Genome-wide identification of long non coding RNAs in *Plasmodium falciparum*

A Major Project dissertation submitted

in partial fulfilment of the requirement for the degree of

Master of Technology
in
Bioinformatics

Submitted by

Vidhi Malik

(2K11/BIO/21)

Delhi Technological University, Delhi, India

Under the supervision of

Dr. YashaHasija

Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)
ShahbadDaulatpur, Main Bawana Road,
Delhi-110042, INDIA

CERTIFICATE

This is to certify that the M. Tech. dissertation entitled **"Genome-wide identification of long non coding RNAs in *Plasmodium falciparum*"**, submitted by **Vidhi Malik (2K11/BIO/21)** in partial fulfillment of the requirement for the award of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate's own work carried out by her under my guidance.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.

Date:

Dr. YashaHasija
(Project Mentor)
Assistant Professor,
Department of Biotechnology,
Delhi Technological University.
(Formerly Delhi College of Engineering, University of Delhi).

ACKNOWLEDGEMENT

Successful completion of any work would be incomplete unless we mention the name of persons who made it possible. Guidance and encouragement served as a beacon of light and crowned my efforts into success.

I would like to take this opportunity to express my deep sense of gratitude to my respected Vice Chancellor Prof. P.B.Sharma for constantly motivating me. I owe my profound gratitude to Dr. YashaHasija, Assistant Professor and Associate Head, Department of Biotechnology, DTU, for her involvement, skilful assistance and guidance during the tenure of this project. She has always been a moral support during my research experience.

It gives me immense pleasure to express my sincere gratitude and respect to my worthy guide, Dr. VinodScaria (Scientist, IGIB, Delhi) for introducing me to the enthralling field of "RNA-Seq". I proclaim my indebtedness to him for the dexterous guidance, untiring efforts, constructive criticism and providing continuous enthusiasm during the entire tenure of my research work. Being always empowered by his graceful and friendly behavior, I enjoyed every moment of my research period. It was a delight to work under his kind supervision.

It is high time to acknowledge Mrs.DeekshaBhartia for her assistance and suggestions related to my research and emotional support during my project.

Vidhi Malik

2K11/BIO/21

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ncRNA	Non coding RNA
lncRNA	Long non coding RNA
ACT	Artemisinin Combination Therapy
G6PD	Glucose 6- Phosphate dehydrogenase
TARE	Telomere associated repetitive elements
RDT	Rapid Diagnostic Test
NIH	National Institute of Health
RNA	Ribonucleic Acid
dNTP	Deoxyribonucleotides
FPKM	Fragments Per Kilobase of transcript per Million fragments mapped

Genome-wide identification of long non coding RNAs in *Plasmodium falciparum*

Vidhi Malik
Delhi Technological University, Delhi, India

1. ABSTRACT

Plasmodium falciparum, causative agent of malaria, is endangering life of millions of people annually. This parasite is highly capable of evading human immune system and developed resistance to many antimalarial drugs. Genome analysis of this parasite has shown signs of involvement of non-coding RNAs in integral biology of parasite aiding in their survival, virulence and resistance development. This has encouraged us to identify the genome-wide repertoire of non-coding RNA especially long non-coding RNA (lncRNA) in this parasite. We have identified 426 lncRNAs in *P. falciparum* using *in-silico* RNA-sequencing analysis tools and *in-house* perl scripts. The identified lncRNAs were annotated to depict their differential expression in different life-cycle stages of *P. falciparum* and the lncRNAs were found to have a stage-dependent expression. Analysis has revealed 5, 2 and 8 unique lncRNAs in late trophozoite, schizont and gametocyte V stage respectively. This study aims to identify and annotate considerable amount of genome-wide lncRNAs in *P. falciparum* and indicates significant roles of these lncRNAs in different life-cycle stages in the human host. The dataset can be used to identify and annotate other lncRNAs and understand the functions of these lncRNAs in the pathogenecity of the parasite.

2. INTRODUCTION

Malaria is a disease caused by *Plasmodium* and transmitted to humans with a bite of *Plasmodium* parasitized female *Anopheles* mosquito. It is a disease of concern due to lack of effective vaccination for its prevention. *P.falciparum,* the most virulent species of genus *Plasmodium* has a very complex life cycle involving two hostsin order to multiply and spread itself in a particular area(Epp *et al.,* 2009). Issue of concern is its ability to evade host's immune system(Florens *et al.,* 2002). It is believed that non coding RNAs (i.e. RNA molecules which do not code for proteins)are involved in providing this parasite benefit of survival. And whole genome studies of *Plasmodium falciparum*also ensure this due to presence of only basal transcription factors, some non coding RNA (ncRNA) and RNA binding protein and lack of various gene regulatory protein, RNA interference and DNA methylation machinery(Scherf *et al.,* 2008).

The expression of highly variant immunodominant erythrocyte surface molecules, PfEMP1 (*Plasmodium falciparum* erythrocyte membrane protein) on erythrocytes surface having domain for cytoadherence, is found to be the reason of capability of parasite to escape host immunity. PfEMP1 is encoded by a member of multigene family, *var* that contain 60 genes but regulated to express in monoallelic manner. Switching between *var* genes for monoallelic expression leads to expression of variant molecule on surface of erythrocytes (Epp *et al.,*2009;Scherf*et al.,*2008).*var*gene also encode two long non coding RNA (lncRNA) which are supposed to be involved in regulation of their monoallelic expression.In addition to *var*gene derived ncRNAs, subtelomeric regions also encode various TARE (telomere associated repetitive elements) ncRNA. These repeats interact with each other to form long and multiple stem loop structure which can bind to histones and bring them at *var* gene loci thereby regulating their expression by assembly or disassembly of heterochromatin at telomere (Sierra-Miranda, M*et al.,* 2012).These lncRNA can create epigenetic memory marks before cell division in order to make epigenetic regulatory information inheritable (Broadbent *et al.,* 2011).

Advent of high throughput sequencing technology has provided a way to evaluate whole transcriptome of an organism by generating short sequence reads termed as RNA-Seq reads.RNA-Seq have a wide application in gene discovery, annotation of coding and as well as non coding genes, quantification of expression of transcripts with the advantage of high speed and low cost over conventional EST sequencing and microarray technology (Roberts *et al.,* 2011). RNA-Seq pipeline is used here to identify ncRNAin *P. falciparum 3D7* and check their expression patternin four stages of parasite's life cycle, namely, late trophozoite, schizont stage, gametocyte II and gametocyte V stage.

Whole pipeline can be categorized into following steps–

1. Mapping of reads to reference genome using TopHat.

2. Transcriptome reconstruction using Cufflinks.
3. Extract non coding transcripts using Coding Potential Calculator and getorf software of EMBOSS suite.
4. Estimating differential expression of noncoding transcripts in different samplesusing cuffdiff and DESeq.

3. REVIEW OF LITERATURE

3.1 *Plasmodium falciparum* 3D7 and Malaria

*Plasmodium falciparum*is the most virulent species that causes malaria and is responsible for killing around 2.7 million people each year. The parasite's ability to develop resistance to available drugs, lack of effective vaccine and its ability to evade host immune response are the serious issues of concern. Therefore, this parasite is attracting attention of many researchers and pharmaceutical companies. There are other species of *Plasmodium* also that can infect humans but *P. falciparum* account for almost 80% of cases.*Plasmodium falciparum* can carry out its lifecycle at temperature above 20ºC.

3.1.1 Life Cycle of *Plasmodium falciparum*

Plasmodium falciparum has a very complicated life cycle and utilizes two hosts to complete its life cycle.With the bite of infected mosquito sporozoites from mosquito's saliva are also released into human subcutaneous tissues from which they passed through blood capillaries to liver cells. In liver cells these sporozoites are transformed into rounded form and start dividing within membrane bound vacuole in order to form many merozoites which are released into the blood by rupture of membrane. These merozoites can either invade red blood cells to start erythrocytic replicative cycle or may differentiate into male and female gametophytes. Within twelve hours after invasion in the erythrocytes the merozoite grow first to a ring-shaped form and then after twenty four hours develop into trophozoite form in which it allow the parasite to change the surface of erythrocytes in order to mediate cytoadherence and transportation of molecules in and out of the cell. Within thirty six hours it enters into late trophozoite stage where it enlarge further followed by division of parasite to produce multiple merozoites within a cell,this stage is termed as schizont. After forty eight hours red blood cell burst and merozoites are released in blood where they infect other cells and cycle goes on (Florens *et al.,* 2002). Clinical symptoms are usually observed in asexual erythrocytic stage of parasite (Epp *et al.,* 2009). After seven to ten days gametocytes form are visible in blood. Gametocytes are classified into five stages based on their morphology. Gametocyte I have round like trophozoite morphology. Gametocyte II has enlarged D form morphology. Gametocyte III has distorted morphology but male and female gametocytes can be distinguished by stain at this stage. Male gametocytes have usually large and lobulated nucleus and less ribosomes, endoplasmic reticulum and mitochondria as compared to female gametocytes. Gametocytes IV have elongated morphology and gametocyte V has banana shaped morphology. When another mosquito bite infected person for blood meal then along with blood gametocytes are also ingested by mosquito. In the cold environment of mosquito gut gametocytes converted into gametes and mate to form zygote which then form ookinete which traverse through the gut wall and develop into oocyst which ultimately divide to give rise to sporozoites again (Florens *et al.,* 2002). The cyst burst to release sprozoites, followed by their

migration into salivary glands. This mosquito releases these sporozoites into human body when attempting to have blood meal. In this way cycle goes on and the two hosts keep on infecting each other. The development of sporozoites in mosquito will take around two weeks after that mosquito is ready to infect humans.

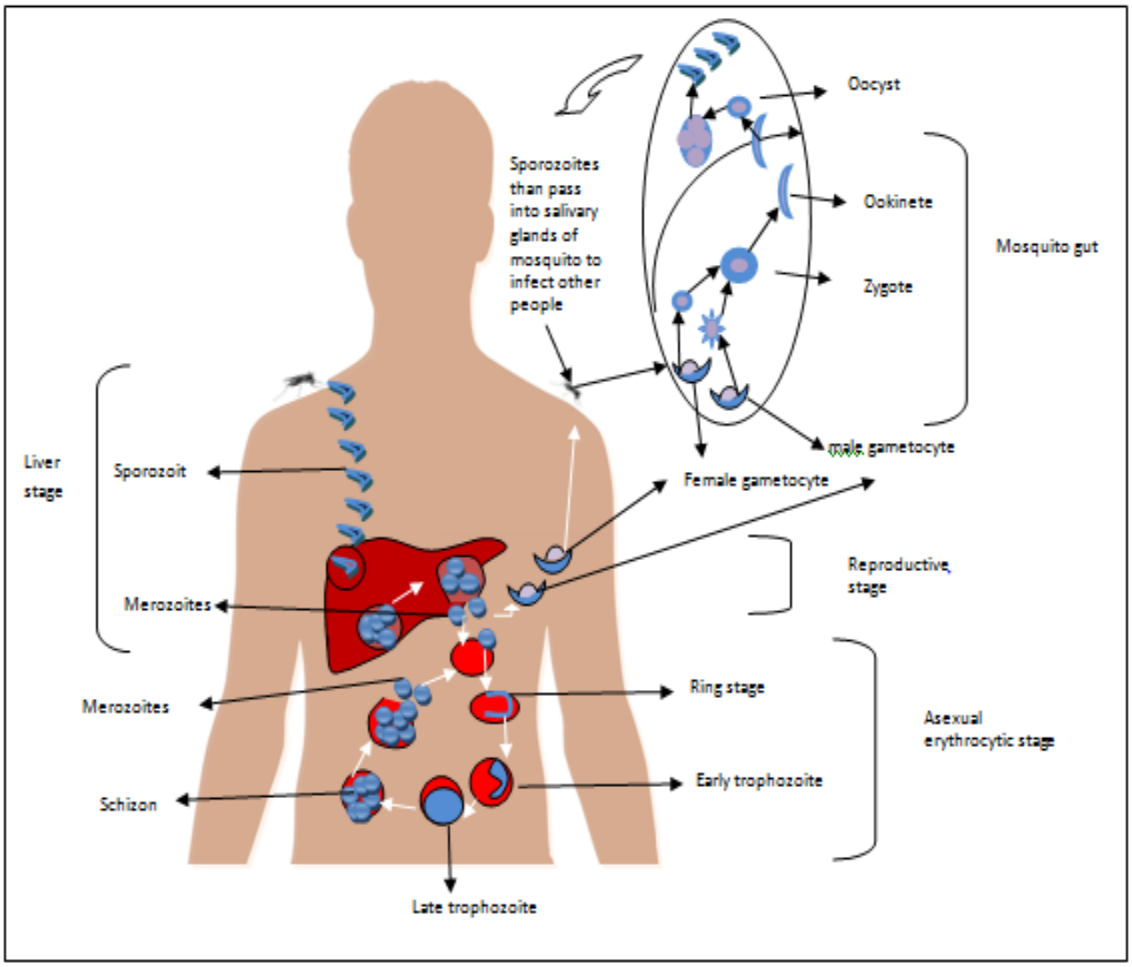


Figure 1: *Plasmodium falciparum* life cycle

3.1.2 Symptoms of Malaria

Symptoms of malaria usually begin to appear after 10-15 days of bitten by infected *anopheles* mosquito. General symptoms of malaria in case of uncomplicated case may include fever, diarrhea, vomiting, muscular pain, headache, weakness and chills. Malaria involves death of lots of red blood cells which may lead to jaundice and anemia. Diagnosis and treatment of this disease should be done as soon as possible, otherwise may result into severe effects. If left untreated for long time then may result into organ failure due to inadequate supply of blood to organs. Coma, kidney failure, difficulty in breathing, decrease in blood platelets, seizure, abnormal behavior, neurological abnormalities and even death can be severe effects of infection

by *P. falciparum* if left untreated. Diagnosis of infection by *P. falciparum* can also be done on basis of detection of elevation in level of bilirubin and aminotransferases, presence of albumin and other abnormal bodies in urine (Clark *et al.,* 2006).
If a woman is infected by this parasite during pregnancy then there is a possibility that infant also develop disease and may lead to paralysis, cerebral malaria, trouble in movement of muscles and even deafness in child. And also pregnant lady can have premature delivery of baby (Clark *et al.,* 2004).

