histolytica*.** *Trans R Soc Trop Med Hyg* 1983, **77**: 248-253.

[63] Acuna-Soto R, Maguire JH, Wirth DF: **Gender distribution in asymptomatic and invasive amebiasis.** *Am J Gastroenterol* 2000, **95**: 1277-1283.

[64] Petri WA Jr, Mondal D, Peterson KM, Duggal P, Haque R: **Association of malnutrition with amebiasis.** *Nutr Rev* 2009, **67**: S207-S215.

[65] Duggal P, Guo X, Haque R, Peterson KM, Ricklefs S, *et al*: **A mutation in the leptin receptor is associated with *Entamoeba histolytica* infection in children.** *J Clin Invest* 2011, **121**: 1191-1198.

[66] Ravdin JI**,** Guerrant RL**: Role of adherence in cytopathogenic mechanisms of *Entamoeba histolytica.* Study with mammalian tissue culture cells and human erythrocytes.** *J Clin Invest* 1981, **68**: 1305-1313.

[67] Chadee K, Meerovitch E: **The pathogenesis of experimentally induced amebic liver abscess in the gerbil (Meriones unguiculatus).** *Am J Pathol* 1984, **117**: 71-80.

[68] Trissl D, Martínez-Palomo A, de la Torre M, de la Hoz R, Pérez de Suárez E: **Surface properties of *Entamoeba*: increased rates of human erythrocyte phagocytosis in pathogenic strains.** *J Exp Med* 1978, **148**: 1137-1143.

[69] Araujo J, García ME, Díaz-Suárez O, Urdaneta H: **Amebiasis: importance of the diagnosis and treatment. Minireview.** *Invest Clin* 2008, **49**: 265-271.

[70] Weedall GD, Clark CG, Koldkjaer P, Kay S, Bruchhaus I, Tannich E, Paterson S, Hall N: **Genomic diversity of the human intestinal parasite *Entamoeba histolytica.*** *Genome Biol* 2012, **13**: R38.

[71] Mattern CF, Keister DB: **Experimental amebiasis. II. Hepatic amebiasis in the newborn hamster.** *Am J Trop Med Hyg* 1977, **26**: 402-411.

[72] Burchard GD, Mirelman D: ***Entamoeba histolytica*: virulence potential and sensitivity to metronidazole and emetine of four isolates possessing nonpathogenic zymodemes.** *Exp Parasitol* 1988, **66**: 231-242.

[73] Vicente JB, Ehrenkaufer GM, Saraiva LM, Teixeira M, Singh U: ***Entamoeba histolytica* modulates a complex repertoire of novel genes in response to oxidative and nitrosative stresses: implications for amebic pathogenesis.** *Cell Microbiol* 2009, **11**: 51-69.

[74] MacFarlane RC, Singh U: **Identification of differentially expressed gene in virulent and nonvirulent *Entamoeba* species: potential implications for amebic pathogenesis.** *Infect Immun* 2006, **74**: 350-351.

[75] Shah PH, MacFarlane RC, Bhattacharya D, Matese JC, Demeter J, *et al.*: **Comparative genomic hybridizations of *Entamoeba* strains reveal unique genetic fingerprints that correlate with virulence.** *Eukaryot Cell* 2005, **4**: 504-515.

[76] Davis PH, Schulze J, Stanley SL Jr: **Transcriptomic comparison of two *Entamoeba histolytica* strains with defined virulence phenotypes identifies new virulence factor candidates and key differences in the expression patterns of cysteine proteinases, lectin light chains, and calmodulin.** *Mol Biochem Parasitol* 2007, **151**: 118-128.

[77] Gilchrist CA, Houpt E, Trapaidze N, Fei Z, Crasta O, *et al.*: **Impact of intestinal colonization and invasion on the *Entamoeba histolytica* transcriptome.** *Mol Biochem Parasitol* 2006, **147**: 163-176.

[78] Petri WA Jr, Haque R, Mann BJ: **The bittersweet interface of parasite and host: lectin-carbohydrate interactions during human invasion by the parasite *Entamoeba histolytica*.** *Annu Rev Microbiol* 2002, **56**: 39-64.

[79] Ankri S, Padilla-Vaca F, Stolarsky T, Koole L, Katz U, Mirelman D: **Antisense inhibition of expression of the light subunit (35 kDa) of the Gal/GalNac lectin complex inhibits *Entamoeba histolytica* virulence.** *Mol Microbiol* 1999, **33**: 327–337.

[80] Katz U, Ankri S, Stolarsky T, Nuchamowitz Y, Mirelman D: ***Entamoeba histolytica* expressing a dominant negative N-truncated light subunit of its gal-lectin are less virulent.** *Mol Biol Cell* 2002, **13**: 4256-4265.

[81] Ravdin JI (ed.): **Amebiasis. Volume 2 in *Tropical Medicine: Science and Practice*, series editors Pasvol G and Hoffman SL.** Imperial College Press, London, 2000.

[82] Thibeaux R, Weber C, Hon CC, Dillies MA, Avé P, *et al.*: **Identification of the virulence landscape essential for *Entamoeba histolytica* invasion of the human colon.** *PLoS Pathog* 2013, **9**: e1003824.

[83] Ullu E, Tschudi C, Chakraborty T: **RNA interference in protozoan parasites.** *Cell Microbiol* 2004, **6**: 509-519.

[84] Kolev NG, Tschudi C, Ullu E: **RNA interference in protozoan parasites: achievements and challenges.** *Eukaryotic cell* 2011, **10**: 1156-1163.

[85] Zhang H, Ehrenkaufer GM, Pompey JM, Hackney JA, Singh U: **Small RNAs with 5′-Polyphosphate Termini Associate with a Piwi-Related Protein and Regulate Gene**

Expression in the Single-Celled Eukaryote *Entamoeba histolytica.* *PLoS Pathog* 2008, **4**: e1000219.

[86] Zhang H, Alramini H, Tran V, Singh U: **Nucleus-localized Antisense Small RNAs with 5′-Polyphosphate Termini Regulate Long Term Transcriptional Gene Silencing in *Entamoeba histolytica* G3 Strain.** *J Biol Chem* 2011, **286**: 44467–44479.

[87] Cerutti H, Casas-Mollano JA: **On the origin and functions of RNA-mediated silencing: from protists to man.** *Curr Genet* 2006, **50**: 81-99.

[88] Baulcombe D: **Molecular biology. Amplified silencing.** *Science* 2007, **315**: 199–200.

[89] Holoch D, Moazed D: **RNA-mediated epigenetic regulation of gene expression.** *Nat Rev Genet* 2015, **16**: 71–84.

[90] Chen X: **Small RNAs – secrets and surprises of the genome.** *Plant J* 2010, **61**: 941–958.

[91] Zamore P, Tuschl T, Sharp P, Bartel D: **RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals.** *Cell* 2000, **101**: 25–33.

[92] Zhang H, Ehrenkaufer GM, Hall N, Singh U: **Small RNA pyrosequencing in the protozoan parasite *Entamoeba histolytica* reveals strain-specific small RNAs that target virulence genes.** *BMC Genomics* 2013, **14**: 53.

[93] Morf L, Pearson RJ, Wang AS, Singh U: **Robust gene silencing mediated by antisense small RNAs in the pathogenic protist *Entamoeba histolytica*.** *Nucleic Acids Res* 2013, **41**: 9424-9437.

[94] Mar-Aguilar F, Trevino V, Salinas-Hernández JE, Taméz-Guerrero MM, Barrón-González MP, *et al*.: **Identification and characterization of microRNAs from *Entamoeba histolytica* HM1-IMSS.** *PLoS One* 2013, **8**: e68202.

[95] De S, Pal D, Ghosh SK: ***Entamoeba histolytica*: computational identification of putative microRNA candidates.** *Exp Parasitol* 2006, **113**: 239-243.

[96] Petri WA Jr, Singh U: **Diagnosis and management of amebiasis.** *Clin Infect Dis* 1999, **29**: 1117-1125.

[97] Saraiya AA, Wang CC: **snoRNA, a Novel Precursor of microRNA in *Giardia lamblia*.** *PLoS Pathog* 2008, **4**: e1000224.

[98] Lin WC, Li SC, Lin WC, Shin JW, Hu SN, *et al*.: **Identification of microRNA in the protist *Trichomonas vaginalis*.** *Genomics* 2009, **93**: 487-493.

[99] Saraiya AA, Li W, Wang CC: **A microRNA derived from an apparent canonical biogenesis pathway regulates variant surface protein gene expression in *Giardia lamblia*.** *RNA* 2011, **17**: 2152-2164.

[100] Li W, Saraiya AA, Wang CC: **Gene regulation in *Giardia lambia* involves a putative microRNA derived from a small nucleolar RNA.** *PLoS Negl Trop Dis* 2011, **5**: e1338.

[101] Okoniewski MJ, Miller CJ: **Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations.** *BMC Bioinformatics* 2006, **7**: 276.

[102] Royce TE, Rozowsky JS, Gerstein MB: **Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification.** *Nucleic Acids Res* 2007, **35**: e99.

[103] Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, *et al.*: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5**: 613-619.

[104] Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, *et al.*: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.** *Mol Ecol* 2008, **17**: 1636-1647.

[105] Emrich SJ, Barbazuk WB, Li L, Schnable PS: **Gene discovery and annotation using LCM-454 transcriptome sequencing.** *Genome Res* 2007, **17**: 69-73.

[106] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science* 2008, **320**: 1344-1349.

[107] Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, *et al.*: **Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.** *Nature* 2008, **453**: 1239-1243.

[108] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**: 621-628.

[109] Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**: 57-63.

[110] Darby AC: **RNA-Seq Introduction to Bioinformatics**, 22nd January, Gene Expression Workshop, The Centre for Genomic Research, University of Liverpool, Merseyside, United Kingdom, 2013.

[111] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, *et al.*: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc* 2012, **7**: 562-578.

[112] Trapnell C, Salzberg SL: **How to map billions of short reads onto genomes.** *Nat Biotechnol* 2009, **27**: 455-457.

[113] Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD: **Count-based differential expression analysis of RNA sequencing data using R and Bioconductor.** *Nat Protoc* 2013, **8**: 1765-1786.

[114] Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**: R25.

[115] Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**: 139-140.

[116] Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet.journal* 2011, **17**: 10-12.

[117] Nelder JA, Wedderburn RWN: **Generalized Linear Models.** *J R Statist Soc A* 1972, **135**: 370-384.

[118] Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**: 1105-1111.

[119] Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high throughput sequencing data.** *Bioinformatics* 2015, **31**: 166-169.

[120] Wilks SS: **The large-sample distribution of the likelihood ratio for testing composite hypotheses.** *Ann Math Statist* 1938, **9**: 60-62.

[121] Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple Testing.** *J R Statist Soc B* 1995, **57**: 289-300.

[122] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, *et al.*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**: R80.

[123] Jones P, Binns D, Chang HY, Fraser M, Li W, *et al.*: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30**: 1236-1240.

[124] Supek F, Bošnjak M, Škunca N, Šmuc T: **REVIGO summarizes and visualizes long lists of gene ontology terms.** *PLoS One* 2011, **6**: e21800.

[125] Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, *et al*.: **Integration of biological networks and gene expression data using Cytoscape.** *Nat Protoc* 2007, **2**: 2366-2382.

[126] Daines B, Wang H, Wang L, Li Y, Han Y, *et al*.: **The *Drosophila melanogaster* transcriptome by paired-end RNA sequencing.** *Genome Res* 2011, **21**: 315-324.

[127] Paterson S: **Gene expression analysis in R**, 22nd January, Gene Expression Workshop, The Centre for Genomic Research, University of Liverpool, Merseyside, United Kingdom, 2013.

[128] Sharma A, Sojar HT, Glurich I, *et al*.: **Cloning, expression, and sequencing of a cell surface antigen containing a leucine-rich repeat motif from *Bacteroides forsythus* ATCC 43037.** *Infect Immun* 1998, **66**: 5703–5710.

[129] Ikegami A, Honma K, Sharma A, Kuramitsu HK: **Multiple functions of the leucine-rich repeat protein LrrA of *Treponema denticola*.** *Infect Immun* 2004, **72**: 4619–4627.

[130] Davis PH, Zhang Z, Chen M, Zhang X, Chakraborty S, Stanley SL Jr: **Identification of a family of BspA like surface proteins of *Entamoeba histolytica* with novel leucine rich repeats.** *Mol Biochem Parasitol* 2006, **145**: 111-116.

[131] Franco E, Manning-Cela R, Meza I: **Signal transduction in *Entamoeba histolytica* induced by interaction with fibronectin: presence and activation of phosphokinase A and its possible relation to invasiveness.** *Arch Med Res* 2002, **33**: 389-397.

[132] Vines RR, Ramakrishnan G, Rogers JB, Lockhart LA, Mann BJ, Petri WA: **Regulation of adherence and virulence by the *Entamoeba histolytica* lectin cytoplasmic domain, which contains a β2 integrin motif.** *Mol Biol Cell* 1998, **9**: 2069–2079.

[133] Ravdin JI, John JE, Johnston LI, Innes DI, Guerrant RL: **Adherence of *Entamoeba histolytica* trophozoites to rat and human colonic mucosa.** *Infect Immun* 1985, **48**: 292–297.

[134] Burchard GD, Bilke R: **Adherence of pathogenic and non-pathogenic *Entamoeba histolytica* strains to neutrophils.** *Parasitol Res* 1992, **78**: 146–153.

[135] Petri WA Jr: **Amebiasis and the *Entamoeba histolytica* Gal/GalNAc lectin: from lab bench to bedside.** *J Invest Med* 1996, **44**: 24–35.

[136] McCoy JJ, Weaver AM, Petri WA Jr: **Use of monoclonal anti-light subunit antibodies to study the structure and function of the *Entamoeba histolytica* Gal/GalNAc adherence lectin.** *Glycoconj J* 1994, **11**: 432–436.

[137] McCoy JJ, Mann BJ, Vedvick TS, Pak Y, Heimark DB, Petri WA Jr: **Structural analysis of the light subunit of the *Entamoeba histolytica* galactose-specific adherence lectin.** *J Biol Chem* 1993, **268**: 24223-24231.

[138] McCoy JJ, Mann BJ, Petri WA Jr: **Adherence and cytotoxicity of *Entamoeba histolytica* or how lectins let parasites stick around.** *Infect Immun* 1994, **62**: 3045–3050.

[139] Boettner DR, Huston C, Petri WA Jr: **Galactose/N-acetylgalactosamine lectin: The coordinator of host cell killing.** *J Biosci* 2002, **27**: 553-557.

[140] Padilla-Vaca F, Ankri S, Bracha R, Koole LA, Mirelman D: **Down regulation of *Entamoeba histolytica* virulence by monoxenic cultivation with *Escherichia coli* O55 is related to a decrease in expression of the light (35-kilodalton) subunit of the Gal/GalNAc lectin.** *Infect Immun* 1999, **67**: 2096–2102.

[141] Ehrenkaufer GM, Haque R, Hackney JA, Eichinger DJ, and Singh U: **Identification of developmentally regulated genes in *Entamoeba histolytica*: insights into mechanisms of stage conversion in a protozoan parasite.** *Cell Microbiol* 2007, **9:** 1426-1444.

[142] Tanabe K, Mackay M, Goman M, Scaife JG: **Allelic dimorphism in a surface antigen gene of the malaria parasite *Plasmodium falciparum*.** *J Mol Biol* 1987, **195**: 273-287.