3.1.3 Diagnosis

Malaria can be diagnosed by two ways –

3.1.3.1 Microscopy

Microscopic examination of blood sample of patient is generally a standard method for detection of malaria. This method provides higher accuracy and is able to distinguish between different strains of *Plasmodium* along with their different stages of life cycle. Also it can detect parasite even at low density in blood sample (Ndyomugyenyi *et al*.,2007).

3.1.3.2 Rapid Diagnostic Test (RDTs)

RDTs are like a dipstick method which provides results within 20 minutes. It is based on detection of antigens of parasite in blood of patient. A strip of nitrocellulose is designed in which labeled antibodies against parasite's antigen is immobilized along with antibody against antigen-antibody complex so that it can capture this complex in a thin test line (Figure 2). One control line is also there at the other end of strip in order to test reliability of antibody-dye conjugate. First of all patients blood sample is mixed with lysing buffer on nitrocellulose strips which ruptures red blood cells so that parasite's antigens can bind the labeled antibody. The antigen-labeled antibody complex and other components of blood travel along nitrocellulose strip by capillary action through fiber-mesh and also by flushing action of buffer applied behind the sample. Antigen-labelled antibody complex will capture by antibodies immobilized at the test line and form a thin visible line. Presence of this visible line shows that antigen is present in blood sample otherwise we will see at the control band if labeled antibodies accumulate there that shows that the complex has travelled whole strip and also antibodies are correctly labeled. Greater concentration of antigen in blood sample will give higher test line intensity and lower control band intensity because most of the antibodies will form conjugate with antigen and bind at test line (Ndyomugyenyi *et al*.,2007; Tarimo*et al.,* 2001).

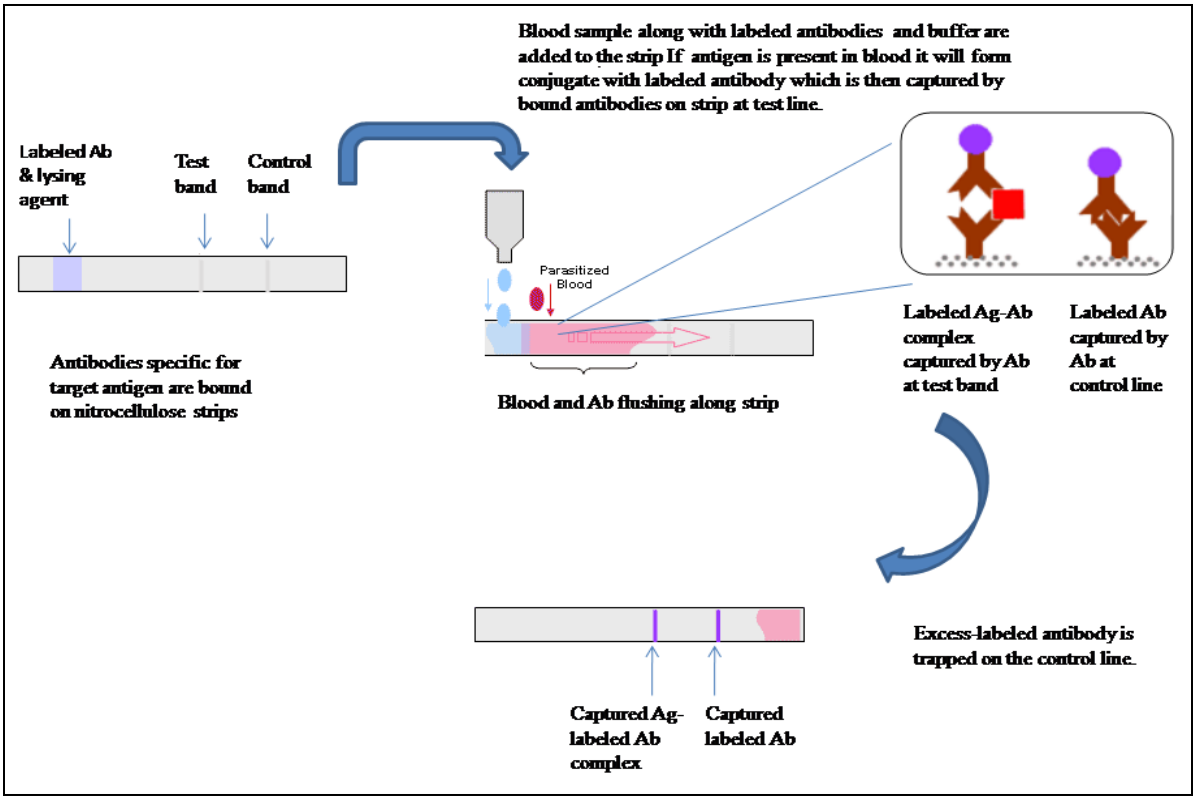


Figure 2: Conceptual design of RDT

3.1.4 Treatment

Before administering medication to malaria patients it is important to check whether a person have any history of travelling to areas where *Plasmodium* species have developed resistance to particular medication. Medication should be provided on the basis of area in which person is living. Because *P. falciparum* and *P. vivax* have developed resistant to many antimalarial drugs for example, chloroquine resistant strains are spread in many endemic areas (Davy *et al., 2010*). Drugs usually recommended by national malaria control programmes are:-

- artemesinin-containing combination treatments (for example, artemether-lumefantrine, artesunate-amodiaquine)
- atovaquone-proguanil
- chloroquine
- doxycycline
- mefloquine
- quinine
- sulfadoxine-pyrimethamine.

3.1.4.1 Treatment of *P. vivax* cases

Chloroquine can be used in treatment of *P. vivax.* Three days dose of 25 mg/kg can be administered in case of patient having all conditions indicating him chloroquine sensitive.But there is a problem of relapse of infection associated with *P. vivax* malaria.Presense of hypnozoites in liver is responsible for this. Primaquine is the drug of choice in case of *vivax* malaria because of their activity against dormant liver forms i.e. hypnozoites. Primaquine should not be administered to infants, pregnant women and to patients of glucose 6 phosphate dehydrogenase deficiency (G6PD) (Tarimo *et al*.,2001).

3.1.4.2 Treatment of *P. falciparum* cases

The treatment of *P. falciparum* malaria is based on areas identified as chloroquine resistant/ sensitive.Artemisinin Combination Therapy (ACT) should be given inresistant areas whereas chloroquine can be used in sensitive areas. ACT is a combination of artemisinin derivative and antimalarial drugs like amodiaquine, lumefantrine, mefloquine or sulfadoxine-pyrimethamine which have quite long lasting effects. Artemisinin derivatives are rapidly acting drugs and should not be prescribed as monotherapy unless the case is complicated because they can lead to development of resistance by parasite. Also before administering ACT to patients it should be confirmed by microscopy or RDT that infection is caused by *P. falciparum* strain (Tarimo*et al*.,2001).

ACT cannot be administered to pregnant women in their first trimester of pregnancy; quinine is the choice of drug in that case. Chloroquine is recommended for patients showing symptoms of malaria but RDT results for *P. falciparum* strains are negative until microscopy test results arrives(Ndyomugyenyi *et al*.,2007).

3.1.4.3 Prevention

Few precaution helps in reducing population of these parasites in environment. Following are some of them (Tarimo*et al*.,2001):-

- Use of Insecticide-treated bed nets while sleeping.
- Indoor spraying of mosquito killers.
- Application of chemical insecticides, insect growth regulators, toxins of *Bacillus thuringiensis* var. *israelensis* (Bti) and oil in water puddles to suppress growth of mosquito larvae into adult forms.
- For pregnant women living in malaria endemic regions, curative dose of antimalarial drugs twice after first trimester of pregnancy and iron and folate supplements in diet are recommended to prevent anemia.

3.2 Long ncRNAs and their role

3.2.1Discovery of LncRNAs

H19 was the first lncRNA gene discovered. It is immediately followed by discovery of X-inactive-specific transcript (XIST) lncRNA, which is involved in X- chromosome inactivation. But after the discovery of first miRNA i.e. lin-14 the focus of research shifts towards discovery of miRNA. During research it has been found that miRNA have a potential of regulating entire gene network and also they were allied with cancer.Approximately 7,000 to 23,000 lncRNA content is estimated in human genome, implying their potential role in gene networks that may be disrupted in metastatis (Gibb *et al*., 2011).

3.2.2 Origin of lncRNA

LncRNAs can be originated from intronic, exonic, intergenic, intragenic, promoter regions, 3′- and 5′- UTR, and enhancer sequences (Figure 3). Most of the lncRNAs are antisense to known protein coding genes known as natural antisense transcripts (NATs) (Nie *et al*., 2012).
NATs can be divided into two subtypes:

a. **Cis-NATs** are those sterile transcripts which are transcribed from opposite DNA strands at the same genomic loci.
b. **Trans-NATs** are the sterile transcripts which are transcribed from distal loci.

More recently, emerging experimental evidence revealed that NATs can also be generated from pseudogenes. Notably, many cancer relevant genes, particularly tumor suppressor genes, produce long antisense ncRNAs (Nie *et al*., 2012).

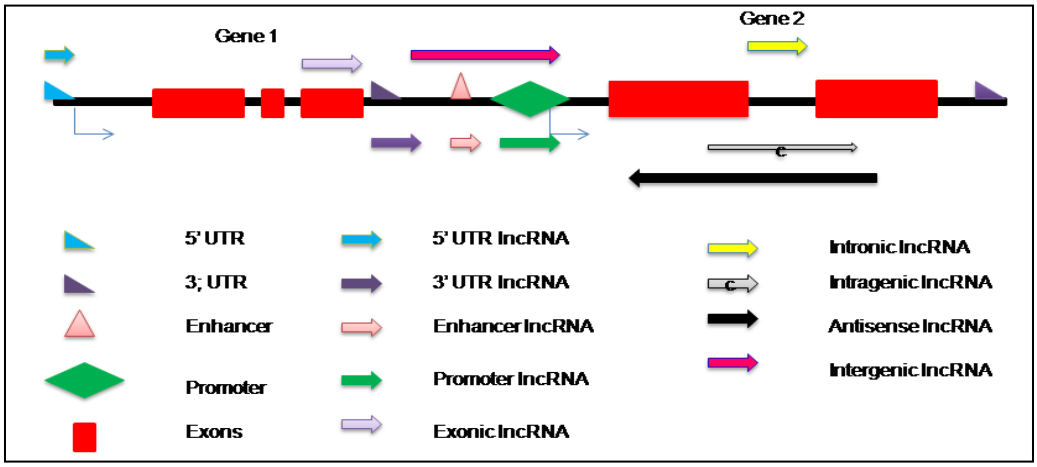


Figure 3:Origins of LncRNA
Arrows represent different types of LncRNA transcripts.

3.2.3 Types of lncRNAs

On the bases of their origin lncRNAs can be classified into six types:-

Class	Symbol	Characteristic	Disease / biological function associations
Long intergenic ncRNAs	lincRNAs	Have a length ranging from 1,000 – 10,000. Lie down at genomic loci between two genes; Regulate transcription of neighbouring genes	Involved in tumorigenesis and cancer metastasis or dosage compensation or imprinting.
Long intronic noncoding RNAs		As name suggest, located within the introns; evolutionary conserved; expression is tissue or subcellular types.	Role in post transcriptional gene silencing; aberrantly expressed in human cancers.
Telomere-associated ncRNAs	TERRAs	100 bp - >9 kb; conserved among eukaryotes; synthesized from C-rich strand; polyadenylated; form inter-molecular G-quadruplex structure with single-stranded telomeric DNA	possible impact on telomere-associated diseases including many cancers / negative regulation of telomere length and activity through inhibition of telomerase
Long non-coding RNAs with dual functions		both protein-coding and functionally regulatory RNA capacity	deregulation has been described in breast and ovarian tumors / modulate gene expressionthrough diverse mechanisms
Pseudogene RNAs		Gene copies that have lost the ability to code for a protein; Potential to regulate their proteincoding cousin; Made through retro-transposition; Tissue specific	Often deregulated during tumorigenesis and cancer progression / regulation of tumor suppressors and oncogenes by acting as microRNA decoys
Transcribed ultraconserved regions	T-UCRs	Longer than 200 bp; Absolutely conserved between orthologous regions of human, rat, and mouse; Located in both intra- and intergenic regions	Expression is often altered in some cancers; possible involvement in tumorigenesis / antisense inhibitors for protein-coding genes or other ncRNAs

Table 1: Types of LncRNA

3.2.4 Mechanisms of lncRNAs action

Wang *et al*. described four different mechanisms of lncRNAs action (Figure 4):-

1) lncRNAs can function as signals and regulate gene expression.
2) lncRNAs can titrate transcription factors and other proteins away from chromatin or they can function as decoy for miRNA target sites.
3) lncRNAs can recruit chromatin modifying enzymes to target genes and therefore function as guides.
4) lncRNAs can bring together multiple proteins to form ribonucleoprotein complexes (Sana *et al*., 2012).