[143] Polley SD, Weedall GD, Thomas AW, Golightly LM, Conway DJ: **Orthologous gene sequences of merozoite surface protein 1 (MSP1) from *Plasmodium reichenowi* and *P. gallinaceum* confirm an ancient divergence of *P. falciparum* alleles.** *Mol Biochem Parasitol* 2005, **142**: 25-31.

[144] Ali IK, Ehrenkaufer GM, Hackney JA, Singh U: **Growth of the protozoan parasite *Entamoeba histolytica* in 5-azacytidine has limited effects on parasite gene expression.** *BMC Genomics* 2007, **8**: 7.

[145] Bernes S, Siman-Tov R, Ankri S: **Epigenetic and classical activation of *Entamoeba histolytica* heat shock protein 100 (EHsp100) expression.** *FEBS Lett* 2005, **579**: 6395-6402.

[146] Ramakrishnan G, Gilchrist CA, Musa H, Torok MS, Grant PA, Mann BJ, Petri WA Jr: **Histone acetyltransferases and deacetylase in *Entamoeba histolytica*.** *Mol Biochem Parasitol* 2004, **138**: 205-216.

[147] Reed SL, Gigli I: **Lysis of complement-sensitive *Entamoeba histolytica* by activated terminal complement components. Initiation of complement activation by an extracellular neutral cysteine proteinase.** *J Clin Investig* 1990, **86**: 1815–1822.

[148] Reed SL, Ember JA, Herdman DS, DiScipio RG, Hugli TE, Gigli I: **The extracellular neutral cysteine proteinase of *Entamoeba histolytica* degrades anaphylatoxins C3a and C5a.** *J Immunol* 1995, **155**: 266–274.

[149] Reed SL, Keene WE, McKerrow JH: **Thiol proteinase expression correlates with pathogenicity of *Entamoeba histolytica*.** *J Clin Microbiol* 1989, **27**: 2772–2777.

[150] Tannich E, Scholze H, Nickel R, Horstmann RD: **Homologous cysteine proteinases of pathogenic and nonpathogenic *Entamoeba histolytica*. Differences in structure and expression.** *J Biol Chem* 1991, **266**: 4798-4803.

[151] Navarro-Garcia F, Chavez-Duenas L, Tsutsumi V, Posadas del Rio F, Lopez-Revilla R: ***Entamoeba histolytica*: increase of enterotoxicity and of 53- and 75-kDa cysteine proteinases in a clone of higher virulence.** *Exp Parasitol* 1995, **80**: 361–372.

[152] Reuber TL, Ausubel FM: **Isolation of *Arabidopsis* genes that differentiate between resistance responses mediated by the RPS2 and RPM1 disease resistance genes.** *Plant Cell* 1996, **8**: 241-249.

[153] Lorenzi H, Thiagarajan M, Haas B, Wortman J, Hall N, *et al*.: **Genome wide survey, discovery and evolution of repetitive elements in three *Entamoeba species*.** *BMC Genomics* 2008, **9**: 595.

[154] Nandi N, Sen A, Banerjee R, Kumar S, Kumar V, Ghosh AN, Das P: **Hydrogen peroxide induces apoptosis-like death in *Entamoeba histolytica* trophozoites.** *Microbiology* 2010, **156**: 1926-1941.

[155] Hughes MA, Lee CW, Holm CF, Ghosh S, Mills A, Lockhart LA, Reed SL, Mann BJ: **Identification of *Entamoeba histolytica* thiol-specific antioxidant as a GalNAc lectin-associated protein.** *Mol Biochem Parasitol* 2003, **127**: 113-120.

[156] Bhattacharya A, Padhan N, Jain R, Bhattacharya S: **Calcium-binding proteins of *Entamoeba histolytica*.** *Arch Med Res* 2006, **37**: 221-225.

[157] Sahoo N, Labruyère E, Bhattacharya S, Sen P, Guillén N, Bhattacharya A: **Calcium binding protein 1 of the protozoan parasite *Entamoeba histolytica* interacts with actin and is involved in cytoskeleton dynamics.** *J Cell Sci* 2004, **117**: 3625-3634.

[158] Gilchrist CA, Baba DJ, Zhang Y, Crasta O, Evans C, *et al*.: **Targets of the *Entamoeba histolytica* transcription factor URE3-BP.** *PLoS Negl Trop Dis* 2008, **2**: e282.

[159] Aslam S, Bhattacharya S, Bhattacharya A: **The Calmodulin-like calcium binding protein EhCaBP3 of *Entamoeba histolytica* regulates phagocytosis and is involved in actin dynamics.** *PLoS Pathog* 2012, **8**: e1003055.

[160] Somlata, Bhattacharya S, Bhattacharya A: **A C2 domain protein kinase initiates phagocytosis in the protozoan parasite *Entamoeba histolytica*.** *Nat Commun* 2011, **2**: 230.

[161] Carbajal ME, Manning-Cela R, Pina A, Franco E, Meza I: **Fibronectin-induced intracellular calcium rise in *Entamoeba histolytica* trophozoites: effect on adhesion and the actin cytoskeleton.** *Exp Parasitol* 1996, **82**: 11–20.

[162] Gilchrist CA, Holm CF, Hughes MA, Schaenman JM, Mann BJ, Petri WA Jr: **Identification and characterization of an *Entamoeba histolytica* upstream regulatory element 3 sequence-specific DNA-binding protein containing EF-hand motifs.** *J Biol Chem* 2001, **276**: 11838–11843.

[163] Gilchrist CA, Leo M, Line CG, Mann BJ, Petri WA Jr: **Calcium modulates promoter occupancy by the *Entamoeba histolytica* Ca2+-binding transcription factor URE3-BP.** *J Biol Chem* 2003, **278**: 4646–4653.

[164] Ravdin JI, Moreau F, Sullivan JA, Petri WA Jr, Mandell GL: **Relationship of free intracellular calcium to the cytolytic activity of *Entamoeba histolytica*.** *Infect Immun* 1988, **56**: 1505–1512.

[165] Ravdin JI, Murphy CF, Guerrant RL, Long-Krug SA: **Effect of antagonists of calcium and phospholipase A on the cytopathogenicity of *Entamoeba histolytica*.** *J Infect Dis* 1985, **152**: 542–549.

[166] Makioka A, Kumagai M, Kobayashi S, Takeuchi T: **Possible role of calcium ions, calcium channels and calmodulin in excystation and metacystic development of *Entamoeba invadens*.** *Parasitol Res* 2002, **88**: 837–843.

[167] Makioka A, Kumagai M, Ohtomo H, Kobayashi S, Takeuchi T: **Effect of calcium antagonists, calcium channel blockers and calmodulin inhibitors on the growth and encystation of *Entamoeba histolytica* and *E. invadens*.** *Parasitol Res* 2001, **87**: 833–837.

[168] Moreno H, Linford AS, Gilchrist CA, Petri WA Jr: **Phospholipid-binding protein EhC2A mediates calcium-dependent translocation of transcription factor URE3-BP to the plasma membrane of *Entamoeba histolytica*.** *Eukaryot Cell* 2010, **9**: 695-704.

[169] Okada M, Huston CD, Oue M, Mann BJ, Petri WA Jr, Kita K, Nozaki T: **Kinetics and strain variation of phagosome proteins of *Entamoeba histolytica* by proteomic analysis.** *Mol Biochem Parasitol* 2006, **145**: 171-183.

[170] Luna-Arias JP, Hernandez-Rivas R, de Dios-Bravo G, Garcia J, Mendoza L, Orozco E: **The TATA-box binding protein of *Entamoeba histolytica*: cloning of the gene and location of the protein by immunofluorescence and confocal microscopy.** *Microbiology* 1999, **145**: 33-40.

[171] Schaenman JM, Gilchrist CA, Mann BJ, Petri WA Jr: **Identification of two *Entamoeba histolytica* sequence-specific URE4 enhancer-binding proteins with homology to the RNA-binding motif RRM.** *J Biol Chem* 2001, **276**: 1602-1609.

[172] Mendoza L, Orozco E, Rodríguez MA, García-Rivera G, Sánchez T, García E, Gariglio P: **Ehp53, an *Entamoeba histolytica* protein, ancestor of the mammalian tumour suppressor p53.** *Microbiology* 2003, **149**: 885-893.

[173] Abhyankar MM, Hochreiter AE, Hershey J, Evans C, Zhang Y, *et al*.: **Characterization of an *Entamoeba histolytica* high-mobility-group box protein induced during intestinal infection.** *Eukaryot Cell* 2008, **7**: 1565-1572.

[174] Iyer LM, Anantharaman V, Wolf MY, Aravind L: **Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes.** *Int J Parasitol* 2008, **38**: 1-31.

[175] Ong SJ, Hsu HM, Liu HW, Chu CH, Tai JH: **Activation of multifarious transcription of an adhesion protein ap65-1 gene by a novel Myb2 protein in the protozoan parasite *Trichomonas vaginalis*.** *J Biol Chem* 2007, **282**: 6716-6725.

[176] Rosinski JA, Atchley WR: **Molecular evolution of the Myb family of transcription factors: evidence for polyphyletic origin.** *J Mol Evol* 1998, **46**: 74-83.

[177] Meneses E, Cárdenas H, Zárate S, Brieba LG, Orozco E, López-Camarillo C, Azuara-Liceaga E: **The R2R3 Myb protein family in *Entamoeba histolytica*.** *Gene* 2010, **455**: 32-42.

[178] Ong SJ, Hsu HM, Liu HW, Chu CH, Tai JH: **Multifarious transcriptional regulation of adhesion protein gene *ap65-1* by a novel Myb1 protein in the protozoan parasite *Trichomonas vaginalis*.** *Eukaryotic Cell* 2006, **5**: 391-399.

[179] MacFarlane RC, Shah PH, Singh U: **Transcriptional profiling of *Entamoeba histolytica* trophozoites.** *Int J Parasitol* 2005, **35**: 533-542.

[180] Van Valen L: **A new evolutionary law.** *Evol Theory* 1973, **1**: 1–30.

[181] Woolhouse ME, Webster JP, Domingo E, Charlesworth B, Levin BR: **Biological and biomedical implications of the co-evolution of pathogens and their hosts.** *Nat Genet* 2002, **32**: 569-577.

[182] Stenseth NC, Smith JM: **Coevolution in ecosystems: Red Queen evolution or stasis?** *Evolution* 1984, **38**: 870–880.

[183] Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, *et al*.: **Antagonistic coevolution accelerates molecular evolution.** *Nature* 2010, **464**: 275-278.

[184] Jeffares DC, Pain A, Berry A, Cox AV, Stalker J, *et al*.: **Genome variation and evolution of the malaria parasite *Plasmodium falciparum*.** *Nat Genet* 2007, **39**: 120-125.

[185] Bhattacharya D, Haque R, Singh U: **Coding and noncoding genomic regions of *Entamoeba histolytica* have significantly different rates of sequence polymorphisms: implications for epidemiological studies.** *J Clin Microbiol* 2005, **43**: 4815-4819.

[186] Mohan R, Venugopal S: **Computational structural and functional analysis of hypothetical proteins of *Staphylococcus aureus*.** *Bioinformation* 2012, **8**: 722–728.

[187] Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, *et al*.: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34**: D247-D251.

[188] Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, *et al*.: **PRINTS and its automatic supplement prePRINTS.** *Nucleic Acids Res* 2003, **31**: 400-402.

[189] Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, *et al*. : **The PROSITE database.** *Nucleic Acids Res* 2006, **34**: D227-D230.

[190] Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P: **SMART 5: domains in the context of genomes and networks.** *Nucleic Acids Res* 2006, **34**: D257-D260.

[191] Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, *et al*.: **The PANTHER database of protein families, subfamilies, functions and pathways.** *Nucleic Acids Res* 2005, **32**: D284-D288.

[192] Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families.** *Nucleic Acids Res* 2003, **31**: 371-373.

[193] Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J: **The SUPERFAMILY database in 2004: additions and improvements.** *Nucleic Acids Res* 2004, **32**: D235-D239.

[194] Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH: **PIRSF family classification system for protein functional and evolutionary analysis.** *Evol Bioinform Online* 2007, **2**: 197-209.

[195] Parsons M, Valentine M, Carter V: **Protein kinases in divergent eukaryotes: identification of protein kinase activities regulated during trypanosome development.** *Proc Natl Acad Sci USA* 1993, **90**: 2656–2660.

[196] Parsons M, Worthey EA, Ward PN, Mottran JC: **Comparative analysis of the kinomes of three pathogenic trypanosomatids: *Leishmania major*, *Trypanosoma brucei*, *Trypanosoma cruzi*.** *BMC Genomics* 2005, **6**: 127.

[197] Anamika K, Bhattacharya A, Srinivasan N: **Analysis of the protein kinome of *Entamoeba histolytica*.** *Proteins* 2008, **71**: 995-1006.

[198] Johnson LN: **The regulation of protein phosphorylation.** *Biochem Soc Trans* 2009, **37**: 627-641.

[199] Meza I: **Extracellular matrix-induced signaling in *Entamoeba histolytica*: its role in invasiveness.** *Parasitol Today* 2000, **16**: 23–28.

[200] Buss SN, Hamano S, Vidrich A, Evans C, Zhang Y, *et al.*: **Members of the *Entamoeba histolytica* transmembrane kinase family play non-redundant roles in growth and phagocytosis.** *Int J Parasitol* 2010, **40**: 833-843.

[201] Goldberg JM, Manning G, Liu A, Fey P, Pilcher KE, Xu Y, Smith JL: **The *Dictyostelium* Kinome—Analysis of the Protein Kinases from a Simple Model Organism.** *PLoS Genet* 2006, **2**: e38.

[202] Bosch DE, Siderovski DP: **G protein signaling in the parasite *Entamoeba histolytica*.** *Exp Mol Med* 2013, **45**: e15.

[203] Goitre L, Trapani E, Trabalzini L, Retta SF: **The Ras superfamily of small GTPases: the unlocked secrets.** *Methods Mol Biol* 2014, **1120**: 1-18.

[204] Wennerberg K, Rossman KL, Der CJ: **The Ras superfamily at a glance.** *J Cell Sci* 2005, **118**: 843–846.

[205] Maugis B, Brugues J, Nassoy P, Guillen N, Sens P, Amblard F: **Dynamic instability of the intracellular pressure drives bleb-based motility.** *J Cell Sci* 2010, **123**: 3884–3892.

[206] Voigt H, Guillen N: **New insights into the role of the cytoskeleton in phagocytosis of *Entamoeba histolytica*.** *Cell Microbiol* 1999, **1**: 195–203.

[207] Tavares P, Sansonetti P, Guillen N: **The interplay between receptor capping and cytoskeleton remodeling in *Entamoeba histolytica*.** *Arch Med Res* 2000, **31**: S140–S142.

[208] Clark CG, Johnson PJ, Adam RD (ed.): **Anaerobic Parasitic Protozoa: Genomics and Molecular Biology.** Caister Academic Press, Norfolk, 2010.

[209] Tavares P, Rigothier MC, Khun H, Roux P, Huerre M, Guillén N: **Roles of cell adhesion and cytoskeleton activity in *Entamoeba histolytica* pathogenesis: a delicate balance.** *Infection and Immunity* 2005, **73**: 1771-1778.

[210] Tavares P, Sansonetti P, Guillén N: **Cell adhesion and polarization in a pathogenic protozoan: role of the *Entamoeba histolytica* Gal/GalNAc lectin.** *Microbes Infect* 2000, **2:** 643-649.

[211] Arhets P, Olivo JC, Gounon P, Sansonetti P, Guillén N: **Virulence and functions of myosin II are inhibited by overexpression of light meromyosin in *Entamoeba histolytica.*** *Mol Biol Cell* 1998, **8:** 1537-1547.

[212] Coudrier E, Amblard F, Zimmer C, Roux P, Olivo JC, Rigothier MC, Guillén N: **Myosin II and the Gal-GalNAc lectin play a crucial role in tissue invasion by *Entamoeba histolytica.*** *Cell Microbiol* 2005, **7:** 19-27.

[213] Bañuelos S, Saraste M, Carugo KD: **Structural comparisons of calponin homology domains: implications for actin binding.** *Structure* 1998, **6**: 1419-1431.

[214] Castresana J, Saraste M: **Does Vav bind to F-actin through a CH domain?** *FEBS Lett* 1995, **374**: 149–151.

[215] Nozaki T, Bhattacharya A (ed.): **Amebiasis: Biology and Pathogenesis of *Entamoeba*.** Springer, Tokyo, 2015.

[216] Kadrmas JL, Beckerle MC: **The LIM domain: from the cytoskeleton to the nucleus.** *Nat Rev Mol Cell Biol* 2004, **5**: 920–931.

[217] Lappalainen P, Paunola E, Mattila PK: **WH2 domain: a small, versatile adapter for actin monomers.** *FEBS Lett* 2002, **513**: 92–97.

[218] Veltman DM, Insall RH: **WASP family proteins: their evolution and its physiological implications**. *Mol Biol Cell* 2010, **21**: 2880–2893.

[219] Machesky LM, Insall RH: **Scar1 and the related Wiskott-Aldrich syndrome protein, WASP, regulate the actin cytoskeleton through the Arp2/3 complex.** *Curr Bio* 1998, **8**: 1347–1356.

[220] Lecker SH, Goldberg AL, Mitch WE: **Protein degradation by the ubiquitin-proteasome pathway in normal and disease states.** *J Am Soc Nephrol* 2006, **17**: 1807-1819.

[221] Bosch DE, Siderovski DP: **Structural determinants of ubiquitin conjugation in *Entamoeba histolytica*.** *J Biol Chem* 2013, **288**: 2290-2302.

[222] Makioka A, Kumagai M, Ohtomo H, Kobayashi S, Takeuchi T: **Effect of proteasome inhibitors on the growth, encystation, and excystation of *Entamoeba histolytica* and *Entamoeba invadens*.** *Parasitol Res* 2002, **88**: 454–459.

[223] The Gene Ontology Consortium, Ashburner M, Ball CA, Blake JA, Botstein D, *et al*.: **Gene Ontology: tool for the unification of biology.** Nat Genet 2000, **25**: 25–29.

[224] The Gene Ontology Consortium: **Gene Ontology Consortium: going forward.** *Nucleic Acids Res* 2015, **43**: D1049–D1056.

[225] Rivals I, Personnaz L, Taing L, Potier MC: **Enrichment or depletion of a GO category within a class of genes: which test?** *Bioinformatics* 2007, **23**: 401–407.

[226] Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**: 406–425.

[227] Pesquita C, Faria D, Falcão AO, Lord P, Couto FM: **Semantic similarity in biomedical ontologies.** *PLoS Comput Biol* 2009, **5**: e1000443.

[228] Schlicker A, Domingues F, Rahnenfuhrer J, Lengauer T: **A new measure for functional similarity of gene products based on Gene Ontology.** *BMC Bioinformatics* 2006, **7**: 302.

[229] López-Casamichana M, Orozco E, Marchat LA, López-Camarillo C: **Transcriptional profile of the homologous recombination machinery and characterization of the EhRAD51 recombinase in response to DNA damage in *Entamoeba histolytica*.** *BMC Mol Biol* 2008, **9**: 35.

[230] Singh N, Bhattacharya A, Bhattacharya S: **Homologous recombination occurs in *Entamoeba* and is enhanced during growth stress and stage conversion.** *PLoS ONE* 2013, **8**: e74465.

[231] Del Socorro Charcas-Lopez M, Garcia-Morales L, Pezet-Valdez M, Lopez-Camarillo C, Zamorano-Carrillo A, Marchat LA: **Expression of EhRAD54, EhRAD51, and EhBLM proteins during DNA repair by homologous recombination in *Entamoeba histolytica*.** *Parasite* 2014, **21**: 7.

[232] Chen J, Cooper DN, Chuzhanova N, Férec C, Patrinos GP: **Gene conversion: mechanisms, evolution and human disease.** *Nature Reviews Genetics* 2007, **8**: 762–775.

[233] Brandsma I, Gent DC: **Pathway choice in DNA double strand break repair: observations of a balancing act.** *Genome Integr* 2012, **3**: 9.

[234] San Filippo J, Sung P, Klein H: **Mechanism of eukaryotic homologous recombination.** *Annu Rev Biochem* 2008, **77**: 229–257.

[235] McEachern MJ, Haber JE: **Break-induced replication and recombinational telomere elongation in yeast.** *Annu Rev Biochem* 2006, **75**: 111-135.

[236] Keeney S, Giroux CN, Kleckner N: **Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family.** *Cell* 1997, **88**: 375–384.

[237] Weedall GD, Hall N: **Sexual reproduction and genetic exchange in parasitic protists.** *Parasitology* 2015, **142**: S120-S127.

[238] Malik SB, Pightling AW, Stefaniak LM, Schurko AM, Logsdon JM Jr: **An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*.** *PLoS ONE* 2007, **3**: e2879.

[239] Schurko AM, Logsdon JM: **Using a meiosis detection toolkit to investigate ancient asexual "scandals" and the evolution of sex.** *Bioessays* 2008, **30**, 579–589.

[240] Ramesh MA, Malik SB, Logsdon JM Jr: **A phylogenomic inventory of meiotic genes; evidence for sex in *Giardia* and an early eukaryotic origin of meiosis.** *Curr Biol* 2005, **15**: 185–191.

[241] Willhoeft U, Tannich E: **The electrophoretic karyotype of *Entamoeba histolytica*.** *Mol Biochem Parasitol* 1999, **99**: 41–53.

[242] Baez-Camargo M, Gharaibeh R, Riveron AM, Hernández FDL, Luna JP, et al.: **Gene amplification in *Entamoeba histolytica*.** *Invasion Metastasis* 1996, **16**: 269–279.

[243] Deitsch KW, del Pinal A, Wellems TE: **Intra-cluster recombination and var transcription switches in the antigenic variation of *Plasmodium falciparum*.** *Mol Biochem Parasitol* 1999, **101**: 107–116.

[244] Conway C, Proudfoot C, Burton P, Barry JD, McCulloch R: **Two pathways of homologous recombination in *Trypanosoma brucei*.** *Mol Microbiol* 2002, **45**: 1687–1700.

[245] Papadopoulou B, Dumas C: **Parameters controlling the rate of gene targeting frequency in the protozoan parasite *Leishmania*.** *Nucleic Acids Res* 1997, **25**: 4278–4286.

[246] Payrastre B, Missy K, Giuriato S, Bodin S, Plantavid M, Gratacap M: **Phosphoinositides: key players in cell signalling, in time and space.** *Cell Signal* 2001, **13**: 377-387.

[247] Gillooly DJ, Simonsen A, Stenmark H: **Phosphoinositides and phagocytosis.** *J Cell Biol* 2001, **155**: 15-18.

[248] Kölsch V, Charest PG, Firtel RA: **The regulation of cell motility and chemotaxis by phospholipid signaling.** *J Cell Sci* 2008, **121**: 551-559.

[249] Blazquez S, Guigon G, Weber C, Syan S, Sismeiro O, *et al*.: **Chemotaxis of *Entamoeba histolytica* towards the pro-inflammatory cytokine TNF is based on PI3K signalling, cytoskeleton reorganization and the Galactose/N-acetylgalactosamine lectin activity.** *Cell Microbiol* 2008, **10**: 1676-1686.

[250] López-Contreras L, Hernández-Ramírez VI, Flores-García Y, Chávez-Munguía B, Talamás-Rohana P: **Src and PI3K inhibitors affect the virulence factors of *Entamoeba histolytica*.** *Parasitology* 2013, **140**: 202–209.

[251] Levin BR, Bull JJ: **Short-sighted evolution and the virulence of pathogenic microorganisms.** *Trends Microbiol* 1994, **2**: 76-81.

[252] Levin BR: **The evolution and maintenance of virulence in microparasites.** *Emerg Infect Dis* 1996, **2**: 93-102.

[253] Hoffman O, Weber RJ: **Pathophysiology and treatment of bacterial meningitis.** *Ther Adv Neurol Disord* 2009, **2**: 1-7.

[254] Nathanson N, Kew OM: **From emergence to eradication: the epidemiology of poliomyelitis deconstructed.** *Am J Epidemiol* 2010, **172**: 1213-1229.

[255] Kuenen WA, Swellengrebel NH: **Die Entamoben des Menschen und ihre praktische Bedeutung.** *Centralbl Baht* 1913, **71**: 378-410.

[256] Frank SA: **Models of parasite virulence.** *Q Rev Biol* 1996, **71:** 37-78.

[257] Alizon S, Hurford A, Mideo N, Van Baalen M: **Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future.** *J Evol Biol* 2009, **22**: 245-259.

[258] Li J, Jiang H, Wong WH: **Modeling non-uniformity in short-read rates in RNA-Seq data.** *Genome Biol* 2010, **11**: R50.

[259] Finotello F, Lavezzo E, Bianco L, Barzon L, Mazzon P, *et al.*: **Reducing bias in RNA sequencing data: a novel approach to compute counts.** *BMC Bioinformatics* 2014, **15**: S7.

[260] Shiroguchi K, Jia TZ, Sims PA, Xie XS: **Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes.** *Proc Natl Acad Sci U S A* 2012, **109**: 1347-1352.

[261] Kulkarni MM: **Digital multiplexed gene expression analysis using the NanoString nCounter system.** *Curr Protoc Mol Biol* 2011, **Chapter 25**: Unit 25B.10.

[262] Abhyankar MM, Hochreiter AE, Hershey J, Evans C, Zhang Y, *et al.*: **Characterization of an *Entamoeba histolytica* High-Mobility-Group Box Protein Induced during Intestinal Infection.** *Eukaryot Cell* 2008, **7**: 1565-1572.

[263] Husain A, Jeelani G, Sato D, Nozaki T: **Global analysis of gene expression in response to L-Cysteine deprivation in the anaerobic protozoan parasite *Entamoeba histolytica*.** *BMC Genomics* 2011, **12**: 275.

[264] Sateriale A, Vaithilingam A, Donnelly L, Miller P, Huston CD: **Feed-forward regulation of phagocytosis by *Entamoeba histolytica*.** *Infect Immun* 2012, **80**: 4456-4462.

[265] Baxt LA, Baker RP, Singh U, Urban S: **An *Entamoeba histolytica* rhomboid protease with atypical specificity cleaves a surface lectin involved in phagocytosis and immune evasion.** *Genes Dev* 2008, **22**: 1636-1646.

[266] Nakada-Tsukui K, Saito-Nakano Y, Husain A, Nozaki T: **Conservation and function of Rab small GTPases in *Entamoeba*: annotation of *E. invadens* Rab and its use for the understanding of *Entamoeba* biology.** *Exp Parasitol* 2010, **126**: 337-347.

[267] R Core Team, **R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.** URL http://www.R-project.org/.

[268] Shilova VY, Garbuz DG, Myasyankina EN, Chen B, Evgen'ev MB, *et al*.: **Remarkable site specificity of local transposition into the Hsp70 promoter of *Drosophila melanogaster*.** *Genetics* 2006, **173**: 809–820.

[269] Minning TA, Weatherly DB, Flibotte S, Tarleton RL: **Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in *Trypanosoma cruzi* strains revealed by array comparative genomic hybridization.** *BMC Genomics* 2011, **12**: 139.

[270] Hurles M: **Gene duplication: the genomic trade in spare parts.** *PLoS biology* 2004, **2**: E206.

[271] Jones C, Todeschini AR, Agrellos OA, Previato JO, Mendonca-Previato L: **Heterogeneity in the biosynthesis of mucin O-glycans from *Trypanosoma cruzi* tulahuen strain with the expression of novel galactofuranosyl-containing oligosaccharides.** *Biochemistry* 2004, **43**: 11889–11897.

[272] Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**: 1754-1760.

[273] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al*.: **The Sequence alignment/map (SAM) format and SAMtools.** *Bioinformatics* 2009, **25**: 2078-2079.

[274] Zurita M, Alagón A, Vargas-Villarreal J, Lizardi PM: **The *Entamoeba histolytica* rDNA episome: nuclear localization, DNAase I sensitivity map, and specific DNA-protein interactions.** *Mol Microbiol* 1991, **5**:1843-1851.

[275] Riveron AM, Lopez-Canovas L, Baez-Camargo M, Flores E, Perez-Perez G, Luna-Arias JP, Orozco E: **Circular and linear DNA molecules in the *Entamoeba histolytica* complex molecular karyotype.** *Eur Biophys J* 2000, **29**: 48–56.

[276] Dhar SK, Choudhury NR, Bhattacharaya A, Bhattacharya S: **A multitude of circular DNAs exist in the nucleus of *Entamoeba histolytica*.** *Mol Biochem Parasitol* 1995, **70**: 203–206.

[277] Bhattacharya S, Som I, Bhattacharya A: **The ribosomal DNA plasmids of entamoeba.** *Parasitol Today* 1998, **14**: 181–185.

[278] Jansson A, Gillin F, Kagardt U, Hagblom P: **Coding of hemolysins within the ribosomal RNA repeat on a plasmid in *Entamoeba histolytica*.** *Science* 1994, **263**: 1440–1443.

[279] Von Hoff DD: **New mechanisms of gene amplification in drug resistance (the episome model). Molecular and Clinical Advances in Anticancer Drug Resistance**. *Cancer Treat Res* 1991, **57**: 1-11.

[280] Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, *et al*.: **Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*.** *Genome Res* 2011, **21**: 2129–2142.

[281] Callejas S, Leech V, Reitter C, Melville S: **Hemizygous subtelomeres of an African trypanosome chromosome may account for over 75% of chromosome length.** *Genome Res* 2006, **16**: 1109-1118.

[282] Ghildiyal M, Zamore PD: **Small silencing RNAs: an expanding universe.** *Nature reviews* 2009, **10**: 94-108.

[283] Malecová B, Morris KV: **Transcriptional gene silencing mediated by non-coding RNAs.** *Curr Opin Mol Ther* 2010, **12**: 214-222.

[284] Kaur G, Lohia A: **Inhibition of gene expression with double strand RNA interference in *Entamoeba histolytica*.** *Biochem Biophys Res Commun* 2004, **320**: 1118-1122.

[285] Vayssié L, Vargas M, Weber C, Guillén N: **Double-stranded RNA mediates homology-dependent gene silencing of gamma-tubulin in the human parasite *Entamoeba histolytica*.** *Mol Biochem Parasitol* 2004, **138**: 21-28.

[286] Linford AS, Moreno H, Good KR, Zhang H, Singh U, Petri WA Jr: **Short hairpin RNA-mediated knockdown of protein expression in *Entamoeba histolytica*.** *BMC Microbiol* 2009, **9**: 38.

[287] Bracha R, Nuchamowitz Y, Anbar M, Mirelman D: **Transcriptional silencing of multiple genes in trophozoites of *Entamoeba histolytica*.** *PLOS pathogens* 2006, **2**: e48.