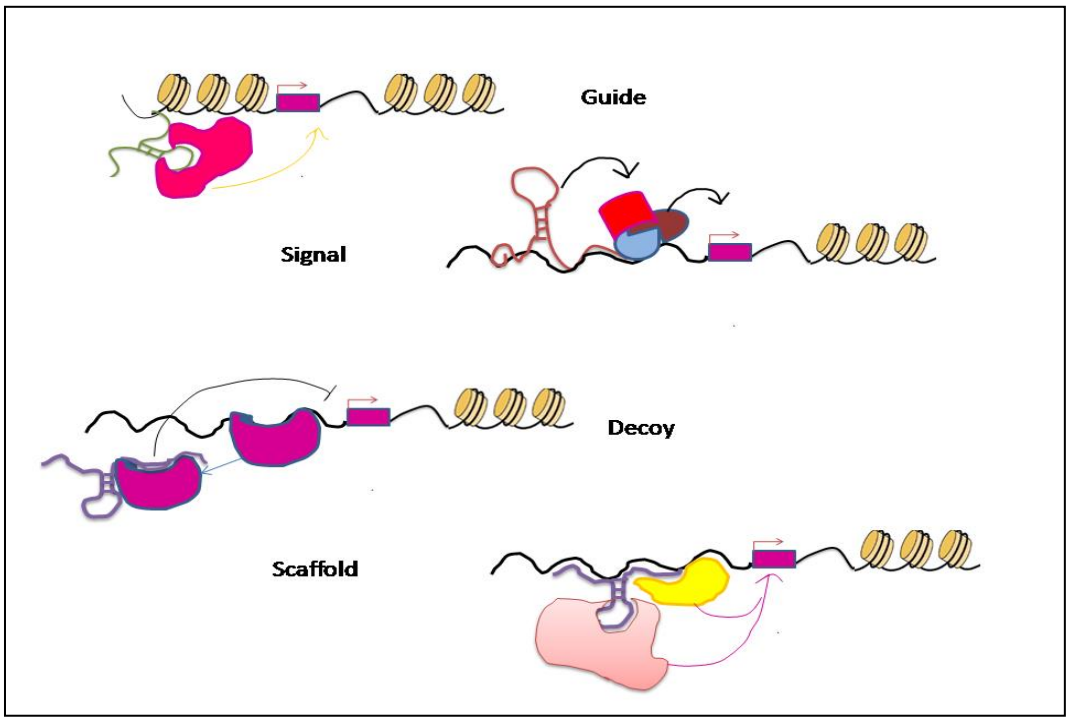


Figure 4: Schematic diagram of the four mechanisms of lncRNAs functioning

3.2.5 Diverse regulatory function of lncRNA

3.2.5.1 LncRNAs in chromatin remodeling

Chromatin remodeling includes changes in organization of chromatin responsible for change in gene expression without changing DNA sequence. Epigenetic modifications, includes histone and DNA methylation, histone acetylation and sumoylation, leads to remodeling of chromatin, which is required for regulation of gene expression. Presence of RNA binding domain in

chromatin remodeling complex was already well known but with the discovery of few lncRNAs and elucidation of their function indicates that these ncRNA are involved in chromatin remodeling. lncRNAs can act as scaffold for the recruitment of chromatin remodeling complex at particular genomic loci. Chromatin remodeling complex than either methylate or acetylate histones/DNA at that loci leads to change in chromatin state at that particular loci (Nie *et al*., 2012).For example, One of these ncRNAs, Hox transcript antisense RNA (*HOTAIR*), originates from the *HOXC* locus recruit Polycomb chromatin 19emodeling complex PRC2 via interacting with EZH2, followed by recruitment of other chromatin modifying complexes such as Mll, PcG, and G9a methyltransferase at HoxD locus, resulting in trimethylation of lysine 27 of histone H3 (Figure 5). This trimethylation change the chromatin state of HoxD locus into heterochromatin and thereby silences transcription across 40 kb of the *HOXD* locus (Sana *et al*., 2012). Similarly lcnRNA such as Xist/RepA, Tsix, ANRIL and Kcnqot1 can recruit Polycomb Repressive Complex (PRC) via direct interaction with EZH2 or other components to their targeted locus in order to silence expression of genes located in that particular region (Nie *et al*., 2012).

Another example of lncRNA involved in epigenetic regulation is Atf1.Expression of fbp1 gene is generally repressed by Tup protein by hindering RNA polymerase II processivity. But under the situation of glucose starvation Atf1 lncRNA bind to UAS1 element followed by binding of Rst2 to UAS2 element, leads to change in chromatin structure around fbp1 initiation site thereby changing accessibility to transcriptional machinery allowing expression of gene (Ponting *et al*., 2009).

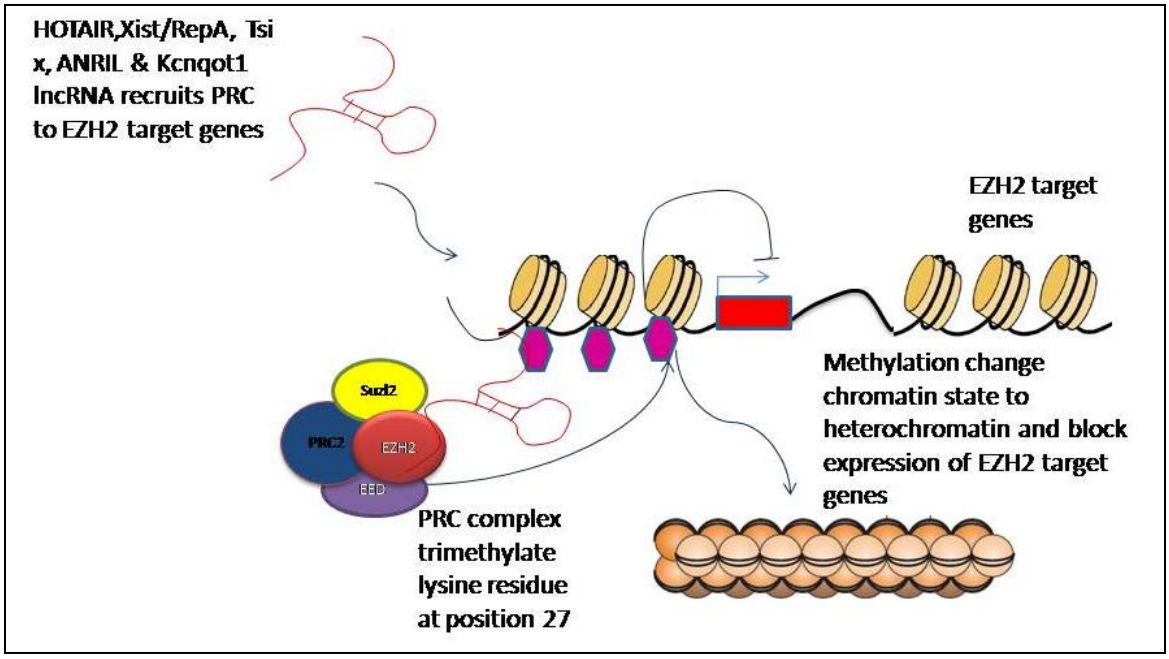


Figure 5: LncRNA in chromatin-remodeling

3.2.5.2 LncRNAs in gene regulation

Double stranded RNA molecule formed due to annealing with antisense sequence or repetitive regions within transcripts can lead to formation of endogenous-siRNAs (endo-siRNAs) molecule by Dicer 2 molecule. These endo-siRNAs play a role in suppressing the spread of mobile transposon elements in germline genome. Also have a role in silencing the expression of genes in case of its origin from antisense transcripts via RNA induced silencing complex (RISC) effector complexes (Chen*et al*., 2007).

3.2.5.3 LncRNAs in transcriptional regulation

Some lncRNAs are directly involved in transcriptional regulation. These lncRNAs can be classified into two categories:-

a. **Promoter-associated ncRNA (paRNAs)**: paRNAs can influence transcription in two ways-

   i. Activation of transcription

   An example of paRNA controlling transcriptional regulation involves an intergenic regionbetween members of the Dlx/dll homeodomain-containing protein family, Dlx-5/6. Feng *et al*. describedthat *Dlx-5/6* ultraconserved region is transcribed to generate an alternatively spliced form of *Evf-1*, the ncRNA *Evf-2*. *Evf-2* specifically cooperates with Dlx-2 to increase the transcriptional activity of the Dlx-5/6 enhancer in a targetand homeodomain-specific manner. A stable complex containing the *Evf-2* ncRNA and the Dlx-2 proteinforms in vivo, suggesting that the *Evf-2* ncRNA activates transcriptional activity by directly influencing Dlx-2activity (Figure 6)(Wang *et al*., 2008)**.**paRNAs do not regulate transcription at RNA level instead regulate transcription at the gene promoters in either *cis* manner in case of overlapping genes or in *trans* manner in case of distant promoter regions., rather than the RNA product of transcription, mediates regulation of the overlapping genes *in cis* or distant enhancer/promoter regions *in trans*(Feng *et al*., 2006).

   ii. Suppression of transcription

   In addition to activation of transcription paRNAs are also involved in suppression of transcription. Wang *et al.*,demonstrated that an RNA-binding protein, TLS, serves as a key transcriptional regulatory sensor of DNA damage signals that, based on its allosteric modulation by RNA, specifically binds to and inhibits CBP/p300 HAT activities on a repressed gene target, cyclin D1 (*CCND1*). Recruitment of TLS to the *CCND1* promoter

to cause gene-specific repression is directed by single stranded, low copy number ncRNA transcripts tethered to the 5′ regulatory regions of *CCND1* that are induced in response to DNA damage signals. Our data suggest that signal-induced ncRNAs localized to regulatory regions of transcription units can act cooperatively as selective ligands, recruiting and modulating the activities of distinct classes of RNA binding co-regulators in response to specific signals, providing an unexpected ncRNA/RNA-binding protein-based strategy to integrate transcriptional programs(Trapnell *et al*., 2009).

In addition, some paRNAs can also regulate transcription suppression by competing with RNA polymerase II for binding with transcription factor. As in the case of the dihydrofolate reductase (DHFR) gene in humans which contains two promoters, a minor and major promoter. In quiescent cells activity of major promotor is suppressed by lncRNA transcribed from minor promoter of DHFR (Figure 6). This lncRNA obstruct the formation of pre-initiation complex at major promoter both by forming triplex structure at major promoter and by binding to transcription factor TFIIB(Ponting *et al*., 2009).

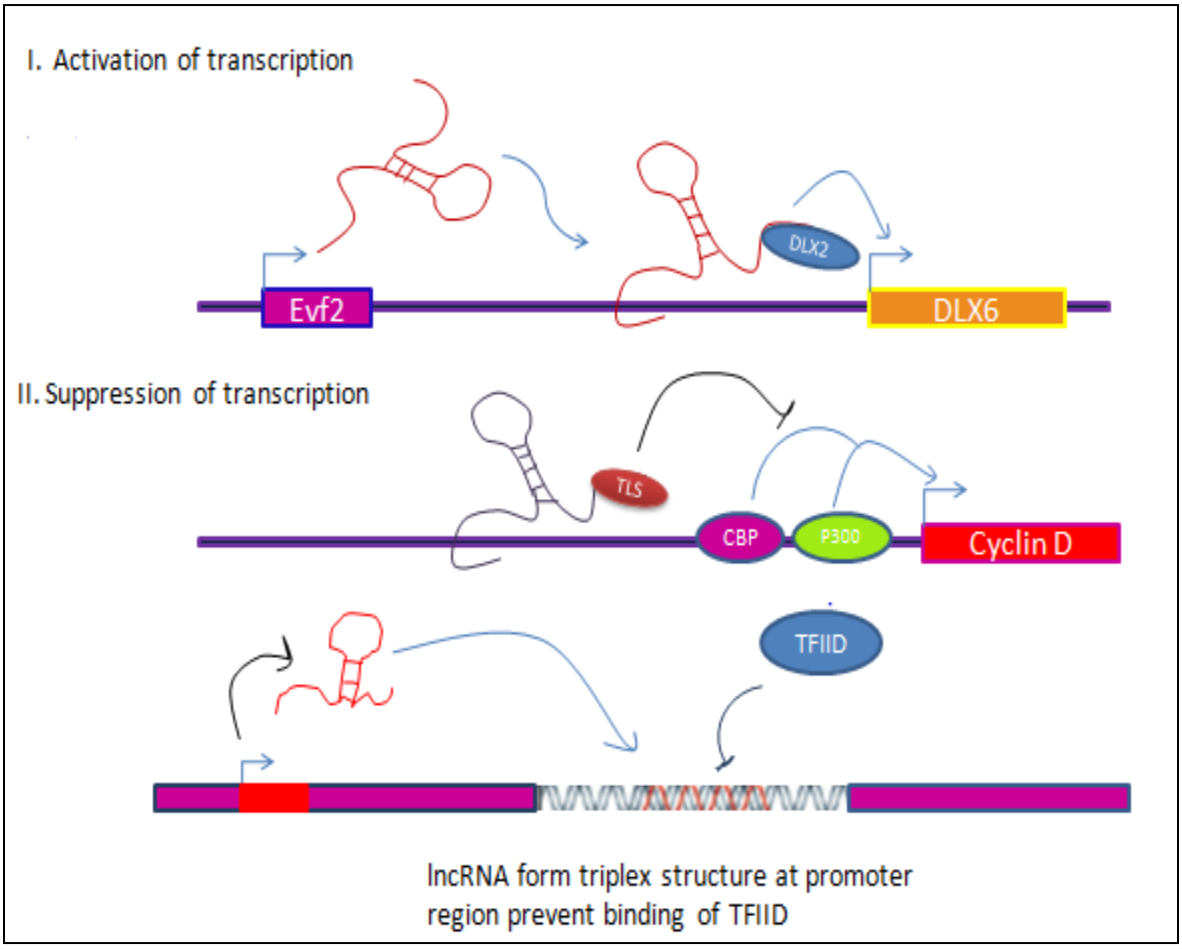


Figure 6: LncRNA in transcriptional regulation

b. Enhancer-like lncRNAs

Depletion of some lncRNA leads to decrease in expression their neighboring genes. This shows that these lncRNAs has a role in enhancing their neighboring gene expression. One of the example of enhancer-like lncRNA (e-RNAs) is ncRNA-a3, its depletion is correlated with decreased expression of its flanking genes i.e. Tal1/SCL (key regulator in hematopoiesis). Another example of e-RNA is ncRNA-a7; enhance expression of its neighboring Snai1 gene known for its function in cell adhesion, migration and epithelial-messenchymal transition (Nie *et al*., 2012).

3.2.5.4 LncRNAs in post transcriptional regulation

LncRNAs are also involved in posttranscriptional processing of mRNAs, including splicing, editing, trafficking, translation and degradation.

3.2.5.4.1 Incease in translational efficiency

Inspite of role of naturally antisense transcript lncRNAs (like Tsix, Air, HOTAIR and Evf-2) in epigenetic regulation, they also have their role in alternative splicing of paired genes by forming RNA duplexes that mask cis-regulatory elements in overlapping genes mRNAs. One of the examples of this type of lncRNA is Zeb2/Sip1 NAT which is involved in regulation of splicing of zinc finger Hox mRNA 'Zeb2' involved in epithelial-mesenchymal transition (EMT). Zeb2 NAT mask the 5'splice site of intron in 5'-UTR region of Zeb2 mRNA and prevent splicing of this intron by inhibiting binding of spliceosome the splice sites (Figure 7). This retained intron have internal ribosome entry site (IRES), which is recognized by translational machinery leading to more efficient translation of Zeb2 mRNA (Nie *et al*., 2012).