[288] Pak J, Fire A: **Distinct populations of primary and secondary effectors during RNAi in *C. elegans*.** *Science* 2007, **315**: 241-244.

[289] Aoki K, Moriguchi H, Yoshioka T, Okawa K, Tabara H: ***In vitro* analyses of the production and activity of secondary small interfering RNAs in *C. elegans*.** *Embo J* 2007, **26**: 5007-5019.

[290] Ambros V: **The functions of animal microRNAs.** *Nature* 2004, **431**: 350–355.

[291] Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 2004, **116**: 281–297.

[292] Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, *et al*.: **The small RNA profile during *Drosophila melanogaster* development.** *Dev Cell* 2003, **5**: 337-350.

[293] Zhang B, Wang Q, Pan X: **MicroRNAs and their regulatory roles in animals and plants.** *J Cell Physiol* 2007, **210**: 279-289.

[294] Pillai RS, Bhattacharyya SN, Filipowicz W: **Repression of protein synthesis by miRNAs: how many mechanisms?** *Trends Cell Biol* 2007, **17**: 118–126.

[295] Zhang H, Pompey JM, Singh U: **RNA interference in *Entamoeba histolytica*: implications for parasite biology and gene silencing.** *Future Microbiol* 2011, **6**: 103-117.

[296] Langmead B, Salzberg S: **Fast gapped-read alignment with Bowtie 2.** *Nature Methods* 2012, **9**: 357-359.

[297] Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N: **Discovering microRNAs from deep sequencing data using miRDeep.** *Nature Biotech* 2008, **26**: 407-415.

[298] Kozomara A, Griffiths-Jones S: **miRBase: annotating high confidence microRNAs using deep sequencing data.** *Nucleic Acids Res* 2014, **42**: D68–D73.

[299] Bonnet E, Wuyts J, Rouzé P, Van de Peer Y: **Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.** *Bioinformatics* 2004, **20**: 2911-2917.

[300] Tarver JE, Donoghue PC, Peterson KJ: **Do miRNAs have a deep evolutionary history?** *Bioessays* 2012, **34**: 857-866.

[301] Friedman Y, Balaga O, Linial M: **Working together: combinatorial regulation by microRNAs.** *Adv Exp Med Biol* 2013, **774**: 317-337.

[302] Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, *et al*.: **Combinatorial microRNA target predictions.** *Nat Genet* 2005, **37**: 495-500.

[303] Quach J, St-Pierre J, Chadee K: **The future for vaccine development against *Entamoeba histolytica*.** *Hum Vaccin Immunother* 2014, **10**: 1514-1521.

[304] Li J, Yang G, Li S, Cao G, Zhao Q, *et al*.: **3′-Poly(A) tail enhances siRNA activity against exogenous reporter genes in MCF-7 cells.** *J RNAi Gene Silencing* 2006, **2**: 195-204.

Appendices

Appendices

Appendix Table 1.1: Functional genes with transcriptomic upregulation ($log_2FC \geq 2$) in all three virulent strains: HM-1:IMSS (n=395), PVBM08B (n=229) and IULA:1092:1 (n=386). These upregulated transcripts can be assigned into 41 functional gene annotations as listed below.

group	Functional gene annotation	Number of upregulated genes, $log_2FC \geq 2$		
		HM-1:IMSS	PVBM08B	IULA:1092:1
1.	leucine-rich repeat protein, BspA family	31	21	23
2.	AIG1 family protein	18	7	5
3.	protein kinase domain-containing protein	8	3	6
4.	surface antigen ariel1, putative	6	2	4
5.	regulator of nonsense transcripts, putative	6	2	3
6.	heat shock protein 70, putative	5	3	3
7.	zinc finger protein, putative	3	3	4
8.	serine-threonine-isoleucine rich protein, putative	3	3	2
9.	cysteine proteinase, putative	3	1	3
10.	iron-sulfur flavoprotein, putative	3	1	3
11.	tyrosine kinase, putative	2	2	5
12.	CXXC-rich protein	2	2	3
13.	proteoglycan-4 precursor, putative	2	2	2
14.	DNA polymerase, putative	2	2	1
15.	peroxiredoxin	2	1	7
16.	26S proteinase regulatory subunit, putative	2	1	3
17.	Skp1 family protein	2	1	2
18.	acetyltransferase, putative	2	1	1
19.	dentin sialophosphoprotein precursor, putative	2	1	1
20.	heat shock protein 70, mitochondrial, putative	2	1	1
21.	predicted protein	2	1	1
22.	RhoGAP domain-containing protein	2	1	1
23.	mucin-like protein 1 precursor, putative	1	1	2
24.	60S ribosome subunit biogenesis protein NIP7, putative	1	1	1
25.	cdc48-like protein, putative	1	1	1
26.	chaperone clpB, putative	1	1	1
27.	Dedicator of cytokinesis domain-containing protein	1	1	1
28.	dextranase precursor, putative	1	1	1
29.	Fe-S cluster assembly protein NifU, putative	1	1	1
30.	heat shock protein, putative	1	1	1

Appendix Table 1.1: Functional genes with transcriptomic upregulation ($log_2FC \geq 2$) in all three virulent strains: HM-1:IMSS (n=395), PVBM08B (n=229) and IULA:1092:1 (n=386). **(Continued)**

group	Functional gene annotation	Number of upregulated genes, $log_2FC \geq 2$		
		HM-1:IMSS	PVBM08B	IULA:1092:1
31.	kinetochore protein Spc25 domain-containing protein	1	1	1
32.	molybdenum cofactor synthesis protein3, putative	1	1	1
33.	Myb family DNA-binding protein, SHAQKYF family	1	1	1
34.	PP-loop family protein	1	1	1
35.	protein tyrosine kinase domain-containing protein	1	1	1
36.	Rap/Ran GTPase-activating protein, putative	1	1	1
37.	replication protein, pseuudogene, putative	1	1	1
38.	serine acetyltransferase 1	1	1	1
39.	tRNA-Leu (anticodon: CAA)	1	1	1
40.	WD domain-containing protein	1	1	1
41.	hypothetical protein	267	149	283
	Total	**395**	**229**	**386**

Appendix Table 1.2: Functional genes with transcriptomic upregulation ($log_2FC \geq 2$), only in two virulent strains: HM-1:IMSS (n=25) and PVBM08B (n=23). These upregulated transcripts can be assigned into 20 functional gene annotations as listed below.

group	Functional gene annotation	Number of upregulated genes, $log_2FC \geq 2$	
		HM-1:IMSS	PVBM08B
1.	Ras family GTPase	3	2
2.	C2 domain-containing protein	2	2
3.	NADPH-dependent FMN reductase domain-containing protein	2	2
4.	P-glycoprotein-2, putative	2	1
5.	60S ribosomal protein L6, putative	1	1
6.	78 kD aglucose-regulated protein homolog precursor, putative	1	1
7.	coiled-coil domain-containing protein 25, putative	1	1
8.	DNA mismatch repair protein Msh2, putative	1	1
9.	endonuclease V, putative	1	1
10.	ethanolamine phosphotransferase, putative	1	1
11.	HEAT repeat domain-containing protein	1	1
12.	hydrolase, alpha/beta fold family domain-containing protein	1	1
13.	peptidyl-prolyl cis-trans isomerase, FKBP-type , putative	1	1
14.	pre-mRNA cleavage factor I 25 kDa subunit, putative	1	1
15.	protein phosphatase domain-containing protein	1	1
16.	rab GDP dissociation inhibitor alpha, putative	1	1
17.	Ras-like protein 1, putative	1	1
18.	signal recognition particle 54 kDa protein, putative	1	1
19.	splicing factor 3B subunit 1, putative	1	1
20.	U3 small nucleolar ribonucleo protein MPP10, putative	1	1
	Total	**25**	**23**

Appendix Table 1.3: Functional genes with transcriptomic upregulation ($log_2FC \geq 2$), only in two virulent strains: HM-1:IMSS (n=40) and IULA:1092:1 (n=39). These upregulated transcripts can be assigned into 31 functional gene annotations as listed below.

group	Functional gene annotation	Number of upregulated genes, $log_2FC \geq 2$	
		HM-1:IMSS	IULA:1092:1
1.	P-glycoprotein 5, putative	3	2
2.	endonuclease/exonuclease/phosphatase family protein	3	1
3.	glutamic acid-rich protein precursor, putative	2	3
4.	cylicin-2, putative	2	2
5.	Rab family GTPase	2	2
6.	RNA recognition motif domain-containing protein	2	2
7.	cysteine synthase A, putative	2	1
8.	long-chain-fatty-acid--CoA ligase, putative	1	2
9.	serine/threonine kinase, putative	1	2
10.	(2r)-phospho-3-sulfolactate synthase, putative	1	1
11.	actobindin, putative	1	1
12.	ADP-ribosylation factor 1, putative	1	1
13.	alcohol dehydrogenase, putative	1	1
14.	aldose reductase, putative	1	1
15.	D-tyrosyl-tRNA(Tyr) deacylase, putative	1	1
16.	Dopey domain protein, putative	1	1
17.	eukaryotic translation initiation factor 4 gamma, putative	1	1
18.	G-box-binding factor, putative	1	1
19.	GTP-binding protein EhRabX29, putative	1	1
20.	histone H2A, putative	1	1
21.	HMG box protein	1	1
22.	malic enzyme, putative	1	1
23.	mucin-5AC, putative	1	1
24.	myb-like DNA-binding domain-containing protein	1	1
25.	protein kinase, putative	1	1
26.	ribosomal RNA methyltransferase, putative	1	1
27.	serine-rich 25 kDa antigen protein, putative	1	1
28.	ThiF family protein	1	1
29.	TolA-like protein, putative	1	1
30.	transporter, auxin efflux carrier (AEC) family	1	1
31.	trichohyalin, putative	1	1
	Total	**40**	**39**

Appendix Table 1.4: Functional genes with transcriptomic upregulation ($log_2FC \geq 2$), only in two virulent strains: PVBM08B (n=8) and IULA:1092:1 (n=8). These upregulated transcripts can be assigned into 8 functional gene annotations as listed below.

group	Functional gene annotation	Number of upregulated genes, $log_2FC \geq 2$	
		PVBM08B	IULA:1092:1
1.	26S proteinase regulatory subunit S10B, putative	1	1
2.	cysteine surface protein, putative	1	1
3.	dTDP-D-glucose 4,6-dehydratase, putative	1	1
4.	glucosamine 6-phosphate N-acetyltransferase. putative	1	1
5.	multidrug resistance-associated protein, putative	1	1
6.	poly(ADP-ribose) polymerase, putative	1	1
7.	protein with DnaJ and myb domains	1	1
8.	Viral A-type inclusion protein repeat, putative	1	1
	Total	**8**	**8**

Appendix Table 1.5: Functional genes with transcriptomic upregulation ($log_2FC \geq 2$), only in virulent HM-1:IMSS strain (n=43). These upregulated transcripts can be assigned into 40 functional gene annotations as listed below.

group	Functional gene annotation	Number of upregulated genes, $log_2FC \geq 2$
1.	actin	3
2.	high mobility group (HMG) box domain-containing protein	2
3.	3' exoribonuclease family protein	1
4.	60S ribosomal protein L30, putative	1
5.	60S ribosomal protein L4, putative	1
6.	AIG family protein	1
7.	aldehyde-alcohol dehydrogenase 2, putative	1
8.	ARP2/3 complex 21 kDa subunit, putative	1
9.	ATP-binding cassette, sub-family C, putative	1
10.	bacterial transferase hexapeptide family protein	1
11.	bifunctional short chain isoprenyl diphosphate synthase, putative	1
12.	calmodulin, putative	1
13.	cortexillin II, putative	1
14.	cysteine desulfurase, putative	1
15.	deoxyuridine 5'-triphosphate nucleotidohydrolase domain-containing protein	1
16.	diacylglycerol kinase, putative	1
17.	DNA methyltransferase, putative	1
18.	DNA mismatch repair protein mutS, putative	1
19.	dual specificity protein phosphatase, putative	1
20.	F-box domain-containing protein	1
21.	formate/nitrite transporter family protein, putative	1
22.	glutamine synthetase, putative	1
23.	heat shock protein 90, putative	1
24.	I/LWEQ domain protein	1
25.	immediate-early protein, putative	1
26.	kinase, PfkB family	1
27.	nucleosome-binding protein 1, putative	1
28.	Nucleotide-binding protein , putative	1
29.	RecF/RecN/SMC domain-containing protein	1
30.	Rho family GTPase	1

Appendix Table 1.5: Functional genes with transcriptomic upregulation ($log_2FC \geq 2$), only in virulent HM-1:IMSS strain (n=43). **(Continued)**

group	Functional gene annotation	Number of upregulated genes, $log_2FC \geq 2$
31.	Rho guanine nucleotide exchange factor, putative	1
32.	serine-rich protein C30B4.01c precursor, putative	1
33.	small GTPase RhoA, putative	1
34.	sulfotransferase, putative	1
35.	suppressor protein SRP40, putative	1
36.	transcription initiation factorTFIID subunitTaf73, putative	1
37.	transketolase, chloroplast, putative	1
38.	translation initiation factor eIF-1A, putative	1
39.	ubiquitin-conjugating enzyme family protein	1
40.	vacuolar sorting protein 26, putative	1
	Total	**43**

Appendix Table 1.6: Functional genes with transcriptomic upregulation ($log_2FC \geq 2$), only in virulent PVBM08B strain (n=9). These upregulated transcripts can be assigned into 9 functional gene annotations as listed below.

group	Functional gene annotation	Number of upregulated genes, $log_2FC \geq 2$
1.	20 kDa antigen, putative	1
2.	actinin-like protein, putative	1
3.	calcineurin catalytic subunit A, putative	1
4.	casein kinase II regulatory subunit family protein	1
5.	cell division control protein 42, putative	1
6.	glutamic acid-rich protein, putative	1
7.	mannosyltransferase, putative	1
8.	valyl-tRNA synthetase, putative	1
9.	viral IAP-associated factor homolog, putative	1
	Total	**9**

Appendix Table 1.7: Functional genes with transcriptomic upregulation ($log_2FC \geq 2$), only in virulent IULA:1092:1 strain (n=45). These upregulated transcripts can be assigned into 41 functional gene annotations as listed below.

group	Functional gene annotation	Number of upregulated genes, $log_2FC \geq 2$
1.	cyst wall-specific glycoprotein Jacob	3
2.	60S ribosomal protein L3, putative	2
3.	zinc finger domain-containing protein	2
4.	Acid sphingomyelinase-like phosphodiesterase, putative	1
5.	beta-amylase, putative	1
6.	caldesmon, putative	1
7.	deoxyuridine 5'-triphosphate nucleotidohydrolase, mitochondrial precursor, putative	1
8.	DNA repair helicase, putative	1
9.	dynamin-1-like protein, putative	1
10.	dynamin-like protein	1
11.	glycosyltransferase, putative	1
12.	HAD hydrolase, family IA, variant 3	1
13.	heat shock protein70, hsp70A2, putative	1
14.	heat shock transcription factor, putative	1
15.	hemolysin-3, putative	1
16.	histone H3, putative	1
17.	homeobox protein, putative	1
18.	interaptin, putative	1
19.	iron-containing superoxide dismutase	1
20.	malate dehydrogenase, putative	1
21.	midasin, putative	1
22.	mitotic chromosome and X-chromosome-associated protein, putative	1
23.	Mob1/phocein family protein	1
24.	molybdenum cofactor sulfurase, putative	1
25.	phospholipase D, putative	1
26.	phospholipid-transporting P-type ATPase, putative	1
27.	PQ loop repeat protein	1
28.	pumilio family RNA-binding protein	1