3.2.5.4.2 Stabilization of mRNA

Pseudogenes mRNA is also a kind of lncRNA that are involved in stabilization of their counterpart mRNA by protecting them against miRNA targeted for their destruction. Pseudogenes compete with miRNA for binding with their counterpart gene mRNA, therefore, they are also known as competing endogenous RNA (ceRNA). For example, PTENP1 is a pseudogene transcript, involved in increasing abundance of its counterpart mRNA PTEN. PTEN and PTENP1 have conserved 3' UTR regions, therefore PTENP1 compete with PTEN for binding with miRNA targeted for destruction of PTEN mRNA, thereby increasing expression and translation of PTEN mRNA.

Mutations in miRNA binding site in PTENP1 are found to be associated to cancer in some cases. Mutation leads to inhibition of binding of miRNA to PTENP1 and all miRNA binds to PTEN

protein in absence of competing PTENP1 action, leads to decrease in translation of PTEN which have a role in tumor suppression (Nie *et al*., 2012).

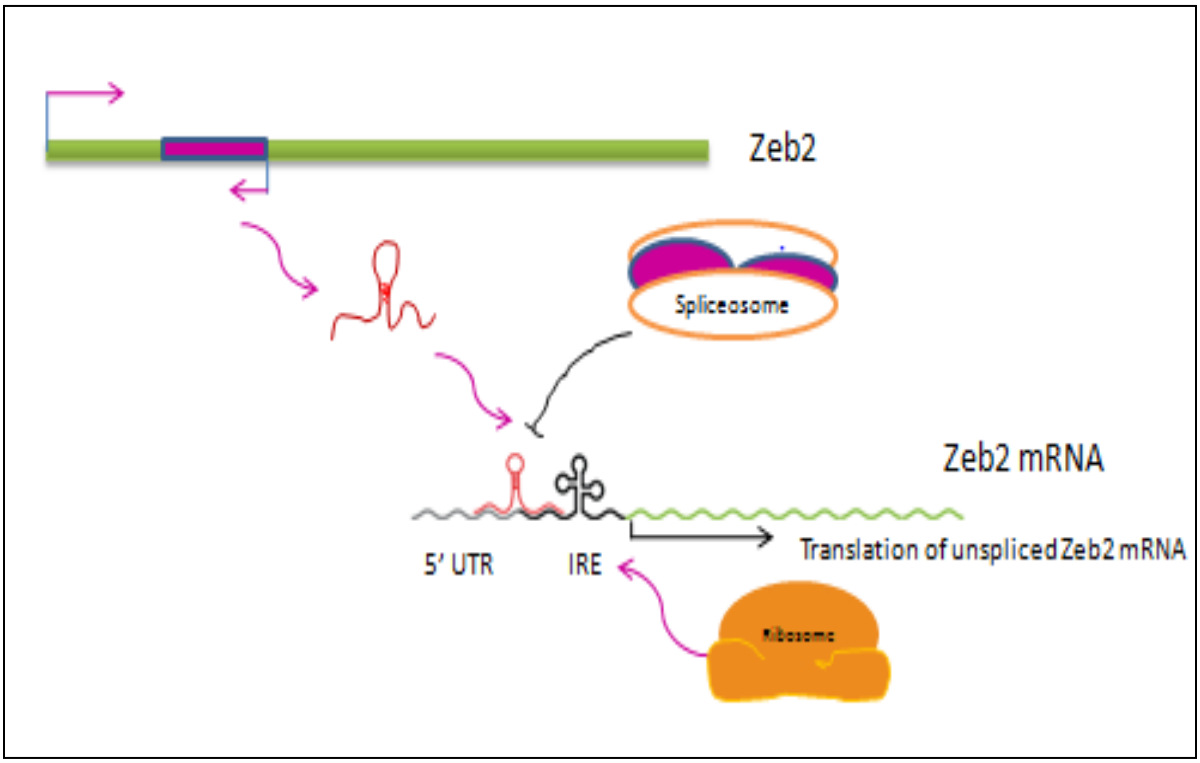


Figure 7: LncRNA in posttranscriptional regulation

3.2.5.4.3 Trafficking regulator

LncRNAs are also involved in controlling expression of genes by regulating subcellular localization of transcription factors. For example, ncRNA NRON (noncoding repressor of NFAT) is regulating activity of transcription factor NFAT(nuclear factor of activated T cells) by restricting it to localize in cytoplasm until calcium dependent signal rupture its binding to importin-beta 1 transporter protein so that NFAT can be imported into nucleus. NRON when bind to importin-beta 1 protein restricts NFAT entry into the nucleus, but do not restrict entry of other transcription factor. Calcium dependent signal aids in translocation of NFAT from cytoplasm to nucleus and thereby activating expression of its target genes(Mercer*et al., 2013*).

3.2.5.4.4 Long ncRNAs in splicing

Like the expression of Zeb2 mRNA antisense transcript regulate the translation of Zeb2 mRNA during mesenchymal development, the expression of antisense transcript Rev-ErbAα2 of ErbAα2 transcript (thyroid hormone receptor) regulate its alternative splicing to form two antagonistic isoforms.

3.2.5.5 LncRNAs in translation

NcRNA may also apply additional regulatory pressures during translation, a property particularly exploited in neurons where the dendritic or axonal translation of mRNA in response to synaptic activity contributes to changes in synaptic plasticity and the respective of neuronal networks. The RNAP III transcribed BC1 and BC200 ncRNAs, that previously derived from tRNAs, are expressed in the mouse and human central nervous system,respectively. BC1 expression is induced in response to synaptic activity and synaptogenesis and is specifically targeted to dendrites in neurons. Sequence complementarity between BC1 and regions of various neuron-specific mRNAs also suggest a role for BC1 in targeted translational repression. Indeed it was recently shown that BC1 is associated with translational repression in dendrites to control the efficiency of dopamine D2 receptor-mediated transmission in the striatum and BC1 RNA-deleted mice exhibit behavioural changes with reduced exploration and increased anxiety (Centonze *et al.,* 2007).

3.3 Role of lncRNAs in Tumor genesis

3.3.1 LncRNAs in metastasis

3.3.1.1 HOTAIR - HOX antisense intergenic RNA

HOTAIR is 2.2 kb gene localized within the human HOXC gene cluster on the long arm of chromosome 2. It has been shown that this lincRNA has a potential to regulate HOXD genes in *trans* via the recruitment of polycomb repressive complex 2 (PRC2), followed by the trimethylation of lysine 27 of histone H3. In general, the 5’ region of the RNA binds the PRC2 complex responsible for H3K27 methylation, while the 3′region of HOTAIR binds LSD1 (flavin-dependent monoamine oxidase), a histone lysine demethylase that mediates enzymatic demethylation of H3K4Me2.

HOTAIR exists in mammals, has poorly conserved sequences and considerably conserved structures, and has evolved faster than nearby HOXC genes. HOTAIR was one of the first metastasis-associated lncRNAs, described to have a fundamental role in cancer (Figure 8). This lncRNA was found to be highly upregulated in both primary and metastatic breast tumors, showing up to 2000-fold increased transcription (Yu *et al*., 2012).

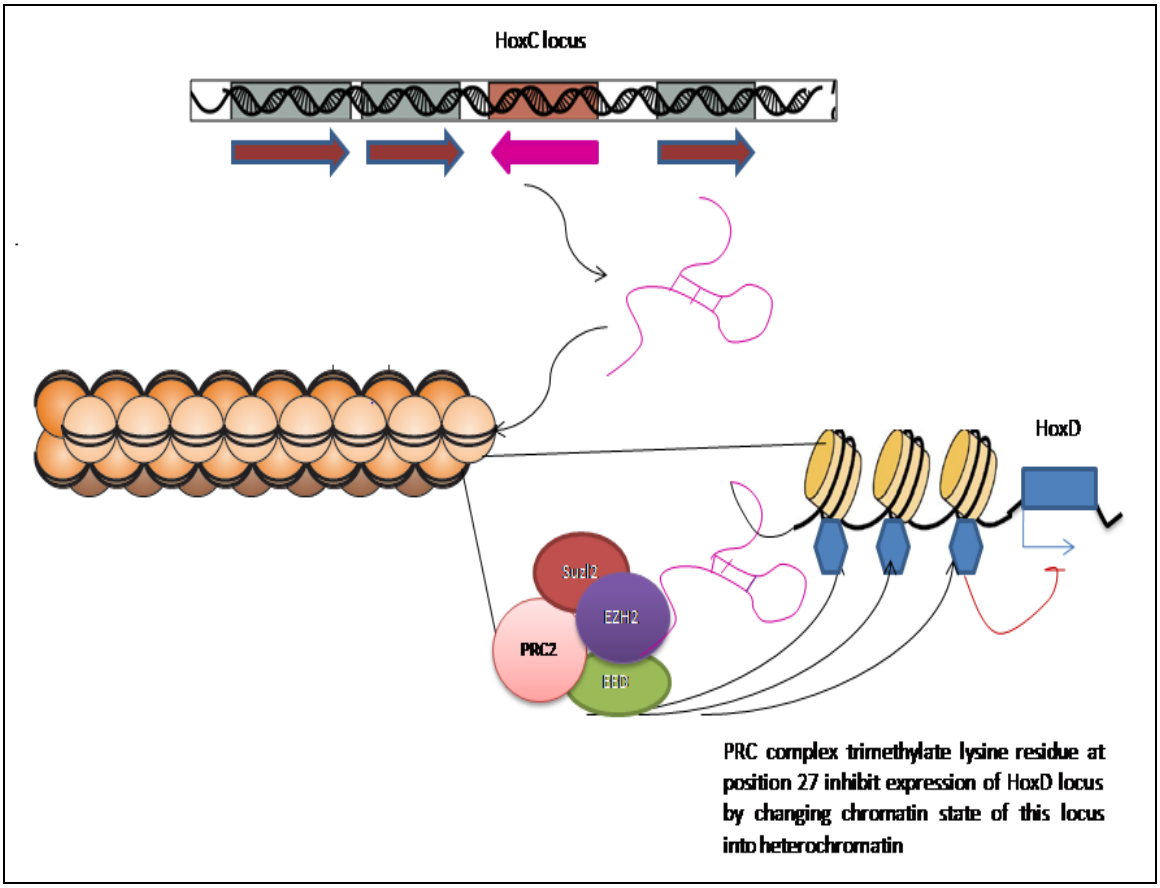


Figure 8: HOTAIR mediated gene silencing of 40 kb of the HOXD locus

3.3.1.2 MALAT1 - Metastasis-associated lung adenocarcinoma transcript 1

This lncRNA is widely expressed in normal human tissues and is found to be upregulated in a variety of human cancers of the breast, prostate, colon, liver and uterus. Cellular MALAT1 transcripts are subject to post-transcriptional processing to yield a short, tRNAlike molecule mascRNA and a long MALAT1 transcript with a poly (A) tail-like moiety. Ribonuclease (RNase) P processing generates the 3' end of the long MALAT1 transcript and the 5' end of the mascRNA. The shorter mascRNA adopts a tRNA clover-leaf structure and is subject to RNaseZ processing and the addition of a CCA to its 3' end before being exported to the cytoplasm (Figure 9). The long MALAT1 transcript is not polyadenylated, but has a poly (A) tail-like sequence that is genome encoded and is putatively present to protect the MALAT1 transcript from degradation. Moreover, MALAT1 localizes to nuclear speckles in a transcription dependent manner(Gibb *et al.*, 2011).

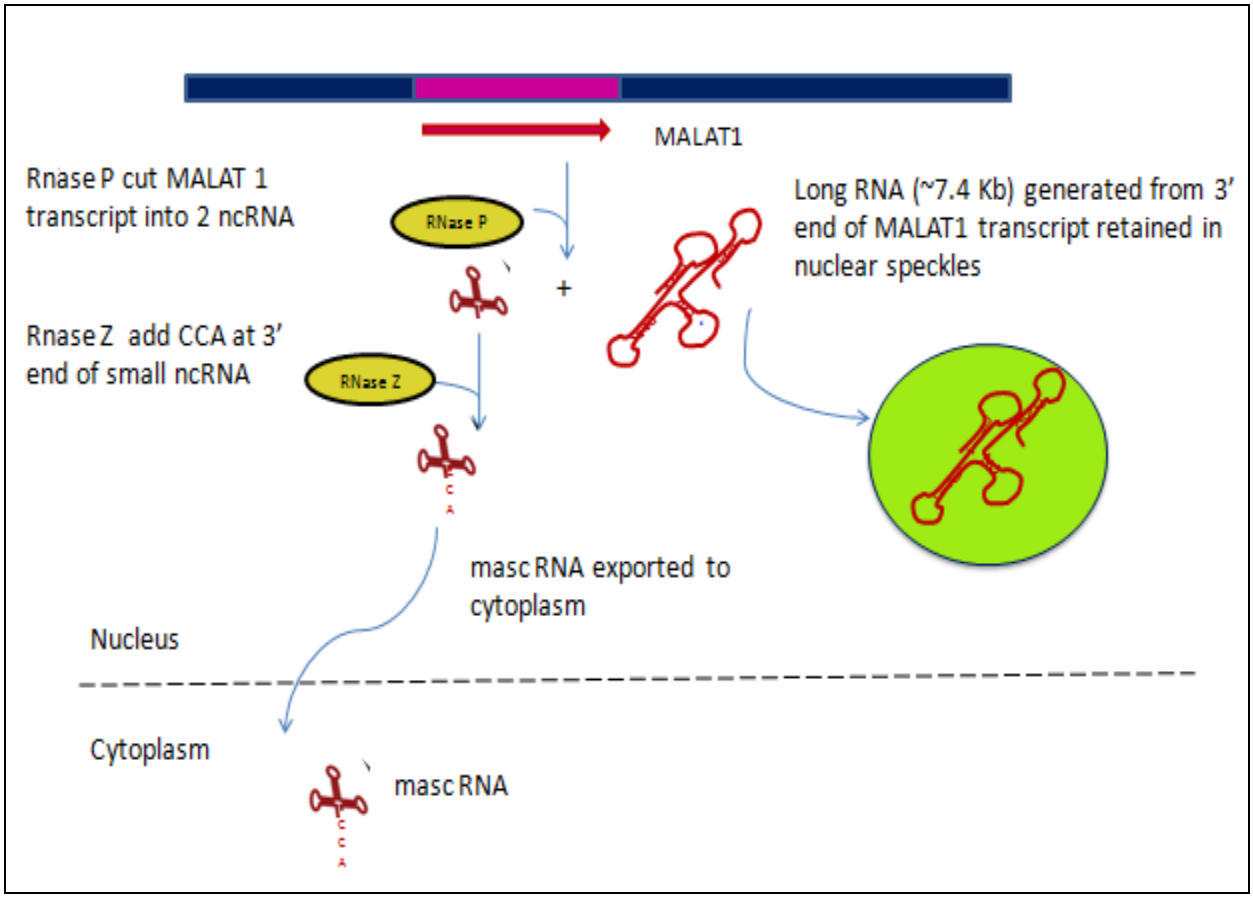


Figure 9: Expression and processing of MALAT1 transcripts

3.3.2 LncRNAs can act as natural 'miRNA sponges' to reduce miRNA levels

HULC - Highly Upregulated in Liver Cancer

HULC is a lncRNA transcribed from chromosome 6p24.3, having highly unregulated expression in liver cancer. This lncRNA consist of all hallmarks of mRNA molecule, like GT-AG intron, polyadenylation signal and nuclear export signal. The kinase PRKACB phosphorylates CREB which turns into active form and associate with RNA polymerase II in order to activate HULC expression.Excess HULC RNA inactivates the suppressive effect of miR-372 molecule. PRKACB levels have a negative relation with translational suppression i.e. increase in PRKACB level is observed with increase in translational suppression of HULC.(Figure 10)(Gibb *et al.*, 2011).