Appendix Table 1.7: Functional genes with transcriptomic upregulation ($log_2FC \geq 2$) in virulent IULA:1092:1 strain (n=45). **(Continued)**

group	Functional gene annotation	Number of upregulated genes, $log_2FC \geq 2$
29.	Ras family protein	1
30.	RNA-binding protein, putative	1
31.	rubrerythrin, putative	1
32.	S-adenosylmethionine synthetase, putative	1
33.	serine palmitoyltransferase, putative	1
34.	serine/threonine- protein phosphatase PP-Z, putative	1
35.	Signal recognition particle receptor alpha subunit, putative	1
36.	syntaxin, putative	1
37.	transitional endoplasmic reticulum ATPase, putative	1
38.	transporter, major facilitator family	1
39.	tRNA-Glu (anticodon: TTC)	1
40.	U2 snRNP auxiliary factor small subunit, putative	1
41.	villidin, putative	1
	Total	**45**

Appendix Table 2.1: Functional genes with transcriptomic downregulation ($log_2FC \leq -2$) in all three virulent strains: HM-1:IMSS (n=120), PVBM08B (n=155) and IULA:1092:1 (n=123). These downregulated transcripts can be assigned into 16 functional annotations as listed below.

group	Functional gene annotation	Number of downregulated genes, $log_2FC \leq -2$		
		HM-1:IMSS	PVBM08B	IULA:1092:1
1.	AIG1 family protein	4	7	5
2.	myb-like DNA-binding domain-containing protein	3	6	4
3.	WD domain-containing protein	3	3	2
4.	leucine-rich repeat protein, BspA family	3	1	4
5.	protein kinase domain-containing protein	2	2	1
6.	protein tyrosine kinase domain-containing protein	2	1	1
7.	rodhanase-like domain-containing protein	1	2	2
8.	surface antigen ariel1, putative	1	1	4
9.	60S ribosomal protein L38, putative	1	1	1
10.	longevity-assurance family protein	1	1	1
11.	metallo-beta-lactamase superfamily protein	1	1	1
12.	nuclear movement protein, putative	1	1	1
13.	RhoGAP domain-containing protein	1	1	1
14.	tyrosine kinase, putative	1	1	1
15.	ubiquitin-conjugating enzyme family protein	1	1	1
16.	hypothetical protein	94	125	93
	Total	**120**	**155**	**123**

Appendix Table 2.2: Functional genes with transcriptomic downregulation ($log_2FC \leq -2$), only in two virulent strains: HM-1:IMSS (n=10) and PVBM08B (n=9). These downregulated transcripts can be assigned into 6 functional gene annotations as listed below.

group	Functional gene annotation	Number of downregulated genes, $log_2FC \leq -2$	
		HM-1:IMSS	PVBM08B
1.	cyst wall-specific glycoprotein Jacob	3	3
2.	Rab family GTPase	3	2
3.	chitinase, putative	1	1
4.	dual specificity protein phosphatase, putative	1	1
5.	Ras GTPase-activating protein, putative	1	1
6.	serine-rich 25 kDa antigen protein, putative	1	1
	Total	**10**	**9**

Appendix Table 2.3: Functional genes with transcriptomic downregulation ($log_2FC \leq -2$), only in two virulent strains: HM-1:IMSS (n=15) and IULA:1092:1 (n=16). These downregulated transcripts can be assigned into 11 functional gene annotations as listed below.

group	Functional gene annotation	Number of downregulated genes, $log_2FC \leq -2$	
		HM-1:IMSS	IULA:1092:1
1.	galactose-specific lectin light subunit, putative	2	3
2.	cysteine proteinase, putative	2	2
3.	methionine gamma-lyase	2	1
4.	serine/threonine protein kinase, putative	2	1
5.	heat shock protein 70, putative	1	2
6.	Ras family GTPase	1	2
7.	N-system amino acid transporter 1, putative	1	1
8.	Rap/Ran GTPase-activating protein, putative	1	1
9.	Ribosomal protein S30, putative	1	1
10.	RNA recognition motif domain-containing protein	1	1
11.	TBC domain-containing protein	1	1
	Total	**15**	**16**

Appendix Table 2.4: Functional genes with transcriptomic downregulation ($log_2FC \leq -2$), only in two virulent strains: PVBM08B (n=7) and IULA:1092:1 (n=5). These downregulated transcripts can be assigned into 5 functional gene annotations as listed below.

group	Functional gene annotation	Number of downregulated genes, $log_2FC \leq -2$	
		PVBM08B	IULA:1092:1
1.	acetyltransferase, GNAT family	2	1
2.	thioredoxin, putative	2	1
3.	Brix domain-containing protein 1, putative	1	1
4.	U5 small nuclear ribonucleoprotein subunit, putative	1	1
5.	zinc finger domain-containing protein	1	1
	Total	**7**	**5**

Appendix Table 2.5: Functional genes with transcriptomic downregulation ($log_2FC \leq -2$), only in virulent HM-1:IMSS strain (n=16). These downregulated transcripts can be assigned into 16 functional gene annotations as listed below.

group	Functional gene annotation	Number of downregulated genes, $log_2FC \leq -2$
1.	1-O-acylceramide synthase precursor, putative	1
2.	ADP-ribosylation factor 1, putative	1
3.	ADP-ribosylation factor, putative	1
4.	aldehyde-alcohol dehydrogenase 2, putative	1
5.	cysteine surface protein, putative	1
6.	EF-hand calcium-binding domain-containing protein	1
7.	heat shock protein 90, putative	1
8.	leucine-rich repeat-containing protein	1
9.	methylene-fatty-acyl-phospholipid synthase, putative	1
10.	nitroreductase family protein	1
11.	peroxiredoxin	1
12.	ser/thr protein phosphatase family protein	1
13.	steroid 5-alpha reductase, putative	1
14.	transporter, major facilitator family	1
15.	tRNA -methyltransferase catalytic subunit, putative	1
16.	tyrosine- protein kinase 2, putative	1
	Total	**16**

Appendix Table 2.6: Functional genes with transcriptomic downregulation ($log_2FC \leq -2$), only in virulent PVBM08B strain (n=49). These downregulated transcripts can be assigned into 44 functional gene annotations as listed below.

group	Functional gene annotation	Number of downregulated genes, $log_2FC \leq -2$
1.	galactose-inhibitable lectin 170 kDa subunit, putative	3
2.	mucin-like protein 1 precursor, putative	3
3.	acetyltransferase, putative	2
4.	40S ribosomal protein S4, putative	1
5.	60S ribosomal protein L37	1
6.	actin, putative	1
7.	actinin-like protein, putative	1
8.	acyl-CoA synthetase, putative	1
9.	alkyl sulfatase, putative	1
10.	amino acid transporter, putative	1
11.	ATP-binding cassette protein, putative	1
12.	calcineurin catalytic subunit A, putative	1
13.	chitinase Jessie, putative	1
14.	cysteine proteinase, pseudogene	1
15.	diaphanous protein, putative	1
16.	DNA mismatch repair protein mutL, putative	1
17.	dTDP-D-glucose 4,6-dehydratase, putative	1
18.	elongation factor 1-alpha 1	1
19.	elongation factor 2	1
20.	F-actin capping protein subunit beta, putative	1
21.	glucosamine--fructose-6-phosphate aminotransferase, putative	1
22.	glycogenphosphorylase, putative	1
23.	grainin, putative	1
24.	heat shock protein 70, mitochondrial, putative	1
25.	heat shock protein, putative	1
26.	hypothetical transmembrane protein	1
27.	inositol polyphosphate kinase, putative	1
28.	NAD-specific glutamate dehydrogenase, putative	1
29.	NADPH-dependent FMN reductase domain-containing protein	1
30.	peroxiredoxin, putative	1

Appendix Table 2.6: Functional genes with transcriptomic downregulation ($log_2FC \leq -2$), only in virulent PVBM08B strain. **(Continued)**

group	Functional gene annotation	Number of downregulated genes, $log_2FC \leq -2$
31.	phosphoserine aminotransferase, putative	1
32.	plasma membrane calcium-transporting ATPase, putative	1
33.	Ras family GTPase, pseudogene	1
34.	Ras guanine nucleotide exchange factor, putative	1
35.	S1 RNA-binding domain-containing protein	1
36.	serine/threonine protein phosphatase PP2A catalytic subunit, putative	1
37.	serine/threonine- protein kinase C823.03, putative	1
38.	sucrose transporter, putative	1
39.	syntaxin, putative	1
40.	transketolase, putative	1
41.	translation initiation factor 4e, putative	1
42.	type A flavoprotein, putative	1
43.	ubiquitin carboxyl-terminal hydrolase domain-containing protein	1
44.	USP6 N-terminal-like protein, putative	1
	Total	**49**

Appendix Table 2.7: Functional genes with transcriptomic downregulation ($log_2FC \leq -2$), only in virulent IULA:1092:1 strain (n=17). These downregulated transcripts can be assigned into 16 functional gene annotations as listed below.

group	Functional gene annotation	Number of downregulated genes, $log_2FC \leq -2$
1.	protein kinase, putative	2
2.	calcium-binding protein 1 (EhCBP1)	1
3.	carbohydrate degrading enzyme, putative	1
4.	carbonic anhydrase, putative	1
5.	casein kinase II regulatory subunit family protein	1
6.	cyclin family protein	1
7.	DNA-directed RNA polymerase subunit N, putative	1
8.	endo-1,4-beta-xylanase, putative	1
9.	endonuclease V, putative	1
10.	high-affinity potassium uptake transporter, putative	1
11.	hydrolase, carbon-nitrogen family	1
12.	N-acetylmuraminidase pseudogene	1
13.	NAD(P) transhydrogenase subunit alpha, putative	1
14.	NLI interacting factor-like phosphatase domain-containing protein	1
15.	Rho family GTPase	1
16.	synapsin, putative	1
	Total	**17**

Appendix Table 3: Summary of intraspecific single nucleotide polymorphisms (SNPs) found in 98 DE genes retrieved from Figure 2.19, across all *E. histolytica* strains available in the AmoebaDB version 4.2 (http://amoebadb.org/amoeba/)

AmoebaDB_ID	Annotation	Total SNPs	NonSynonymous SNPS	Synonymous SNPs	Non-coding SNPs	Stop Codon SNPs	Nonsyn/Syn SNP ratio	SNPs per kb (CDS)
EHI_012330	serine-threonine-isoleucine rich protein, putative	80	54	26	0	0	2.08	10.14
EHI_025700	serine-threonine-isoleucine rich protein, putative	61	40	21	0	0	1.9	7.81
EHI_004340	serine-threonine-isoleucine rich protein, putative	56	42	14	0	0	3	7.89
EHI_018010	DNA polymerase, putative	32	12	20	0	0	0.6	13.10
EHI_164190	DNA polymerase, putative	30	14	15	0	1	0.93	7.81
EHI_150770	heat shock protein 70, putative	30	22	8	0	0	2.75	19.04
EHI_129470	AIG1 family protein	25	17	8	0	0	2.13	28.53
EHI_119040	AIG1 family protein, putative	24	19	5	0	0	3.8	25.88
EHI_018840	leucine-rich repeat protein, BspA family	21	13	8	0	0	1.63	12.5
EHI_102380	leucine-rich repeat protein, BspA family	21	15	5	0	1	3	19.66
EHI_072850	AIG1 family protein, putative	20	10	4	6	0	2.5	40.65
EHI_127710	leucine-rich repeat protein, BspA family	15	9	6	0	0	1.5	10.14
EHI_133950	heat shock protein 70, putative	12	5	7	0	0	0.71	7.63
EHI_022600	NADPH-dependent FMN reductase domain-containing protein	12	4	7	0	1	0.57	19.80
EHI_005260	surface antigen ariel1, putative	12	7	5	0	0	1.4	10.78
EHI_102600	AIG1 family protein	11	10	1	0	0	10	20.48
EHI_069320	C2 domain-containing protein	11	3	0	8	0	0	19.40
EHI_127700	heat shock protein 70, mitochondrial, putative	11	7	3	0	1	2.33	6.12
EHI_142700	endonuclease V, putative	10	6	3	1	0	2	13.88
EHI_034610	leucine-rich repeat protein, BspA family	8	8	0	0	0	0	7.51

Appendix Table 3: Summary of intraspecific single nucleotide polymorphisms (SNPs) found in 98 DE genes retrieved from Figure 2.19, across all *E. histolytica* strains. **(Continued)**

AmoebaDB_ID	Annotation	Total SNPs	NonSynonymous SNPS	Synonymous SNPs	Non-coding SNPs	Stop Codon SNPs	Nonsyn/Syn SNP ratio	SNPs per kb (CDS)
EHI_105370	leucine-rich repeat protein, BspA family	8	6	2	0	0	3	10.41
EHI_123830	DNA mismatch repair protein Msh2, putative	7	1	6	0	0	0.17	3.69
EHI_126550	AIG1 family protein, putative	6	4	1	0	1	4	10.52
EHI_161300	leucine-rich repeat protein, BspA family	6	6	0	0	0	0	11.36
EHI_084730	multidrug resistance-associated protein, putative	6	1	5	0	0	0.2	10.25
EHI_186600	P-glycoprotein-2, putative	6	4	2	0	0	2	4.83
EHI_058520	Ras family GTPase	6	6	0	0	0	0	10.10
EHI_055140	ethanolamine phophotransferase, putative	5	2	3	0	0	0.67	12.25
EHI_058550	Ras family GTPase, pseudogene	5	4	0	0	1	0	8.29
EHI_050150	HEAT repeat domain-containing protein	4	1	3	0	0	0.33	1.94
EHI_145840	peroxiredoxin	4	2	1	1	0	2	5.69
EHI_021780	heat shock protein 70, putative	3	3	0	0	0	0	3.10
EHI_022730	signal recognition particle 54 kDa protein, putative	3	0	3	0	0	0	3.84
EHI_031350	60S ribosome subunit biogenesis protein NIP7, putative	2	1	1	0	0	1	3.62
EHI_059860	C2 domain-containing protein	2	1	1	0	0	1	3.18
EHI_021490	coiled-coil domain-containing protein 25, putative	2	1	1	0	0	1	3.94
EHI_164890	Rab GDP dissociation inhibitor alpha, putative	2	1	1	0	0	1	1.51
EHI_199570	RhoGAP domain-containing protein	2	0	2	0	0	0	1.50
EHI_049170	splicing factor 3B subunit 1, putative	2	1	1	0	0	1	0.72

Appendix Table 3: Summary of intraspecific single nucleotide polymorphisms (SNPs) found in 98 DE genes retrieved from Figure 2.19, across all *E. histolytica* strains. **(Continued)**

AmoebaDB_ID	Annotation	Total SNPs	NonSynonymous SNPS	Synonymous SNPs	Non-coding SNPs	StopCodon SNPs	Nonsyn/Syn SNP ratio	SNPs per kb (CDS)
EHI_093580	60S ribosomal protein L6, putative	1	1	0	0	0	0	1.62
EHI_118600	calcineurin catalytic subunit A, putative	1	1	0	0	0	0	0.65
EHI_051870	peptidyl-prolyl cis-trans isomerase, FKBP-type, putative	1	1	0	0	0	0	0.87
EHI_077000	pre-mRNA cleavage factor I 25 kDa subunit, putative	1	1	0	0	0	0	1.40
EHI_048860	U3 small nucleolar ribonucleo protein MPP10, putative	1	1	0	0	0	0	1.54
EHI_092070	WD domain-containing protein	1	1	0	0	0	0	1.62
EHI_148530	leucine-rich repeat protein, BspA family	0	0	0	0	0	0	0
EHI_144590	protein kinase domain-containing protein	0	0	0	0	0	0	0
EHI_128800	hypothetical protein	67	52	15	0	0	3.47	17.69
EHI_077290	hypothetical protein, conserved	54	30	24	0	0	1.25	19.39
EHI_087110	hypothetical protein, conserved	41	38	3	0	0	12.67	47.45
EHI_130550	hypothetical protein	36	15	21	0	0	0.71	23.62
EHI_047510	hypothetical protein	30	21	9	0	0	2.33	18.90
EHI_062320	hypothetical protein	26	16	9	0	1	1.78	18.47
EHI_028770	hypothetical protein	13	9	4	0	0	2.25	18.59
EHI_097750	hypothetical protein, conserved	12	4	8	0	0	0.5	11.2
EHI_001730	hypothetical protein	11	7	4	0	0	1.75	15.60
EHI_011260	hypothetical protein, conserved	11	8	3	0	0	2.67	15.27
EHI_193690	hypothetical protein	11	9	2	0	0	4.5	9.96
EHI_193790	hypothetical protein, conserved	11	5	4	1	1	1.25	16.51
EHI_017780	hypothetical protein	10	3	7	0	0	0.43	9.36
EHI_145460	hypothetical protein	10	2	8	0	0	0.25	19.60

Appendix Table 3: Summary of intraspecific single nucleotide polymorphisms (SNPs) found in 98 DE genes retrieved from Figure 2.19, across all *E. histolytica* strains. **(Continued)**

AmoebaDB_ID	Annotation	Total SNPs	NonSynonymous SNPS	Synonymous SNPs	Non-coding SNPs	StopCodon SNPs	Nonsyn/Syn SNP ratio	SNPs per kb (CDS)
EHI_077510	hypothetical protein	10	6	4	0	0	1.5	9.08
EHI_101410	hypothetical protein	9	4	5	0	0	0.8	13.95
EHI_033890	hypothetical protein	9	5	4	0	0	1.25	13.95
EHI_142680	hypothetical protein	8	3	5	0	0	0.6	12.40
EHI_046150	hypothetical protein	8	3	2	1	2	1.5	23.59
EHI_142690	hypothetical protein	8	7	1	0	0	7	17.77
EHI_112390	hypothetical protein	7	4	3	0	0	1.33	12.08
EHI_033450	hypothetical protein	7	3	3	1	0	1	6.78
EHI_145490	hypothetical protein	7	6	1	0	0	6	7.77
EHI_087740	hypothetical protein	6	0	6	0	0	0	2.98
EHI_089460	hypothetical protein	6	3	2	0	1	1.5	15.03
EHI_101400	hypothetical protein	5	5	0	0	0	0	11.11
EHI_081110	hypothetical protein	5	2	3	0	0	0.67	5.86
EHI_004460	hypothetical protein	4	3	1	0	0	3	5.31
EHI_166040	hypothetical protein	4	0	4	0	0	0	4.61
EHI_196070	hypothetical protein	4	3	0	1	0	0	3.68
EHI_049760	hypothetical protein	3	0	3	0	0	0	2.00
EHI_037690	hypothetical protein	3	1	2	0	0	0.5	8.47
EHI_039590	hypothetical protein, conserved	2	1	1	0	0	1	1.37
EHI_068610	hypothetical protein	2	1	0	0	1	0	7.49
EHI_080880	hypothetical protein	2	2	0	0	0	0	3.31
EHI_083380	hypothetical protein, conserved	2	0	2	0	0	0	1.46
EHI_097490	hypothetical protein	2	1	1	0	0	1	1.56
EHI_104220	hypothetical protein	2	1	1	0	0	1	2.87
EHI_120250	hypothetical protein	2	1	1	0	0	1	6.06
EHI_062300	hypothetical protein	2	1	1	0	0	1	6.11
EHI_074080	hypothetical protein	1	0	1	0	0	0	3.14
EHI_079240	hypothetical protein	1	1	0	0	0	0	1.62
EHI_022300	hypothetical protein, conserved	1	0	1	0	0	0	1.17

Appendix Table 3: Summary of intraspecific single nucleotide polymorphisms (SNPs) found in 98 DE genes retrieved from Figure 2.19, across all *E. histolytica* strains. **(Continued)**

AmoebaDB_ID	Annotation	Total SNPs	NonSynonymous SNPS	Synonymous SNPs	Non-coding SNPs	Stop Codon SNPs	Nonsyn/Syn SNP ratio	SNPs per kb (CDS)
EHI_118230	hypothetical protein	1	0	1	0	0	0	2.38
EHI_099710	hypothetical protein	1	1	0	0	0	0	3.03
EHI_146120	hypothetical protein	1	1	0	0	0	0	1.57
EHI_192240	hypothetical protein	1	0	1	0	0	0	2.12
EHI_067090	hypothetical protein	1	1	0	0	0	0	1.56
EHI_152360	hypothetical protein	0	0	0	0	0	0	0
EHI_062310	hypothetical protein	0	0	0	0	0	0	0
EHI_039600	hypothetical protein	0	0	0	0	0	0	0

Appendix Table 4: Gene Ontology Biological Process terms that are enriched in 1,162 upregulated DE transcripts in the three virulent *E. histolytica* strains, i.e. PVBM08B, HM-1:IMSS and IULA:1092:1.

Term_ID	description	Genes in the bkgd	Genes in the sample	Percent of bkgd Genes in the result	Fold enrichment	Odds ratio	*P*-value	FDR-adjusted *P*-value (Benjamini)	Bonferroni
GO:0030163	protein catabolic process	72	29	40.3	2.57	2.63	4.39E-05	1.51E-03	1.54E-03
GO:0044257	cellular protein catabolic process	65	25	38.5	2.45	2.5	2.59E-04	1.51E-03	9.07E-03
GO:0043632	modification-dependent macromolecule catabolic process	65	25	38.5	2.45	2.5	2.59E-04	1.51E-03	9.07E-03
GO:0019941	modification-dependent protein catabolic process	65	25	38.5	2.45	2.5	2.59E-04	1.51E-03	9.07E-03
GO:0006511	ubiquitin-dependent protein catabolic process	65	25	38.5	2.45	2.5	2.59E-04	1.51E-03	9.07E-03
GO:0051603	proteolysis involved in cellular protein catabolic process	65	25	38.5	2.45	2.5	2.59E-04	1.51E-03	9.07E-03
GO:0009057	macromolecule catabolic process	88	30	34.1	2.17	2.22	3.62E-04	1.81E-03	1.27E-02
GO:0044265	cellular macromolecule catabolic process	68	25	36.8	2.34	2.39	4.49E-04	1.96E-03	1.57E-02
GO:0006310	DNA recombination	14	10	71.4	4.55	4.6	6.46E-04	2.51E-03	2.26E-02
GO:0071840	cellular component organisation or biogenesis	173	47	27.2	1.73	1.78	7.76E-04	2.72E-03	2.72E-03
GO:0016043	cellular component organisation	131	38	29	1.85	1.89	8.97E-04	2.85E-03	3.14E-02
GO:0006996	organelle organisation	103	30	29.1	1.86	1.89	2.87E-03	8.38E-03	1.01E-01
GO:0006508	proteolysis	162	39	24.1	1.53	1.56	1.17E-02	2.62E-02	4.09E-01
GO:0006412	translation	270	59	21.9	1.39	1.43	1.26E-02	2.62E-02	4.42E-01
GO:0007015	actin filament organisation	17	8	47.1	3	3.02	1.43E-02	2.62E-02	4.99E-01
GO:0044419	interspecies interaction between organisms	4	4	100	6.37	6.4	1.50E-02	2.62E-02	5.24E-01
GO:0016032	viral process	4	4	100	6.37	6.4	1.50E-02	2.62E-02	5.24E-01
GO:0044764	multiorganism cellular process	4	4	100	6.37	6.4	1.50E-02	2.62E-02	5.24E-01

Appendix Table 4: Gene Ontology Biological Process terms that are enriched in 1,162 upregulated DE transcripts in the three virulent *E. histolytica* strains. **(Continued)**

Term_ID	description	Genes in the bkgd	Genes in the sample	Percent of bkgd Genes in the result	Fold enrichment	Odds ratio	*P*-value	FDR-adjusted *P*-value (Benjamini)	Bonferroni
GO:0044403	symbiosis, encompassing mutualism through parasitism	4	4	100	6.37	6.4	1.50E-02	2.62E-02	5.24E-01
GO:0051704	multi-organism process	4	4	100	6.37	6.4	1.50E-02	2.62E-02	5.24E-01
GO:0033554	cellular response to stress	61	18	29.5	1.88	1.9	1.68E-02	2.80E-02	5.89E-01
GO:0051276	chromosome organisation	43	14	32.6	2.07	2.1	1.76E-02	2.80E-02	6.17E-01
GO:0006281	DNA repair	58	17	29.3	1.87	1.89	2.08E-02	3.17E-02	7.30E-01
GO:0007010	cytoskeleton organisation	45	14	31.1	1.98	2	2.36E-02	3.34E-02	8.25E-01
GO:0030029	actin filament-based process	32	11	34.4	2.19	2.21	2.51E-02	3.34E-02	8.79E-01
GO:0030036	actin cytoskeleton organisation	32	11	34.4	2.19	2.21	2.51E-02	3.34E-02	8.79E-01
GO:0006974	cellular response to DNA damage stimulus	60	17	28.3	1.8	1.82	2.66E-02	3.34E-02	9.31E-01
GO:0008154	actin polymerisation or depolymerisation	12	6	50	3.18	3.2	2.67E-02	3.34E-02	9.36E-01
GO:0006259	DNA metabolic process	137	32	23.4	1.49	1.51	2.91E-02	3.51E-02	1.00E+00
GO:0006950	response to stress	66	18	27.3	1.74	1.76	3.04E-02	3.54E-02	1.00E+00
GO:0000278	mitotic cell cycle	3	3	100	6.37	6.39	3.62E-02	3.84E-02	1.00E+00
GO:0006020	inositol metabolic process	3	3	100	6.37	6.39	3.62E-02	3.84E-02	1.00E+00
GO:0007067	mitotic nuclear division	3	3	100	6.37	6.39	3.62E-02	3.84E-02	1.00E+00
GO:0044249	cellular biosynthetic process	496	94	19	1.21	1.24	4.53E-02	4.66E-02	1.00E+00
GO:1901576	organic substance biosynthetic process	504	95	18.8	1.2	1.23	4.90E-02	4.90E-02	1.00E+00

Appendix Table 5: Gene Ontology Cellular Component terms that are enriched in 1,162 upregulated DE transcripts in the three virulent *E. histolytica* strains.

Term_ID	description	Genes in the bkgd	Genes in the sample	Percent of bkgd Genes in the result	Fold enrichment	Odds ratio	*P*-value	FDR-adjusted *P*-value (Benjamini)	Bonferroni
GO:0000502	proteasome complex	24	15	62.5	3.98	4.04	8.92E-05	1.34E-03	1.34E-03
GO:0043228	non-membrane-bounded organelle	283	69	24.4	1.55	1.61	6.96E-04	3.48E-03	1.04E-02
GO:0043232	intracellular non-membrane-bounded organelle	283	69	24.4	1.55	1.61	6.96E-04	3.48E-03	1.04E-02
GO:0005839	proteasome core complex	19	11	57.9	3.69	3.73	1.25E-03	3.75E-03	1.87E-02
GO:0030529	ribonucleoprotein complex	223	56	25.1	1.6	1.65	1.25E-03	3.75E-03	1.88E-02
GO:0032991	macromolecular complex	466	100	21.5	1.37	1.43	2.09E-03	5.23E-03	3.14E-02
GO:0044424	intracellular part	744	146	19.6	1.25	1.31	4.5E-03	9.82E-03	6.87E-02
GO:0019773	proteasome core complex, alpha-sububit complex	11	7	63.6	4.05	4.08	6.73E-03	1.21E-02	1.01E-01
GO:0005840	ribosome	203	48	23.6	1.51	1.54	7.28E-03	1.21E-02	1.09E-01
GO:0005694	chromosome	39	13	33.3	2.12	2.14	1.89E-02	2.83E-02	2.83E-01
GO:0005838	proteasome regulatory particle	5	4	80.0	5.1	5.12	2.41E-02	3.01E-02	3.61E-01
GO:0022624	proteasome accessory complex	5	4	80.0	5.1	5.12	2.41E-02	3.01E-02	3.61E-01
GO:0043226	organelle	555	105	18.9	1.21	1.24	3.66E-02	3.92E-02	5.49E-01
GO:0043229	intracellular organelle	555	105	18.9	1.21	1.24	3.66E-02	3.92E-02	5.49E-01
GO:0015629	actin cytoskeleton	14	6	42.9	2.73	2.74	4.41E-02	4.41E-02	6.62E-01

Appendix Table 6: Gene Ontology Molecular Function terms that are enriched in 1,162 upregulated transcripts in the three virulent *E. histolytica* strains.

Term_ID	description	Genes in the bkgd	Genes in the sample	Percent of bkgd Genes in the result	Fold enrichment	Odds ratio	*P*-value	FDR-adjusted *P*-value (Benjamini)	Bonferroni
GO:0005515	protein binding	951	189	19.9	1.27	1.36	5.99E-04	4.99E-03	7.19E-03
GO:0070003	threonine-type peptidase activity	19	11	57.9	3.69	3.73	1.25E-03	4.99E-03	1.50E-02
GO:0004298	threonine-type endopeptidase activity	19	11	57.9	3.69	3.73	1.25E-03	4.99E-03	1.50E-02
GO:0016887	ATPase activity	170	43	25.3	1.61	1.65	4.01E-03	1.20E-02	4.81E-02
GO:0003735	structural constituent of ribosome	211	50	23.7	1.51	1.55	6.01E-03	1.44E-02	7.22E-02
GO:0005198	structural molecule activity	226	52	23.0	1.47	1.5	8.35E-03	1.67E-02	1.00E-01
GO:0003779	actin binding	50	16	32.0	2.04	2.06	1.32E-02	2.26E-02	1.58E-01
GO:0016874	ligase activity	99	25	25.3	1.61	1.63	2.51E-02	3.77E-02	3.02E-01
GO:0008092	cytoskeletal protein binding	56	16	28.6	1.82	1.84	2.92E-02	3.89E-02	3.50E-01
GO:0003950	NAD+ ADP-ribosyltransferase activity	6	4	66.7	4.25	4.26	3.59E-02	4.15E-02	4.31E-01
GO:0042623	ATPase activity, coupled	130	30	23.1	1.47	1.49	3.81E-02	4.15E-02	4.57E-01
GO:0005488	binding	2649	440	16.6	1.06	1.15	4.90E-02	4.90E-02	5.88E-01

Appendix Table 7: Gene Ontology Biological Process terms that are enriched in 997 downregulated DE transcripts in the three virulent *E. histolytica* strains.