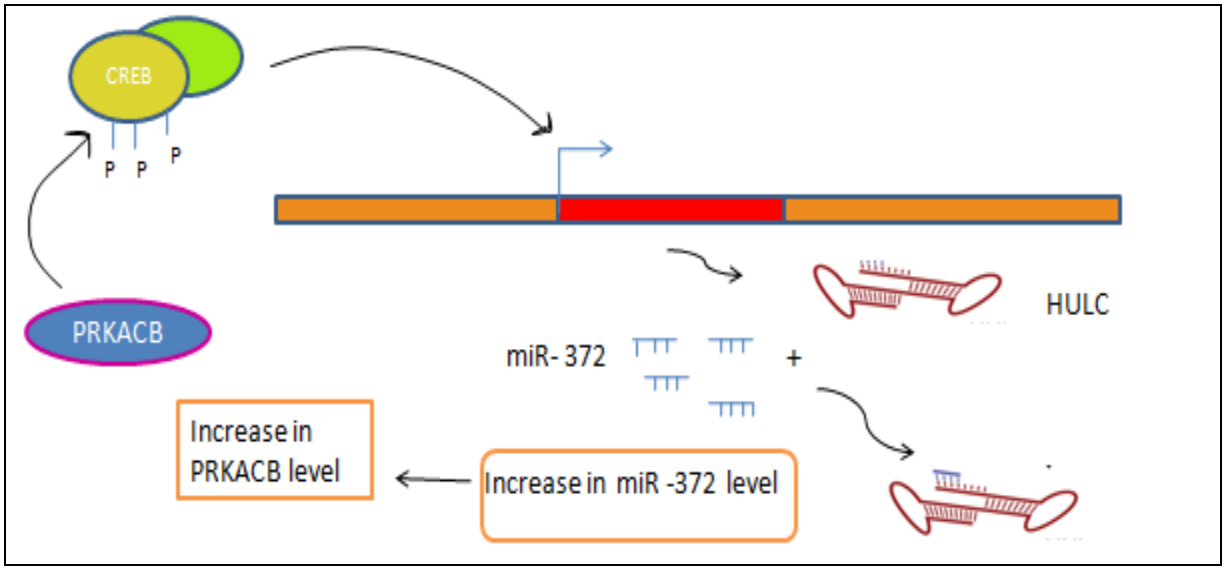


Figure 10: Proposed mechanism of HULC upregulation in hepatocellular carcinoma

3.3.3 Aberrant T-UCR expression in human carcinoma

Transcribed ultraconserved Regions (T-UCRs) are evolutionary conserved sequences found in both intergenic and intragenic regions of the human genome. The transcription products of T-UCRs are 200-779 nt in length and were originally classed into three categories, non-exonic, exonic and possibly exonic, according to their overlap with known protein-coding genes. More recently, T-UCRs have been re-annotated into a more descriptive set of five categories (Gibb *et al.*, 2011):

- Intergenic (38.7%),
- Intronic (42.6%),
- Exonic (4.2%),
- Partly exonic (5%) or
- Exon containing (5.6%).

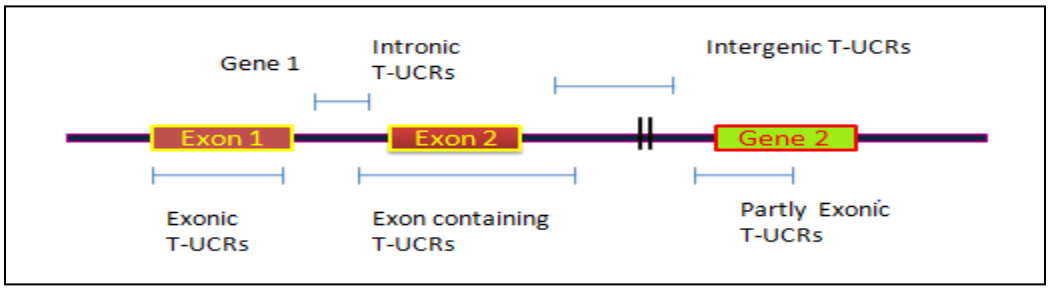


Figure 11: Genomic locations of the five classes of T-UCRs

The high degree of conservation of T-UCRs, combined with their tissue-specific expression, suggests these ncRNAs may play a critical role in cellular metabolism and development. The expression of many T-UCRs is significantly altered in cancer, notably in adult chronic lymphocytic leukemias, colorectal and hepatocellular carcinomas and neuroblastomas. Their aberrant transcription profiles can be used to differentiate types of human cancers and have been linked to patient outcome (Cheetham *et al*., 2013;Gibb *et al.,* 2011). Various lncRNAs involved in tumors and cancers are listed in Table 2 along with their function.

LncRNA	**Function**
H19	Imprinted at the lgf2 locus; controls igf2 expression *in cis*, implicated in both tumor suppressors and oncogenes.
HOTAIR	Intergenic transcript of HoxC locus, gene silencing in trans through interacting with PCR2 and LSD1 complex, involved in breast cancer metastasis
AIR	Imprinted, monoallelically expressed from the paternal allele, interacts with histone methyltransferase G9a
ANRIL	Antisense transcript of INK4n/ARF/INK4a and p15/CDKN2B, required for the PRC2 recruitment to and silencing p15INK4b tumor suppressor gene
LincRNA ROR	Expressed in the induced pluripotent stem cells (iPSCs), involved in the conversionof lineage-committed cells by interacting with reprogramming complexes
TERRA	Telomeric UUAGG repeat-containing RNA, inhibits telomerase activity, also regulates Xist and HOTAIR
PTENP1	Transcript of PTEN tumor suppressor pseudogene, PTENP1 3'-UTR exerts a tumor suppressive function by acting as a decoy for PTEN-targeting miRNAs
HAR1	REST target gene, decreased in the neurons of Huntington's disease
CCND1/ Cyclin D1	Transcribed from 5' end of Cyclin D1 gene, induced by DNA damage and binding to TLS protein, leading to allosteric changes and repression of Cyclin D1 and anti-sense transcripts of tie-1 related to vascular malformation
SRA-1	Alternative splicing of SRA-1, loss of coding frame, an increased expression is associated with tumor metastasis
MALAT-1/ NEAT2	Expressed in many cancers, regulates alternative splicing of pre-mRNA and promotes cell motility through transcriptional and post-transcriptional regulatior of motility related gene expression
Xist	Mosaic expression, spreads on Xi *in cis*, interacts with BRCA1, correlated with breast cancer, cervical, ovarian, and testis tumors
Tsix	Antisense transcript to Xist , prevents Xist stabilization and inhibits the interaction between Rep A and PRC2, silencing Xist expression

Table 2: Identified tumor and disease-associated LncRNA

3.4LncRNAs in *P. falciparum* 3D7

It has been found during research that *P. falciparum* has been evolved to escape host immunity due to expression of highly variant immunodominant erythrocyte surface molecules, PfEMP1 (*Plasmodium falciparum* erythrocyte membrane protein). This variant surface molecule is encoded by a member of multigene family, *var*, that contain 60 genes and they are regulated in such a way that at a time only one gene is expressed and rest 59 genes remain in silenced state. Switching between *var* genes for monoallelic expression leads to expression of variant molecule on surface of erythrocytes. Another interesting feature of PfEMP1 is the presence of a cytoadherence domain in it, which aids parasitized erythrocytes to adhere to vascular endothelium thereby escaping clearance by spleen (Epp*et al.,* 2009;Florens*et al.,* 2002).
Apart from PfEMP1, members of other variant gene families are also expressed on erythrocytes surface i.e. *rif* (repetitive interspersed family), *stevor*, *Pfmc-2TM*, and *surf* (surface associated interspersed gene). Out of these*rif* ,*stevor* and *Pfmc-2TM* has a two transmembrane topology and show clonally variant expression just like *var* gene family (Scherf *et al.,* 2008). Still PfEMP1 is considered to be the cause of antigenic variation of *Plasmodium* because of two reasons. Firstly PfEMP1 expression is quite earlier in erythrocyte stage than other gene families and secondly till now no evidence of surface reactivity of other variant gene families with antibodies is reported (Florens *et al.,* 2002).

When comes to genome organization of *var* gene family it has been observed that almost 60 % of *var* genes are located in subtelomeric region adjacent to six distinct TARE elements. A typical *var* gene has two exons, exon 1 (3.5 – 9 Kb) and exon 2 (1 - 1.3 Kb), one conserved intron and a conserved 5' flanking region (Figure 12). Exon 1 codes for polymorphic extracellular domain and have variable number of duffy-binding-like (DBL) domain which contributes to adhesive property of PfEMP1. Exon 2 codes for intracellular domain (Florens *et al.,* 2002). Two promoter regulate *var* gene expression of which one is located upstream of the exon 1 which is responsible for *var* gene transcription and other is located in the conserved intron. Intron promotor shows bidirectional transcriptional activity and recruits RNA Polymerase II to produce sense ncRNA of exon 2 and antisense ncRNA of *var* exon 1 in trophozoite and schizont stage (Figure 12). These ncRNA associate with chromatin or various chromatin associated factors having RNA binding domain at *var* loci in order to regulate its monoallelic expression by silencing all other genes by chromatin modification (Epp *et al.,* 2009). But it has been observed that intron promotor is active in both on and off state of *var* gene. Also *var* genes are flanked by insulator sequences to prevent silencing of one *var* gene to spread to adjacent *var* gene. This suggests that both intron promotor and noncoding transcripts are required for *var* gene regulation (Florens *et al.,* 2002).

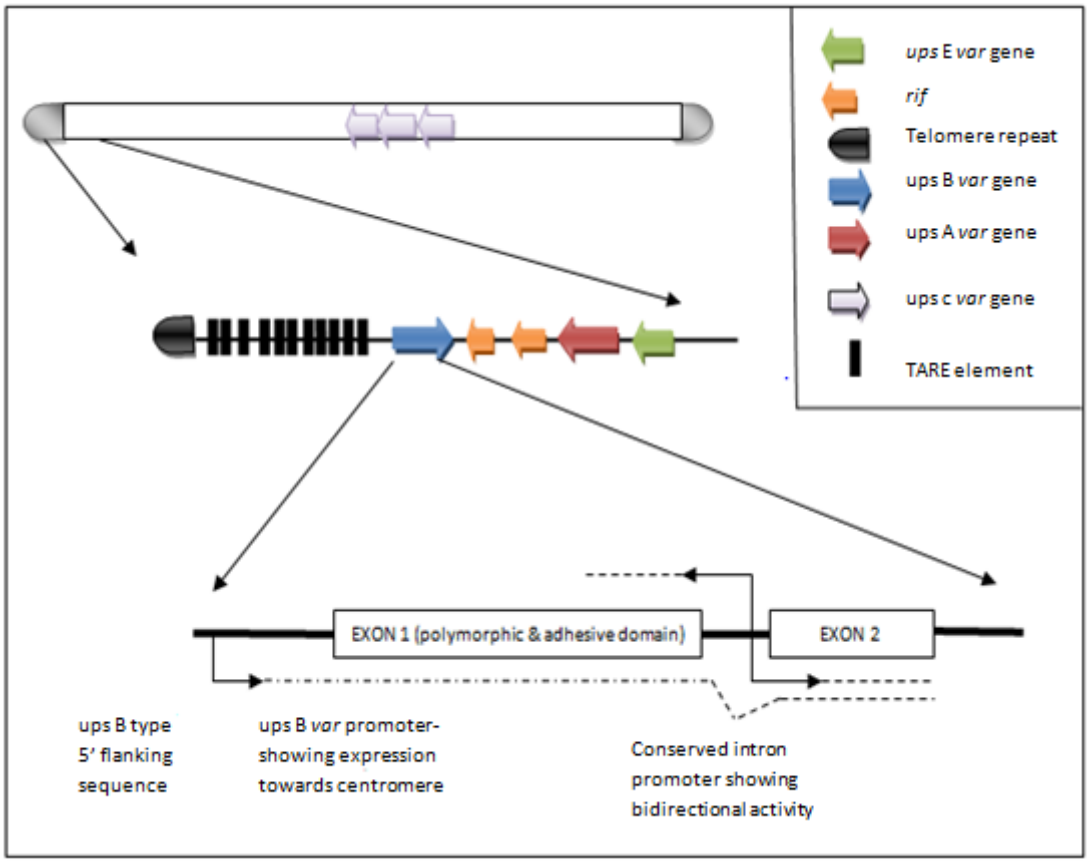


Figure 12: Genomic organization and structure of *var* gene

Majourity of *var* genes are located in subtelomeric regions next to telomere-associated repeat elements (TARE) which are found just after telomere repeat ends. TARE elements are followed by *var* genes. *var* gene contain two exons- exon 1(coding for polymorphic and duffy binding –like adhesive domain) and exon 2 (coding for intracellular region), one conserved intron and a 5' flanking conserved sequence.*var* genes can be classified into three types based on 5' flanking sequence present in their gene structure i.e. ups A type (expression towards telomere), ups B type (expression towards centromere) and ups C type (internal var cluster). Between exon 1 and exon 2 lies a conserved intron that contain a bidirectional promoter.