Term_ID	description	Genes in the bkgd	Genes in the sample	Percent of bkgd Genes in the result	Fold enrichment	Odds ratio	*P*-value	FDR-adjusted *P*-value (Benjamini)	Bonferroni
GO:0065007	biological regulation	742	124	16.7	1.36	1.46	4.00E-04	3.66E-03	1.76E-02
GO:0023051	regulation of signaling	272	55	20.2	1.65	1.72	5.63E-04	3.66E-03	2.48E-02
GO:0010646	regulation of cell communication	272	55	20.2	1.65	1.72	5.63E-04	3.66E-03	2.48E-02
GO:0009966	regulation of signal transduction	272	55	20.2	1.65	1.72	5.63E-04	3.66E-03	2.48E-02
GO:0048583	regulation of response to stimulus	272	55	20.2	1.65	1.72	5.63E-04	3.66E-03	2.48E-02
GO:0050794	regulation of cellular process	725	120	16.6	1.35	1.44	7.14E-04	3.66E-03	3.14E-02
GO:0051056	regulation of small GTPase mediated signal transduction	269	54	20.1	1.64	1.7	7.35E-04	3.66E-03	3.23E-02
GO:1902531	regulation of intracellular signal transduction	269	54	20.1	1.64	1.7	7.35E-04	3.66E-03	3.23E-02
GO:0050789	regulation of biological process	726	120	16.5	1.35	1.44	7.48E-04	3.66E-03	3.29E-02
GO:0006793	phosphorus metabolic process	659	110	16.7	1.36	1.45	9.59E-04	3.84E-03	4.22E-02
GO:0006796	phosphate-containing compound metabolic process	659	110	16.7	1.36	1.45	9.59E-04	3.84E-03	4.22E-02
GO:0007154	cell communication	396	69	17.4	1.42	1.48	3.86E-03	1.38E-02	1.70E-01
GO:0035556	intracellular signal transduction	267	50	18.7	1.53	1.58	4.07E-03	1.38E-02	1.79E-01
GO:0007165	signal transduction	388	67	17.3	1.41	1.46	5.29E-03	1.53E-02	2.33E-01
GO:0044700	single organism signaling	389	67	17.2	1.4	1.46	5.57E-03	1.53E-02	2.45E-01
GO:0023052	signaling	389	67	17.2	1.4	1.46	5.57E-03	1.53E-02	2.45E-01
GO:0007264	small GTPase mediated signal transduction	245	45	18.4	1.5	1.54	8.50E-03	2.17E-02	3.74E-01
GO:0006468	protein phosphorylation	342	59	17.3	1.41	1.45	8.88E-03	2.17E-02	3.91E-01
GO:0016310	phosphorylation	355	60	16.9	1.38	1.42	1.20E-02	2.79E-02	5.30E-01
GO:0046578	regulation of Ras protein signal transduction	135	27	20.0	1.63	1.66	1.58E-02	2.83E-02	6.94E-01
GO:1900542	regulation of purine nucleotide metabolic process	68	16	23.5	1.92	1.95	1.80E-02	2.83E-02	7.91E-01

Appendix Table 7: Gene Ontology Biological Process terms that are enriched in 997 downregulated DE transcripts in the three virulent *E. histolytica* strains. **(Continued)**

Term_ID	description	Genes in the bkgd	Genes in the sample	Percent of bkgd Genes in the result	Fold enrichment	Odds ratio	*P*-value	FDR-adjusted *P*-value (Benjamini)	Bonferroni
GO:0051336	regulation of hydrolase activity	68	16	23.5	1.92	1.95	1.80E-02	2.83E-02	7.91E-01
GO:0006140	regulation of nucleotide metabolic process	68	16	23.5	1.92	1.95	1.80E-02	2.83E-02	7.91E-01
GO:0030811	regulation of nucleotide catabolic process	68	16	23.5	1.92	1.95	1.80E-02	2.83E-02	7.91E-01
GO:0009118	regulation of nucleoside metabolic process	68	16	23.5	1.92	1.95	1.80E-02	2.83E-02	7.91E-01
GO:0043087	regulation of GTPase activity	68	16	23.5	1.92	1.95	1.80E-02	2.83E-02	7.91E-01
GO:0033124	regulation of GTP catabolic process	68	16	23.5	1.92	1.95	1.80E-02	2.83E-02	7.91E-01
GO:0033121	regulation of purine nucleotide catabolic process	68	16	23.5	1.92	1.95	1.80E-02	2.83E-02	7.91E-01
GO:0031329	regulation of cellular catabolic process	69	16	23.2	1.89	1.92	2.00E-02	3.01E-02	8.80E-01
GO:0009894	regulation of catabolic process	70	16	22.9	1.86	1.89	2.22E-02	3.01E-02	9.76E-01
GO:0008033	tRNA processing	36	10	27.8	2.27	2.29	2.41E-02	3.01E-02	1.00E+0
GO:0051174	regulation of phosphorus metabolic process	71	16	22.5	1.84	1.86	2.45E-02	3.01E-02	1.00E+0
GO:0019220	regulation of phosphate metabolic process	71	16	22.5	1.84	1.86	2.45E-02	3.01E-02	1.00E+0
GO:0050790	regulation of catalytic activity	71	16	22.5	1.84	1.86	2.45E-02	3.01E-02	1.00E+0
GO:0006399	tRNA metabolic process	65	15	23.1	1.88	1.91	2.46E-02	3.01E-02	1.00E+0
GO:0032318	regulation of Ras GTPase activity	65	15	23.1	1.88	1.91	2.46E-02	3.01E-02	1.00E+0
GO:0065009	regulation of molecular function	72	16	22.2	1.81	1.84	2.71E-02	3.22E-02	1.00E+0
GO:0034660	ncRNA metabolic process	79	17	21.5	1.75	1.78	2.93E-02	3.29E-02	1.00E+0
GO:0032483	regulation of Rab protein signal transduction	49	12	24.5	2.0	2.02	2.99E-02	3.29E-02	1.00E+0
GO:0032313	regulation of Rab GTPase activity	49	12	24.5	2.0	2.02	2.99E-02	3.29E-02	1.00E+0
GO:0034470	ncRNA processing	50	12	24.0	1.96	1.98	3.36E-02	3.60E-02	1.00E+0

Appendix Table 7: Gene Ontology Biological Process terms that are enriched in 997 downregulated DE transcripts in the three virulent *E. histolytica* strains. **(Continued)**

Term_ID	description	Genes in the bkgd	Genes in the sample	Percent of bkgd Genes in the result	Fold enrichment	Odds ratio	*P*-value	FDR-adjusted *P*-value (Benjamini)	Bonferroni
GO:0006464	cellular protein modification process	474	72	15.2	1.24	1.27	4.52E-02	4.62E-02	1.00E+0
GO:0036211	protein modification process	474	72	15.2	1.24	1.27	4.52E-02	4.62E-02	1.00E+0
GO:0043412	macromolecule modification	505	76	15.0	1.23	1.26	4.70E-02	4.70E-02	1.00E+0

Appendix Table 8: Gene Ontology Molecular Function terms that are enriched in 997 downregulated DE transcripts in the three virulent *E. histolytica* strains.

Term_ID	description	Genes in the bkgd	Genes in the sample	Percent of bkgd Genes in the result	Fold enrichment	Odds ratio	*P*-value	FDR-adjusted *P*-value (Benjamini)	Bonferroni
GO:0005085	guanyl-nucleotide exchange factor activity	119	28	23.5	1.92	1.97	2.13E-03	2.58E-02	5.11E-02
GO:0030234	enzyme regulator activity	193	39	20.2	1.65	1.7	3.59E-03	2.58E-02	8.62E-02
GO:0016772	transferase activity, transferring phosphorus-containing groups	529	88	16.6	1.36	1.42	3.62E-03	2.58E-02	8.70E-02
GO:0008047	enzyme activator activity	182	36	19.8	1.61	1.65	6.77E-03	2.58E-02	1.63E-01
GO:0016740	transferase activity	730	113	15.5	1.26	1.33	7.86E-03	2.58E-02	1.89E-01
GO:0016773	phosphotransferase activity, alcohol group as acceptor	403	68	16.9	1.38	1.43	7.93E-03	2.58E-02	1.90E-01
GO:0016301	kinase activity	411	69	16.8	1.37	1.42	8.29E-03	2.58E-02	1.99E-01
GO:0004672	protein kinase activity	343	59	17.2	1.4	1.45	9.36E-03	2.58E-02	2.25E-01
GO:0043167	ion binding	1332	191	14.3	1.17	1.26	9.67E-03	2.58E-02	2.32E-01
GO:0060589	nucleoside-triphosphatase regulator activity	182	35	19.2	1.57	1.61	1.08E-02	2.60E-02	2.60E-01
GO:0008270	zinc ion binding	245	44	18.0	1.46	1.5	1.28E-02	2.79E-02	3.07E-01
GO:0005096	GTPase activator activity	180	34	18.9	1.54	1.57	1.48E-02	2.93E-02	3.56E-01
GO:0030695	GTPase regulator activity	181	34	18.8	1.53	1.57	1.59E-02	2.93E-02	3.81E-01
GO:0046914	transition metal ion binding	271	47	17.3	1.41	1.45	1.72E-02	2.96E-02	4.14E-01
GO:0043169	cation binding	389	62	15.9	1.3	1.34	2.93E-02	3.69E-02	7.02E-01
GO:0005097	Rab GTPase activator activity	49	12	24.5	2.0	2.02	2.99E-02	3.69E-02	7.18E-01
GO:0005083	small GTPase regulator activity	67	15	22.4	1.83	1.85	3.02E-02	3.69E-02	7.25E-01
GO:0008746	NAD(P)+ transhydrogenase activity	4	3	75.0	6.12	6.14	3.24E-02	3.69E-02	7.78E-01
GO:0008750	NAD(P)+ transhydrogenase (AB-specific) activity	4	3	75.0	6.12	6.14	3.24E-02	3.69E-02	7.78E-01
GO:0030246	carbohydrate binding	22	7	31.8	2.59	2.61	3.29E-02	3.69E-02	7.89E-01

Appendix Table 8: Gene Ontology Molecular Function terms that are enriched in 997 downregulated DE transcripts in the three virulent *E. histolytica* strains. **(Continued)**

Term_ID	description	Genes in the bkgd	Genes in the sample	Percent of bkgd Genes in the result	Fold enrichment	Odds ratio	*P*-value	FDR-adjusted *P*-value (Benjamini)	Bonferroni
GO:0005099	Ras GTPase activator activity	50	12	24.0	1.96	1.98	3.36E-02	3.69E-02	8.06E-01
GO:0046872	metal ion binding	378	60	15.9	1.29	1.33	3.38E-02	3.69E-02	8.11E-01
GO:0004849	uridine kinase activity	5	3	60.0	4.89	4.91	4.77E-02	4.77E-02	1.00E+0
GO:0016652	oxidoreductase activity, acting on NAD(P)H, NAD(P) as acceptor	5	3	60.0	4.89	4.91	4.77E-02	4.77E-02	1.00E+0

Appendix Table 9: REVIGO's summarisation of 35 upregulated biological process ontologies in the three virulent strains, i.e. PVBM08B, HM-1:IMSS and IULA:1092:1, compared to the nonvirulent Rahman strain. Higher frequency of proteins annotated in the UniProt database reflects a more general GO term. Twenty-three cluster representatives are shown in black letters and their cluster members are listed in gray italics and indented. The thirty-five terms could be summarised into twenty-one clusters and fifteen of which have only a single term.

Term_ID	description	frequency	plot_X	plot_Y	plot_size	log10_FDR	uniqueness	dispensability
GO:0006950	response to stress	4.12%	0.53	-0.541	6.312	-1.451	0.84	0
GO:0009057	macromolecule catabolic process	1.64%	-5.841	3.702	5.912	-2.7423	0.722	0
GO:0016032	viral process	1.74%	3.586	-3.804	5.938	-1.5817	0.758	0
└ GO:0044764	*multi-organism cellular process*	*2.06%*	*null*	*null*	*6.01*	*-1.5817*	*0.761*	*0.879*
└ GO:0044403	*symbiosis, encompassing mutualism through parasitism*	*1.78%*	*null*	*null*	*5.948*	*-1.5817*	*0.794*	*0.979*
└ GO:0044419	*interspecies interaction between organisms*	*1.78%*	*null*	*null*	*5.948*	*-1.5817*	*0.8*	*0.881*
GO:0016043	cellular component organisation	4.29%	0.238	7.208	6.329	-2.5452	0.733	0
└ GO:0006996	*organelle organisation*	*0.93%*	*null*	*null*	*5.666*	*-2.0768*	*0.712*	*0.704*
GO:0051704	multi-organism process	2.77%	5.03	1.372	6.14	-1.5817	0.944	0
GO:0071840	cellular component organisation or biogenesis	5.43%	1.473	-5.472	6.431	-2.5654	0.945	0
GO:0030029	actin filament-based process	0.08%	-0.89	-6.775	4.582	-1.4763	0.891	0.025
GO:0006020	inositol metabolic process	0.02%	4.572	-1.271	3.969	-1.4157	0.871	0.064
GO:0000278	mitotic cell cycle	0.09%	3.404	4.431	4.637	-1.4157	0.863	0.071
GO:0006310	DNA recombination	1.84%	-7.401	-0.095	5.962	-2.6003	0.752	0.185
GO:0006412	translation	4.70%	-6.213	-0.946	6.369	-1.5817	0.653	0.254
GO:0006259	DNA metabolic process	6.34%	-6.351	-0.024	6.499	-1.4547	0.746	0.379
GO:0044249	cellular biosynthetic process	28.21%	-5.393	-5.227	7.147	-1.3316	0.819	0.385
GO:0030163	protein catabolic process	0.36%	-6.804	3.474	5.251	-2.821	0.629	0.442
└ GO:0044265	*cellular macromolecule catabolic process*	*1.11%*	*null*	*null*	*5.742*	*-2.7077*	*0.6*	*0.798*

Appendix Table 9: REVIGO's summarisation of 35 upregulated biological process ontologies in the three virulent strains. **(Continued)**

Term_ID	description	frequency	plot_X	plot_Y	plot_size	log10_FDR	uniqueness	dispensability
GO:0006508	proteolysis	3.71%	-7.346	1.556	6.266	-1.5817	0.735	0.578
GO:0051276	chromosome organisation	0.34%	0.637	6.948	5.223	-1.5528	0.668	0.623
└ GO:0007015	*actin filament organisation*	*0.04%*	*null*	*null*	*4.285*	*-1.5817*	*0.658*	*0.951*
└ GO:0030036	*actin cytoskeleton organisation*	*0.07%*	*null*	*null*	*4.559*	*-1.4763*	*0.648*	*0.717*
└ GO:0007010	*cytoskeleton organisation*	*0.15%*	*null*	*null*	*4.876*	*-1.4763*	*0.682*	*0.76*
└ GO:0008154	*actin polymerisation or depolymerisation*	*0.03%*	*null*	*null*	*4.141*	*-1.4763*	*0.663*	*0.93*
GO:0006281	DNA repair	1.95%	-4.175	0.368	5.988	-1.4989	0.62	0.652
└ GO:0033554	*cellular response to stress*	*2.34%*	*null*	*null*	*6.065*	*-1.5528*	*0.784*	*0.795*
└ GO:0006974	*cellular response to DNA damage stimulus*	*1.98%*	*null*	*null*	*5.993*	*-1.4763*	*0.778*	*0.934*
GO:1901576	organic substance biosynthetic process	28.89%	-5.012	-5.487	7.157	-1.3098	0.841	0.663
GO:0051603	proteolysis involved in cellular protein catabolic process	0.22%	-6.643	3.011	5.036	-2.821	0.571	0.687
└ GO:0019941	*modification-dependent protein catabolic process*	*0.14%*	*null*	*null*	*4.826*	*-2.821*	*0.579*	*0.983*
└ GO:0044257	*cellular protein catabolic process*	*0.22%*	*null*	*null*	*5.039*	*-2.821*	*0.578*	*0.916*
└ GO:0006511	*ubiquitin-dependent protein catabolic process*	*0.11%*	*null*	*null*	*4.741*	*-2.821*	*0.584*	*0.944*
GO:0007067	mitotic nuclear division	0.05%	1.684	6.031	4.367	-1.4157	0.658	0.693
GO:0043632	modification-dependent macromolecule catabolic process	0.14%	-6.217	4.4	4.828	-2.821	0.637	0.699