In addition to intron derived ncRNAs, subtelomeric regions also encode at least two TARE (telomere associated repetitive elements) ncRNA i.e. lncTARE-3-Telomere (~ 4 Kb) produced from TARE-3 repeat till telomeric repeat and lncTARE-6 (8.4 – 21 Kb) that contain 21 bp TARE-6 repeat. These repeats interact with each other to form long and multiple stem loop structure which can bind to histones and bring them at *var* gene loci thereby regulating their expression by assembly or disassembly of heterochromatin at telomere. Infact TARE-6 can be called as histone chaperone (Scherf *et al.*, 2008).

Presence of these lncRNA is evidenced in the schizont stage which later processed into smaller and stable ncRNA in the ring stage and also these lncRNAs are not localized in any of known nuclear subcompartment instead localize in novel perinuclear compartment at nuclear periphery (Scherf *et al.*, 2008).

Brodbent *et al.* identified 60 putative lncRNAs in *P. falciparum*, further characterization of which leads to identification of a family of 22 lncRNA-TARE transcripts localized in TARE 2-3 subtelomeric repeats. Adjacent to these lncRNAs, ups B type *var* genes are located and like ups B *var* gene upstream element they also contain SPE2 binding sites of transcription factor PfSip2 (*P. falciparum* SPE2- interacting protein) suggesting their role in ups B var gene regulation (Broadbent *et al.,* 2011).

Both lncRNA-TAREs and intron promoter derived sense and antisense ncRNA show post S phase expression. So that these lncRNA can create epigenetic memory marks before cell division in order to make epigenetic regulatory information inheritable (Broadbent *et al.,* 2011).

3.5 Next generation sequencing techniques

Traditional sanger sequencing method was not only slow but also very costly. To overcome the limited scalability of traditional sequencing approach massively parallel sequencing methods are developed. Massively parallel sequencing approaches involve sequencing of millions of reads in parallel by attaching DNA sample to be sequenced onto a bead or any solid surface or by creating microreactor (Jorge S Reis-Filho, 2009). These techniques are capable of providing DNA/RNA sequence from single template molecule. It has advantages of providing millions of read in a single run very cheaply. The reads generated are shorter in ranging from ~21 to ~400 base pairs. Their quantification can allow copy number estimation of each genomic region, identification of somatic mutation in non-modal population of cells.When applied to sequencing of RNA molecules, aid in identification of novel gene rearrangements, novel splice variants, novel fusion genes, read-throughs, etc. At present, there are four technologiescommercially available, namely, 454 pyrosequencing, illumina genome analyzer, AB SOLiD and HeliScope (Jorge S Reis-Filho, 2009). All of these technique follow same workflow broadly grouped as template/library preparation, sequencing and imaging and data analysis but are quite different in there biochemistry and protocol.

3.5.1 Template preparation and immobilization strategies

Genomic DNA is fragmented into smaller fragments. These fragments are usually then immobilized on some solid surface/ beads. Two types of templates can be used in next generation sequencing approach (Michael L. Metzker, 2010):

I. Clonally amplified template

Imaging system is not efficient in recognizing single fluorescent labeled template. Therefore there is need to amplify template molecule. Generally two types of methods are employed for amplification, namely, emulsion PCR (emPCR) and solid phase PCR (Figure 14).

Emulsion PCR involves creation of a library of fragments or mate-pair targets followed by ligation of adapters to the end of fragments. ssDNA is then captured onto beads such that one bead have one DNA molecule ligated on it. Amplification of these DNA strands is then carried out at these beads. After amplification each bead will have millions of copies of single DNA molecule (Shendure *et al*., 2008).

Solid-phase amplificationinvolves clonally amplification of fragments on glass slides on which forward and reverse primers are immobilized. 100- 200 million templates clusters can be obtained from solid-phase amplification, which can be amplified by using a universal primer (Shendure *et al*., 2008).

II. Single molecule template

Quantitative applications, such as RNA–seq, perform more effectively with non-amplified template sources, which do not alter the representational abundance of mRNA molecules (M. L. Metzker, 2005).
Singlemolecule templates are usually immobilized on solid supports using one of following three approaches (Figure 13) (Michael L. Metzker, 2010).

a) **Immobilization by primer**: In this approach, spatially distributed primers are immobilized on solid support onto which adaptor ligated genomic fragments are hybridized.

b) **Immobilization by template**: In this approach, spatially distributed single- molecule templates are immobilized onto the solid support. A universal primer is then hybridized to the template to carry out NGS reaction.

c) **Immobilization by polymerase**: In this approach, spatially distributed polymerase molecules are immobilized onto the solid surface which can bind primed templates and carry out their amplification. Pacific Biosciences use this approach for immobilization of templates. Also this approach can be used with real time methods therefore provide longer read length.

First two approaches are used by Helicos BioSciences.

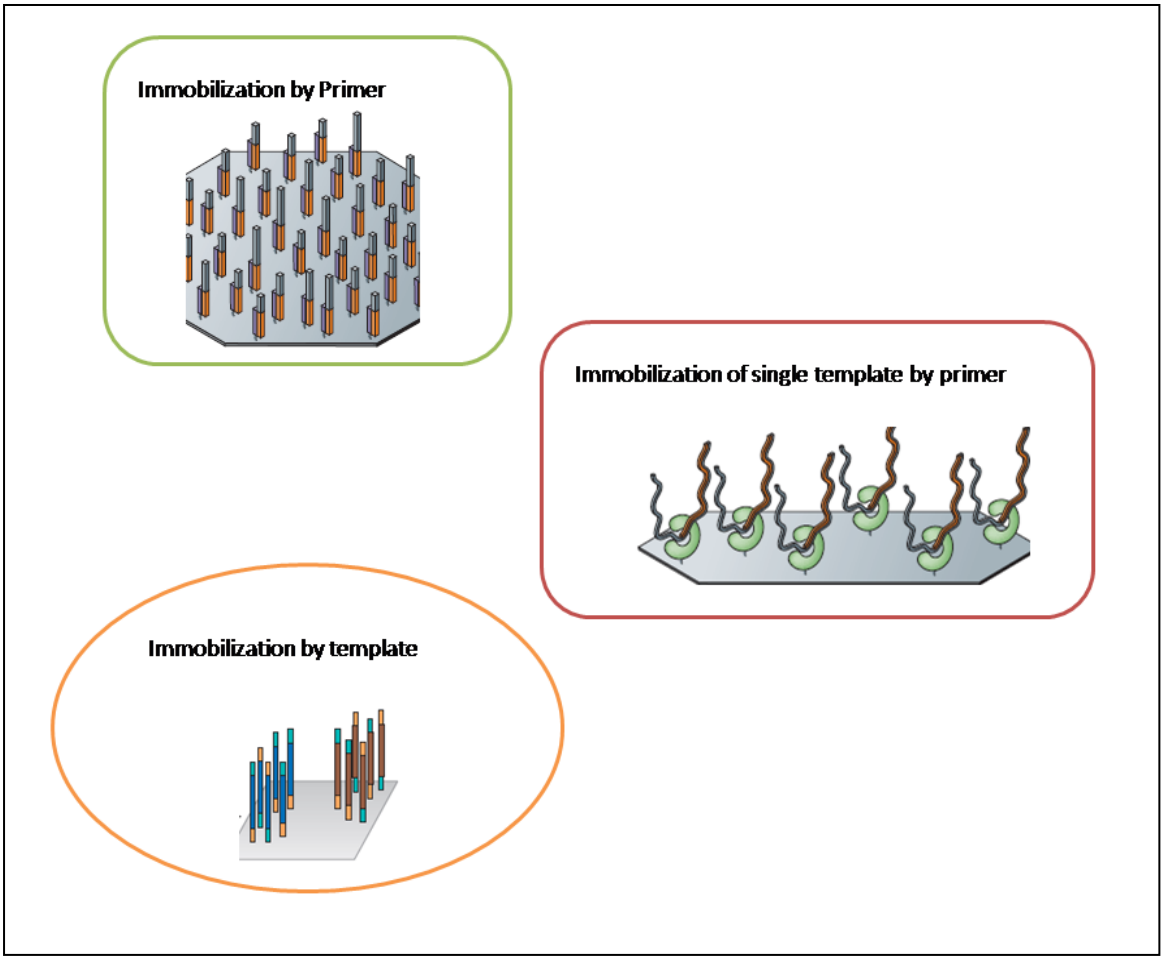


Figure 13: Various template immobilization strategies

3.5.2 Commercially available next generation sequencing technologies

3.5.2.1 454 pyrosequencing

It is the first commercially available next generation sequencing platform.After library preparation, clonally amplification of template is performed using water-in-oil emulsion PCR. Primers are immobilized on 28µm beads. These beads are placed in water-in-oil emulsion drop along with adaptor ligated template to carry out amplification of template. The amplicon-bearing beads are then isolated from emulsion. A universal adaptor is used as a template for designing primer for sequencing reaction.Pyrosequencing technique is used for sequencing clonally amplified fragments. Sequencing reaction is carried out in picolitrescale wells to give it an semi-ordered form of array where one side utilized for introduction and removal of reagents and other side is fitted with fibre-optic bundle for CCD- based signal detection. Firstly, amplicon-bearing beads are incubated with *Bacillusstearothermophilus* (*Bst*) polymerase and single-stranded binding protein followed by their placement in picoliter wells along with smaller beads

immobilized with enzymes required for pyrosequencing (ATP sulfurylase and luciferase). CCD device will detect signal when nucleotide is incorporated thereby revealing sequence of template represented by individual bead (Shendure *et al*., 2008).

3.5.2.2 Illumina Genome Analyzer/ Solexa

This technique utilizes bridge PCR to carry out amplification of template library. In this approach, both forward and reverse PCR primers are immobilized onto a solid substrate with flexible linkers attached at 5' end. Due to which amplicons originating from single template during amplification remain immobilized in a single cluster on a solid surface. Millions of spatially distributed clusters can be amplified in different lanes in a single run. Of which each cluster contain ~1000 clonal amplicons. After cluster generation, the amplicons are denatured and a primer complementary to known sequence flanking the region of interest is hybridized to single stranded clonal amplicons. Pyrosequencing is performed on bridge PCR surface itself. In each cycle incorporation of a single base is detected by adding polymerase and four modified deoxyribonucleotides (with blocking agent attached at 3' hydroxyl residue that allow incorporation of only a single base) each labeled with different fluorescent dye. After single base extension and acquisition of images, fluorescent dye and blocking agent are cleaved to carry out next cycle (Figure 14). This technique provides a read length of around 36 bp. Although longer reads are possible but error rate will be higher in that case (Shendure *et al*., 2008).

Average raw error rates are on the order of 1–1.5%, but higher accuracy bases with error rates of 0.1% or less can be identified through quality metrics associated with each base-call (Sundquist *et al*.,2007).

3.5.2.3 HeliScope

This technique does not require clonal amplification of templates instead use single DNA molecule as template for sequencing. It also uses pyrosequencing to sequence DNA molecule but need high efficiency fluorescent detection system due to use of single template. Templates are hybridized onto a solid surface at which poly-T oligomers are immobilized. In each cycle single fluorescently labeled nucleotides along with polymerase is added resulting in either single base extension if fluorescent signal detected or another base is added. Hundreds of cycle will provide read length of 25 bp or more (Figure 14) (Shendure *et al*., 2008).

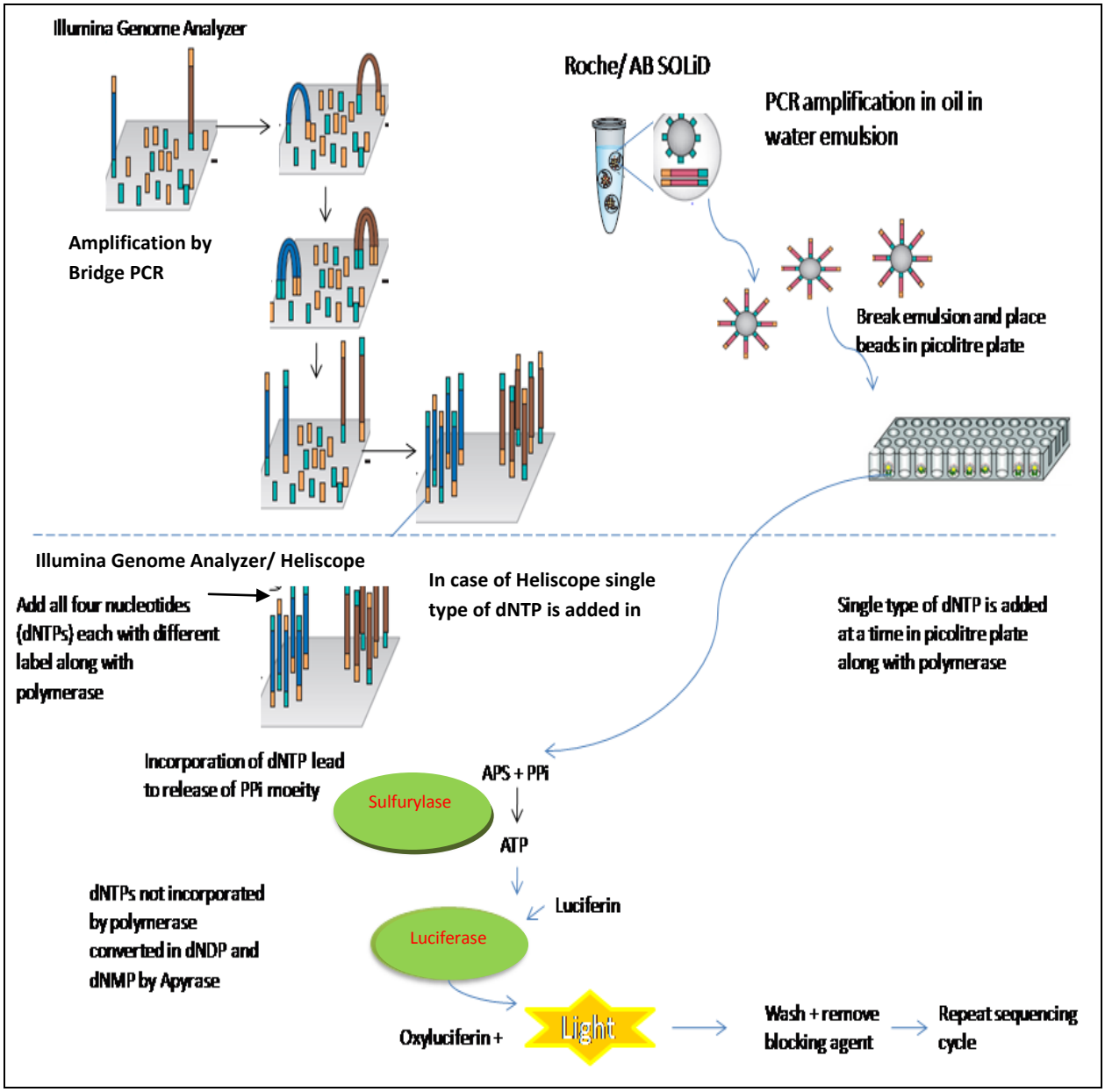


Figure 14: Pyrosequencing by illumina genome analyzer, Roche and HeliScope

3.5.2.4 AB SOLiD

This technique use emulsion PCR to carry out clonal amplification of constructed libraries, which results in millions of amplicons immobilized on micro-bead. Beads bearing amplification products are extracted from emulsion and immobilized on another solid surface(Figure 14). It also employs sequencing by synthesis approach but in spite of polymerase involves use of DNA ligase. A universal primer complementary to adaptor sequence is hybridized to the array of

amplicon-bearing beads. Each cycle of sequencing involves ligation of fluorescently labeled population of octamers. Population of octamers is designed in such a way that central 2 bp identities will be correlated with the label which enables two-base encoding. After images are acquired of ligation process the octamer is cleaved to remove fluorescent label at position 5 and 6. Iteration of ligation process will provide sequence of every 5^{th} base. After few cycles the primer is denatured and new primer (hybridize one or ore bp back to adaptor-insert junction) /set of octamers are designed in which different position is correlated with label for example at position 3. Next iteration will provide sequence of every 5^{th} base from position 3. This technique will provide highly accurate result due to two-base encoding scheme(Sundquist*et al*.,2007).

3.6 RNA-Seq Annotation Pipeline

RNA-Seq reads alignment to the refernence genome is the most crucial and challenging task in RNA-Seq analysis. Because of short length (~36-125 bases), high error rate, large number of reads (hundreds of millions) per experiments and due to span of reads along exon-exon junction. Typical RNA-Seq analysis include four major steps –

3.6.1 Mapping short RNA-Seq reads to reference genome

There are two major algorithmic approaches to map RNA-Seq reads to a reference genome.

3.6.1.1 Unspliced read aligners

These types of aligners do not allow large gaps while aligning reads to reference genome. They can also be classified in two categories based on algorithm used for unspliced alignment (Garber *et al*., 2011).

i. Seed methods

This algorithm works by finding short subsequences that exactly matches the reference transcriptome which is then extended in both directions by more sensitive alignment algorithm like Smith-Waterman algorithm to get full alignment. Short subsequences are termed as seeds thereby method is known as seed method. Seed methods provide high sensitivity results in case of alignment to polymorphic regions in reference transcriptome in order to quantify allele-specific expression and also when transcriptome on which reads are aligned is taken from distant species. Packages such as MAQ, Stampy and Short-read mapping package (SHRiMP) are based on seed method based unspliced aligners algorithm (Garber *et al*., 2011).

ii. Burrows-Wheeler transform methods

Burrows-Wheeler transform methods indexes the reference genome or transcriptome using a scheme based on the Burrows-Wheeler transform (BWT) approach. BWT-based indexing compact the genome into index files that aids in fast and accurate alignment utilizing small memory. It performs better when mismatches are not allowed, with each allowed mismatch the performance of algorithm is compromised in exponential manner. They perform faster than seed based method and can be opt as a choice of algorithm for alignment in case of availability of exact reference transcriptome or genome. BWT based unspliced read aligners cannot be used for identification of novel splicing events and are limited to identification of already known exons. Bowtie and BWA are the freely available packages meant for BWT based unspliced read alignment (Langmead *et al*., 2009).

3.6.1.2 Spliced read aligners

Spliced aligners allow gaps in order to align intron spanning reads. Spliced read aligners can be classified into two categories:-

i. Exon first aligners

TopHat, MapSplice and SpliceMap belongs to a category of spliced aligner and works in two step –

1. In the first step, all reads are mapped to the genome using Bowtie, an unspliced read aligner. It map all non-junction reads to the genome using Burrows-Wheeler transform method. This Burrows- Wheeler indexing approach makes it ultrafast and memory-efficient programme for alignment of short reads when exact reference genome is available(Langmead *et al*., 2009).

2. In the second step, unmapped reads are split into smaller segments which are then independently aligned to reference genome. These mapped reads are then extended to find possible splice sites to determine spliced alingnment of initially unmappable read (Trapnell*et al*., 2009) (Figure 15).

These aligners work very fast in case if only few unmappable reads are left after first step because second step is quite computationally intensive. When compared to seed and extend based spliced aligners exon first methods are always faster and less computationally intensive. Exon-first approaches can miss some spliced alignments

especially for reads that map to genes that have pseudogenes counterpart spread in genome(Trapnell*et al*., 2009).

ii. Seed and extend aligners

As the name is suggesting these types of aligners break the reads into shorter subsequences, termed as seeds, followed by alignment of these reads onto the genome to find exact matches. These seeds are then extending in both directions to get full alignment via using more sensitive alignment algorithm like Smith-Waterman algorithm or iteratively extend and merge in order to get correct spliced alignment (Figure 15)(Garber *et al*., 2011).

Some of the softwares for seed and extend based spliced read aligners are GSNAP (genomic short-read nucleotide alignment program) and QPALMA (computing accurate spliced alignments). Seed extend method is slow when compared to exon-first method but can find more spliced alignment because it perform both spliced and unspliced alignment in a single step therefore there are less chances of biasness towards reads unspliced alignment. Also seed extend method give better performance in case of alignment to polymorphic sites (Garber *et al*., 2011).

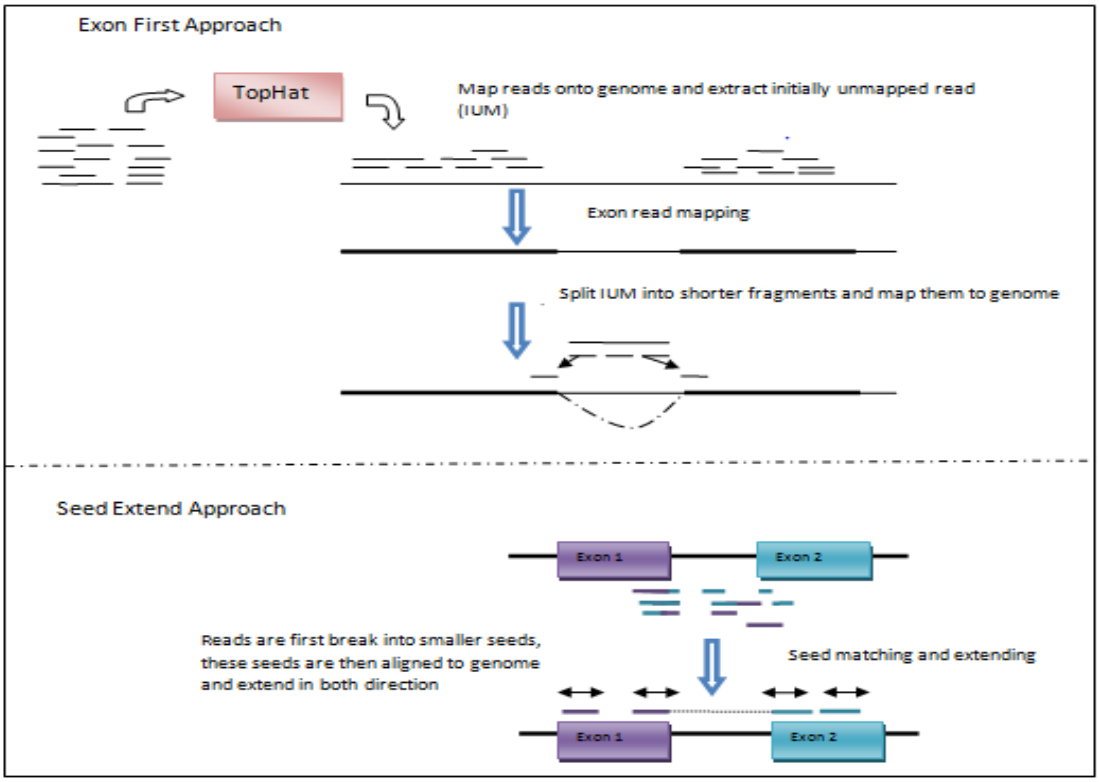


Figure 15: Strategies adopted by spliced read aligners

3.6.2Transcriptome Reconstruction

Transcriptome reconstruction involves assembly of aligned reads into transcripts in the form of overlap graph, provided reads are overlapping and compatible with genome. Methods for reconstruction of transcriptome can be classified into two main classes-

3.6.2.1 Genome-guided reconstruction

Genome-guided methods involves mapping of reads firest to a reference genome, after which mapped reads are assembled into transcripts. Genome guided methods itself can be classified into two types-

1. Exon identification

Exon identification method relies on identification of exon as coverage island and then define boundaries of exons based on reads spanning across these coverage islands. This method is suitable for transcriptome reconstruction in case of short reads (~36 bases) with few aligned exon-exon junction.G.mor.se is one of the algorithms available for defining transcriptome based on this approach. In case of long and alternatively spliced genes with low level of expression, it does not provide good results. Genome –guided assembly approaches should be considered in such cases(Figure 16) (Wu*et al*., 2005).

2. Genome-guided assembly approaches

Cufflinks and Scriptures are based on genome-guided approaches; useful for transcriptome reconstruction in case of longer read length. These methodsuse spliced reads directly to reconstructthe transcriptome.

Scripture solve transcript reconstruction problems by transforming the genome into a graph topology which represents all possible connectionsof bases in either consecutive manner or connected by spliced reads. It provides increased sensitivity results in predicting low level expressed transcripts by analyzing all possible paths through the graph and report all isoforms possible in a given read data.

Other software based on genome-guided approach is Cufflinks that works with maximum precision by reporting minimal number of paths that can describe all isoforms possible in a given data set. Cufflink is one of the freely available software used for constructing the whole transcriptome map of sample. It first divides the fragments into non-overlapping loci and then assembles each locus independently. Each fragment is treated as a node in an overlap graph and a directed edge is placed between each pair of compatible fragment

(overlapping fragments having characteristics of identical inherent intron) followed by filtering out of incompatible fragments that may be originated from different transcript isoform (Trapnell *et al*., 2010). Cufflinks implements a proof of Dilworth's Theorem by producing minimum set of path that can cover all fragments in directed overlap graph. For which it first find largest set of incompatible reads by finding maximum cardinality matching in a bipartite graph and assign nodes that do not have incident edges as a member of antichain. Extension of these members into path provides a minimum path cover (Garber *et al*., 2011).

Both Scripture and Cufflinks utilize similar computational power and gives almost similar number of highly expressed transcripts. But when it comes to prediction of lower expressed genes Cufflinks report three times more transcripts than Scripture whereas Scripture reports more isoforms per locus than Cufflinks (Wu *et al*., 2005).

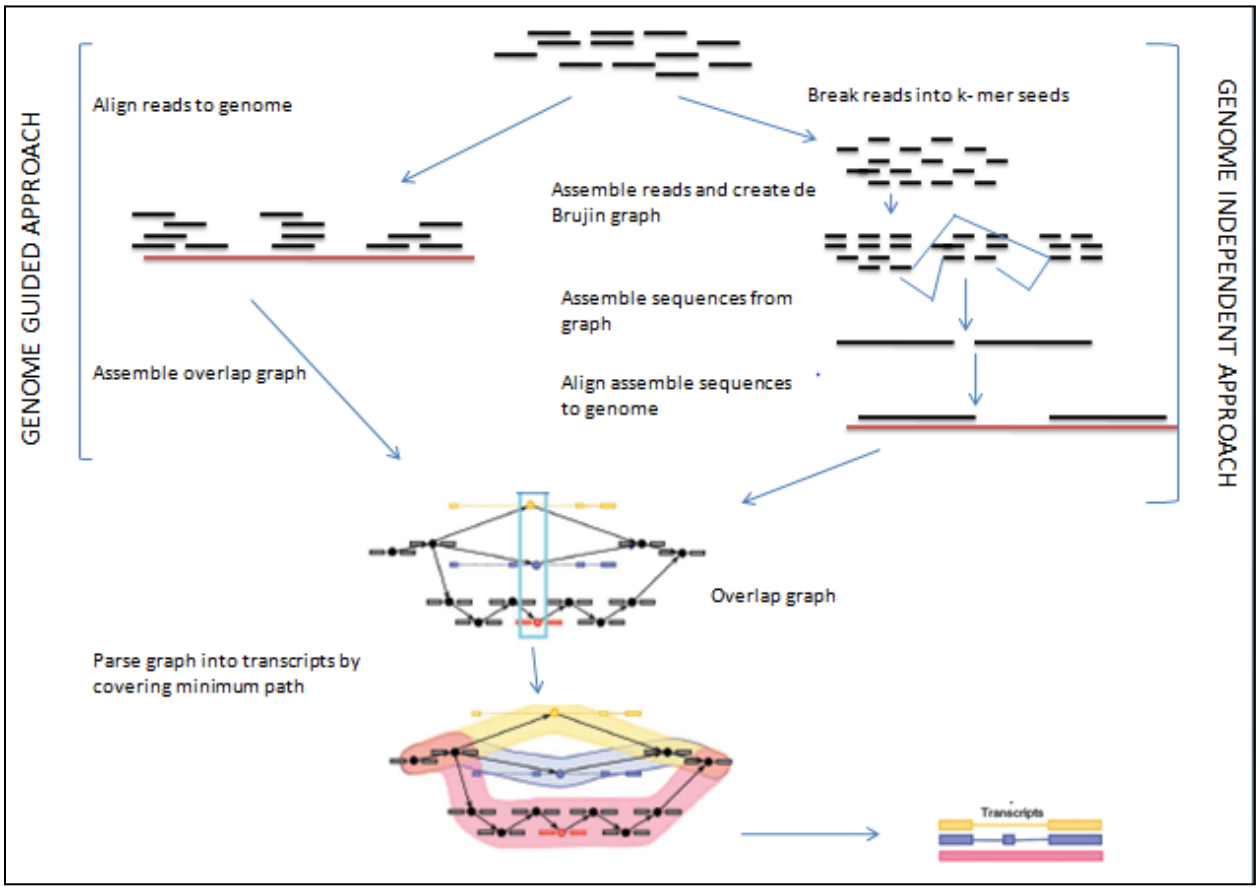


Figure 16: Transcriptome reconstruction methods

3.6.2.2 Genome-independent reconstruction

Genome-independent methods do not utilize reference genome and directly assemble reads into transcripts.transAbyss is one of the example of algorithm based on this approach. transAbyss first build consensus transcripts from reads and then these consensus transcripts are mapped onto the genome or protein database (Wu*et al*., 2005).