Appendix Table 10: REVIGO's summarisation of 15 upregulated cellular component ontologies in the three virulent strains. Eleven cluster representatives are shown in black letters and their cluster members are listed in gray italics and indented. The 15 terms could be summarised into 11 clusters and nine of which have only one singleton.

term_ID	description	frequency	plot_X	plot_Y	plot_size	log10_FDR	uniqueness	dispensability
GO:0000502	proteasome complex	0.28%	-4.466	-4.87	4.822	-2.8729	0.552	0
GO:0032991	macromolecular complex	14.46%	1.904	5.059	6.531	-2.2815	0.882	0
GO:0043226	organelle	16.72%	-3.272	3.477	6.594	-1.4067	0.885	0
GO:0005694	chromosome	0.97%	5.286	-3.642	5.359	-1.5482	0.488	0.089
GO:0044424	intracellular part	43.66%	1.13	-7.539	7.011	-2.0079	0.693	0.153
GO:0005839	proteasome core complex	0.14%	-5.008	-3.47	4.506	-2.426	0.426	0.367
└ GO:0005838	*proteasome regulatory particle*	*0.01%*	*null*	*null*	*3.354*	*-1.5214*	*0.48*	*0.739*
└ GO:0022624	*proteasome accessory complex*	*0.02%*	*null*	*null*	*3.738*	*-1.5214*	*0.463*	*0.782*
└ GO:0019773	*proteasome core complex, alpha-subunit complex*	*0.03%*	*null*	*null*	*3.881*	*-1.9172*	*0.456*	*0.8*
GO:0030529	ribonucleoprotein complex	6.09%	-2.299	-5.4	6.155	-2.426	0.514	0.382
GO:0043228	non-membrane-bounded organelle	8.44%	6.254	-2.758	6.297	-2.4584	0.562	0.419
GO:0015629	actin cytoskeleton	0.15%	5.613	-1.608	4.549	-1.3556	0.545	0.426
GO:0043232	intracellular non-membrane-bounded organelle	7.68%	4.264	-3.9	6.256	-2.4584	0.41	0.629
└ GO:0005840	*ribosome*	*5.76%*	*null*	*null*	*6.131*	*-1.9172*	*0.285*	*0.835*
GO:0043229	intracellular organelle	15.79%	3.57	-4.381	6.569	-1.4067	0.444	0.69

Appendix Table 11: REVIGO's summarisation of 12 upregulated molecular function ontologies in the three virulent strains. Eleven cluster representatives are shown in black letters and their cluster members are listed in gray italics and indented. Total 12 terms could be summarised into 11 clusters and only one of which contains two members (GO:0042623 and GO:0016887).

term_ID	description	frequency	plot_X	plot_Y	plot_size	log10_FDR	uniqueness	dispensability
GO:0003735	structural constituent of ribosome	2.61%	-5.09	-4.316	6.091	-1.8416	0.831	0
GO:0004298	threonine-type endopeptidase activity	0.07%	-5.922	3.54	4.506	-2.3019	0.677	0
GO:0005198	structural molecule activity	3.64%	-0.759	-5.894	6.236	-1.7773	0.832	0
GO:0005488	binding	55.59%	-1.481	8.137	7.42	-1.3098	0.921	0
GO:0005515	protein binding	2.48%	6.456	0.109	6.07	-2.3019	0.814	0
GO:0003950	NAD+ ADP-ribosyltransferase activity	0.01%	4.096	-4.359	3.721	-1.382	0.813	0.014
GO:0016874	ligase activity	3.87%	0.956	-0.94	6.262	-1.4237	0.811	0.021
GO:0003779	actin binding	0.08%	4.054	5.232	4.579	-1.6459	0.733	0.048
GO:0042623	ATPase activity, coupled	2.88%	-6.326	1.2	6.134	-1.382	0.637	0.22
└ GO:0016887	*ATPase activity*	*5.23%*	*null*	*null*	*6.394*	*-1.9208*	*0.636*	*0.758*
GO:0070003	threonine-type peptidase activity	0.07%	-5.194	3.2	4.506	-2.3019	0.677	0.465
GO:0008092	cytoskeletal protein binding	0.16%	3.518	5.713	4.885	-1.4101	0.733	0.532

Appendix Table 12: REVIGO's summarisation of 44 downregulated biological process ontologies in the three virulent strains. Twenty-three cluster representatives are shown in black letters and their cluster members are listed in gray italics and indented. Total 44 GO terms could be summarised into 23 clusters and 15 of which have only a single term.

term_ID	description	frequency	plot_X	plot_Y	plot_size	log10_FDR	uniqueness	dispensability
GO:0007154	cell communication	4.36%	-3.171	-7.036	6.336	-1.8601	0.838	0
GO:0023052	signaling	3.84%	5.337	1.818	6.281	-1.8153	0.956	0
GO:0048583	regulation of response to stimulus	0.69%	-7.106	1.529	5.535	-2.4365	0.558	0
GO:0065007	biological regulation	14.92%	5.057	-1.051	6.87	-2.4365	0.961	0
GO:0006468	protein phosphorylation	1.21%	1.655	3.465	5.778	-1.6635	0.669	0.037
└ GO:0036211	*protein modification process*	*2.90%*	*null*	*null*	*6.158*	*-1.3354*	*0.815*	*0.71*
└ GO:0006464	*cellular protein modification process*	*2.90%*	*null*	*null*	*6.158*	*-1.3354*	*0.769*	*0.864*
GO:0006793	phosphorus metabolic process	16.89%	3.825	5.903	6.924	-2.4157	0.861	0.092
GO:0006399	tRNA metabolic process	2.53%	1.308	-4.521	6.1	-1.5214	0.723	0.221
└ GO:0034470	*ncRNA processing*	*2.26%*	*null*	*null*	*6.05*	*-1.4437*	*0.724*	*0.892*
└ GO:0008033	*tRNA processing*	*1.59%*	*null*	*null*	*5.897*	*-1.5214*	*0.731*	*0.851*
GO:0043412	macromolecule modification	5.09%	2.843	-3.917	6.404	-1.3279	0.863	0.225
GO:0009118	regulation of nucleoside metabolic process	0.17%	-4.23	-0.202	4.935	-1.5482	0.541	0.292
GO:0023051	regulation of signaling	0.38%	-7.033	0.421	5.275	-2.4365	0.496	0.313
└ GO:0044700	*single organism signaling*	*3.84%*	*null*	*null*	*6.281*	*-1.8153*	*0.621*	*0.96*
└ GO:0035556	*intracellular signal transduction*	*2.72%*	*null*	*null*	*6.131*	*-1.8601*	*0.389*	*0.912*
└ GO:0007165	*signal transduction*	*3.80%*	*null*	*null*	*6.277*	*-1.8153*	*0.373*	*0.709*
GO:0010646	regulation of cell communication	0.38%	-5.792	-0.486	5.276	-2.4365	0.51	0.313
GO:0065009	regulation of molecular function	0.84%	-5.579	1.169	5.619	-1.4921	0.595	0.332

Appendix Table 12: REVIGO's summarisation of 44 downregulated biological process ontologies in the three virulent strains. **(Continued)**

term_ID	description	frequency	plot_X	plot_Y	plot_size	log10_FDR	uniqueness	dispensability
GO:0019220	regulation of phosphate metabolic process	0.39%	-3.916	3.572	5.289	-1.5214	0.379	0.358
└ GO:0006140	*regulation of nucleotide metabolic process*	*0.18%*	*null*	*null*	*4.955*	*-1.5482*	*0.387*	*0.931*
GO:0009894	regulation of catabolic process	0.28%	-6.266	3.216	5.138	-1.5214	0.379	0.359
GO:0031329	regulation of cellular catabolic process	0.26%	-5.515	2.447	5.116	-1.5214	0.443	0.371
└ GO:0032318	*regulation of Ras GTPase activity*	*0.08%*	*null*	*null*	*4.619*	*-1.5214*	*0.221*	*0.965*
└ GO:0030811	*regulation of nucleotide catabolic process*	*0.17%*	*null*	*null*	*4.933*	*-1.5482*	*0.351*	*0.994*
└ GO:0033124	*regulation of GTP catabolic process*	*0.13%*	*null*	*null*	*4.822*	*-1.5482*	*0.359*	*0.96*
└ GO:0033121	*regulation of purine nucleotide catabolic process*	*0.17%*	*null*	*null*	*4.933*	*-1.5482*	*0.35*	*0.979*
└ GO:0032313	*regulation of Rab GTPase activity*	*0.03%*	*null*	*null*	*4.169*	*-1.4828*	*0.254*	*0.835*
└ GO:1900542	*regulation of purine nucleotide metabolic process*	*0.18%*	*null*	*null*	*4.951*	*-1.5482*	*0.367*	*0.994*
└ GO:0043087	*regulation of GTPase activity*	*0.13%*	*null*	*null*	*4.822*	*-1.5482*	*0.332*	*0.935*
└ GO:0032483	*regulation of Rab protein signal transduction*	*0.03%*	*null*	*null*	*4.169*	*-1.4828*	*0.456*	*0.999*
└ GO:0046578	*regulation of Ras protein signal transduction*	*0.10%*	*null*	*null*	*4.682*	*-1.5482*	*0.426*	*0.982*
GO:0051174	regulation of phosphorus metabolic process	0.39%	-3.686	3.133	5.29	-1.5214	0.517	0.384
GO:0050790	regulation of catalytic activity	0.65%	-6.448	2.528	5.51	-1.5214	0.497	0.388
└ GO:0051336	*regulation of hydrolase activity*	*0.30%*	*null*	*null*	*5.173*	*-1.5482*	*0.518*	*0.875*
GO:0006796	phosphate-containing compound metabolic process	16.69%	0.67	6.51	6.919	-2.4157	0.704	0.476
GO:0050794	regulation of cellular process	13.66%	-6.142	1.526	6.832	-2.4365	0.465	0.478
└ GO:0050789	*regulation of biological process*	*14.47%*	*null*	*null*	*6.857*	*-2.4365*	*0.488*	*0.825*
GO:0034660	ncRNA metabolic process	3.21%	1.636	-3.836	6.203	-1.4828	0.752	0.482
GO:1902531	regulation of intracellular signal transduction	0.28%	-6.35	-0.03	5.144	-2.4365	0.408	0.547
└ GO:0009966	*regulation of signal transduction*	*0.36%*	*null*	*null*	*5.257*	*-2.4365*	*0.401*	*0.966*
└ GO:0051056	*regulation of small GTPase mediated signal transduction*	*0.11%*	*null*	*null*	*4.746*	*-2.4365*	*0.432*	*0.884*
GO:0007264	small GTPase mediated signal transduction	0.23%	-6.595	-0.316	5.067	-1.6635	0.476	0.588
GO:0016310	phosphorylation	6.30%	0.164	7.076	6.496	-1.5544	0.725	0.649

Appendix Table 13: REVIGO's summarisation of 24 downregulated molecular function ontologies in the three virulent strains. Sixteen cluster representatives are shown in black letters and their cluster members are listed in gray italics and indented. Total 24 terms could be summarised into 16 clusters and 13 of which have only a single term.

term_ID	description	frequency	plot_X	plot_Y	plot_size	log10_FDR	uniqueness	dispensability
GO:0005085	guanyl-nucleotide exchange factor activity	0.06%	1.25	1.547	4.462	-1.5884	0.915	0
GO:0008047	enzyme activator activity	0.12%	2.87	5.583	4.736	-1.5884	0.572	0
└ GO:0005083	*small GTPase regulator activity*	*0.06%*	*null*	*null*	*4.469*	*-1.433*	*0.532*	*0.964*
└ GO:0030695	*GTPase regulator activity*	*0.09%*	*null*	*null*	*4.62*	*-1.5331*	*0.541*	*0.9*
└ GO:0005099	*Ras GTPase activator activity*	*0.04%*	*null*	*null*	*4.239*	*-1.433*	*0.539*	*0.93*
└ GO:0005097	*Rab GTPase activator activity*	*0.03%*	*null*	*null*	*4.144*	*-1.433*	*0.543*	*0.917*
└ GO:0005096	*GTPase activator activity*	*0.08%*	*null*	*null*	*4.547*	*-1.5331*	*0.529*	*0.774*
└ GO:0060589	*nucleoside-triphosphatase regulator activity*	*0.16%*	*null*	*null*	*4.887*	*-1.585*	*0.566*	*0.82*
GO:0008270	zinc ion binding	3.46%	-3.964	-5.458	6.213	-1.5544	0.791	0
GO:0016773	phosphotransferase activity, alcohol group as acceptor	4.04%	3.944	-5.083	6.281	-1.5884	0.743	0
└ GO:0016301	*kinase activity*	*5.08%*	*null*	*null*	*6.381*	*1.5884*	*0.74*	*0.701*
GO:0030234	enzyme regulator activity	0.44%	-5.996	3.345	5.315	-1.5884	0.915	0
GO:0016652	oxidoreductase activity, acting on NAD(P)H, NAD(P) as acceptor	0.09%	-2.169	5.3	4.627	-1.3215	0.811	0.022
GO:0016740	transferase activity	22.12%	5.493	3.294	7.02	-1.5884	0.909	0.048
GO:0030246	carbohydrate binding	0.83%	6.975	0.34	5.591	-1.433	0.884	0.064
GO:0043167	ion binding	33.31%	-6.112	-0.93	7.197	-1.5884	0.885	0.117
GO:0043169	cation binding	15.81%	-4.986	-3.428	6.874	-1.433	0.825	0.281
GO:0004849	uridine kinase activity	0.03%	3.122	-6.216	4.134	-1.3215	0.792	0.382
GO:0016772	transferase activity, transferring phosphorus-containing groups	9.19%	4.657	-4.406	6.638	-1.5884	0.796	0.42
GO:0008746	NAD(P)+ transhydrogenase activity	0.06%	-1.685	5.493	4.433	-1.433	0.785	0.586
└ GO:0008750	*NAD(P)+ transhydrogenase (AB-specific) activity*	*0.04%*	*null*	*null*	*4.245*	*-1.433*	*0.787*	*0.913*

Appendix Table 13: REVIGO's summarisation of 24 downregulated molecular function ontologies in the three virulent strains. **(Continued)**

term_ID	description	frequency	plot_X	plot_Y	plot_size	log10_FDR	uniqueness	dispensability
GO:0046872	metal ion binding	15.49%	-4.546	-4.326	6.865	-1.433	0.773	0.594
GO:0004672	protein kinase activity	1.88%	3.805	-5.594	5.948	-1.5884	0.745	0.603
GO:0046914	transition metal ion binding	7.34%	-4.3	-4.815	6.54	-1.5287	0.78	0.694