Most challenging task in genome independent approaches is to report alignment into set of tanscripts that represent all possible isoform of genes in a genome. They model overlapping sequences into *k*- mers using de Bruijn graph. The overlaps of $k - 1$base between these *k*-mers constitute the graph of all possiblesequences that can be constructed. Paths are then traversed in the graph to eliminate false branch point that are not supported by paired end reads and may be introduced by *k*-mers belong to different transcripts. Each recovered path after removal of false *k*-mer is then reported as a set of transcripts. The length of *k*-mer should be decided according to the level of coverage because it greatly affects the assembly quality. Length of *k*-mer should be small in case of low coverage; small value leads to large number of overlapping nodes resulting in more complex graph pattern. Whereas in case of high coverage large length is preferred that yields simpler graph with less number of overlapping reads.transABySS deals with this variability in expression level, it uses a variable *k*-mer length strategy to assemble transcripts utilizing more computational power(Garber *et al*., 2011).

3.6.3Reference Annotation Based Transcript (RABT) Assembly

All cufflink assemblies are then merged together using reference annotation based transcripts (RABT) assembly method. RABT assembler employs cuffmerge with –g/--GTF <reference annotation.gtf> option to incorporate reference annotation into assembly. It works in following steps (Figure 17)-

1. It first generates faux reads alignment tiling the reference transcripts in order to cover all reference transcript positions by multiple reads.
2. These faux reads are then merged with aligned sequenced read using cuffmerge. This step generates fewest possible trasfrags, having the ability to explain both types of reads.
3. These transfrags are then merged with reference transcripts to filter out noisy read mapping and identify novel transcripts (Roberts *et al*., 2011). Transfrags are discarded if reference transcripts found with following characteristics-

    i. Transfrag's 5' endpoint is found in reference transcripts..
    ii. Transfrag's 3' endpoint cannot be extended more than 600 bp outside the reference transcript.
    iii. Both transfrag's and reference transcripts does not contain an intron.

iv. It contained all introns in the reference transcript that fully lay within its boundaries.

v. Its endpoints extended no more than the mean fragment length into the intron of the reference transcripts (Roberts *et al*., 2011).

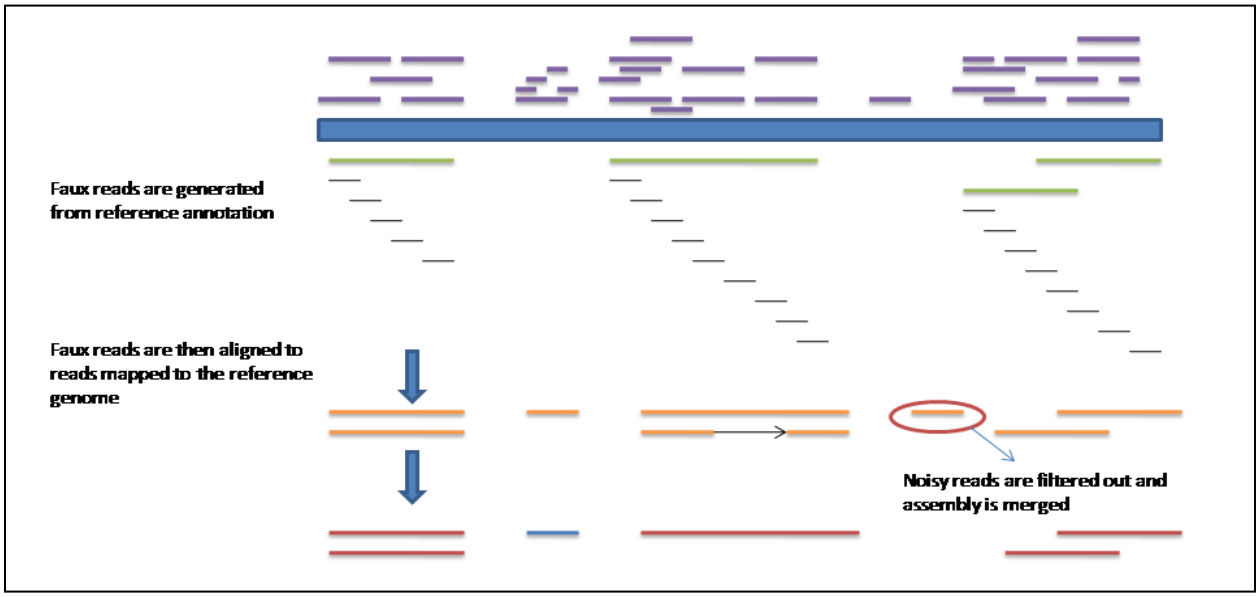


Figure 17: Overview of RABT assembly method

3.6.4Differential analysis with Cuffdiff

Cuffdiff is software for calculating differential expression in two or more samples included in Cufflinks package. It not only calculates expression level but also test the statistical significance of observed differential expression between two or more samples. It works on the assumption that number of reads produced by each transcript is directly proportional to its expression level. Cuffdiff deals with biasness in RNA-Seq experiments by modeling of large fraction of bias followed by eliminating them from each transcript. This biasness may have arisen due to difference in library preparation and due to variability in different replicates of same experiment. Cuffdiff can take multiple biological or technical replicate as a single condition in a comma separated format and set of conditions a separate by space. The purpose of recognizing multiple replicates as single condition is to estimate the variance caused in read counts for each and every gene in replicates which is then used to calculate change in expression (Trapnell *et al*., 2012).

Cuffdiff reportsFPKM (Fragments Per Kilobase of exon model per Million mapped fragments) values of each transcripts, gene and primary transcripts in separate tab separated output files. Transcripts isoforms FPKM are reported in isoforms.fpkm_tracking file. Primary transcript FPKM values are reported in cds.fpkm_tracking file by summing up fpkm values of all transcripts belong to each transcripts group. Gene's FPKM values are reported in

genes.fpkm_tracking file by summing transcripts FPKM values of all transcripts belongs to each gene group(Trapnell*et al*., 2013).

Cuffdiff also report differential expression between different samples and different specified conditions in four files-

i.	isoform exp.diff	reports transcripts differential FPKM.
ii.	gene_exp.diff	reports differential gene FPKM (difference in summed FPKM reported in cds.fpkm file).
iii.	tss_group_exp.diff	reports primary transcripts differential expression's FPKM.
iv.	Cds_exp.diff	reports coding sequence differential expression in FPKM.

These files contain familiar statistics such as fold change (in $\log_2$ scale), *P* values (both raw and corrected for multiple testing) and gene- and transcript-related attributes such as common name and location in the genome.

Cuffdiff can also identify genes that are differentially regulated by promoter switching or differentially spliced genes(Trapnell*et al*., 2013).

3.6.5 Visualization with CummeRbund

Cuffdiff provides results of differential expression in different tab separated files that can be viewed and analysed in Microsoft Excel sheets. But analyzing this output given I multiple files can't be done manually. Therefore different programme, CummeRbund, is designed to carry out analysis of cuffdiff output. CummeRbund is a user-friendly tool designed for advanced statistical analysis of cuffdiff output, for plotting and cluster analysis of differentially expressed data (Trapnell*et al*., 2012).

3.6.6Estimating coding potential of transcripts

Two softwares are used to calculate coding potential of transcripts in order to identify non coding transcripts and then their results are compared with each other to get more accurate results. But before that transcript having length more than or equal to 200 nucleotides are extracted and rest are discarded. Reverse complement of these extracted sequences is also created in a separate file.

3.6.6.1 CodingPotential Calculator (CPC)

CPC is a support vector machine based classifier, designed to classify transcripts into protein coding RNAs and ncRNAs. It assesses the protein coding potential of transcripts based on six features (Ponting *et al*., 2009) –

a) LOG-ODD Score: indicates the quality of ORF. Its value should be higher for protein coding transcripts.

b) Coverage of predicted ORF: An indicator of ORF length. Protein coding transcripts usually have long length ORF.

First two features are derived using framefinder software. It has ability to identify longest reading frame in three frames with high error tolerance.

c) Integrity of predicted ORF: indicatespresence/ absence of start and in-frame stop codon in the transcripts.

d) Number of hits found in BLAST X search against UniProt reference clusters: It also perform BLAST X search for transcripts against UniProt reference clusters in order to utilize the information stored in protein sequence databases to identify protein coding transcripts. Larger the number of hits for particular transcript higher is the possibility of transcript having coding potential.

Last two parameters are also calculated based on BLAST X search.

e) Hit Score: measures the quality of hits. Hit score is calculated using following equations-

$$\mathbf{S_i} = \sum_{n=1}^{j} (-\mathbf{log10\ Eij})$$

Where S_i is the measure of the average quality of high scoring fragment pairs (HSPs) in i^{th} frame and E_{ij} is the E-Value of j^{th} HSP in i^{th} frame.

Hit score is a measure of the average of S_i in all three frames i.e.

$$\mathbf{HIT\ Score} = \sum_{i=1}^{3} \frac{\mathbf{Si}}{\mathbf{3}}$$

f) Frame Score: It is a measure of distribution of HSPs among three reading frames.

$$\mathbf{FRAME\ SCORE} = \sum_{i=1}^{3} \frac{Si - \bar{S}}{3}$$

Frame score distinguish protein coding transcripts from noncoding transcripts by finding matches of transcripts in all three frames. Non coding transcripts are usually scattered in all three frames whereas protein coding transcripts are likely to concentrate in one frame only. Higher the frame score higher are the chances that hits are concentrate in one frame.

In all, these six features contribute to faster and higher accuracy predictions of CPC (Ponting *et al*., 2009).

3.6.6.2getorf - Jemboss

getorf is also an ORF predictor tool of an European Molecular Open Software Suite (EMBOSS) package. We used EMBOSS graphical user interface (GUI), Jemboss for ORF prediction.

4. METHODOLOGY

4.1 Data Downloaded

4.1.1 RNA-Seq Dataset

RNA-Seq data for *P. falciparum* 3D7 is downloaded from SRA (Sequence Read Archive) through DNAnexus from URLhttp://sra.dnanexus.com .

RNA-Seq data for *Plasmodium falciparum* 3D7 is submitted by two different groups in SRA database i.e. University of California- San Francisco and NIH (National Institue of Health) using Illumina Genome Analyzer II. We have downloaded both datasets to carry out our research (Table 3). Details of both studies are provided in Appendix I.

NIH		University of California	
Single end reads	**Paired end reads**	**Single end reads**	**Paired end reads**
SRR364834	SRR364836	SRR066576	SRR066581
SRR364838	SRR364841	SRR066577	SRR066582
SRR364840	SRR364842	SRR066578	SRR066583
SRR364843	SRR364846	SRR066579	SRR066584
SRR364847		SRR066580	SRR066585
SRR364848			SRR066586
SRR364849			SRR066587
			SRR066588
			SRR066589
			SRR066590

Table 3: Accession number of runs downloaded from DNAnexus

4.1.2 Reference Genome and annotation Data

Reference genome for *Plasmodium falciparum* 3D7 is downloaded from ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/3D7/3D7.latest_version/September_2011/

Annotation file is downloaded fromhttp://www.broadinstitute.org/annotation/genome/plasmodium_falciparum_spp

4.2Conversion of sra files downloaded from DNAnexus to fastq format files

RNA-Seq reads downloaded are in sra format and this format is not recognized by TopHat which we will use for alignment of reads to reference genome. So this sra format is converted into the fastq format using fastq-dump programme of SRA toolkit.

Single ended reads are converted using simple command:-

fastq-dump<sra file>

In case of paired-ended reads, sra files need to split into two fastq files. fastq-dumpcan perform this task by providing split argument that will split the sra file into three files, two paired end files and one single end file. The command for splitting is:-

fastq-dump --split-3 <sra file>

4.3Quality check using FastQC

Since error rates are quite high in RNA-Seq data, therefore we have decided to check the quality of reads before start working on this data.FastQC software is used to check the quality of raw sequence data in fastq format. FastQC gives per base sequence quality of reads and the quality score at which base-calling error is too high is used as phredcutoff. And we found that read's quality is not good and requires trimming. For trimming phredcutoff value '20' is provided which can be deciphered from FastQC graphical output.

4.4Trimming of low quality reads

SolexaQA software package was used for trimming of low quality reads. It takes .fastq files as input. This package includes three programmes i.e. SolexaQA, DynamicTrim and LengthSort. Out of which we have used DynamicTrim with "h" option i.e.by providing phredcutoff (h = 20) followed by LengthSort with length cutoff taken as '36'.

DynamicTrim trim each read to its longest contiguous segment until exceed quality cutoff specified by user and generate trimmed files.

These trimmed files are then used by LengthSort which create three files, namely, *.discard file and *.single in case of single ended files. Reads smaller than length cutoff are stored in *.discard file and rest are stored in *.single in case of single ended reads. Out of which *.single file is used as input file in TopHat.

And for paired ended files it take both paired ended file as input and provide output on the basis of length cutoff in three files – *.paired1, *.paired2 and *.single file. Both *.paired1 and *.paired2 will be used as input fastq file while running TopHat.

4.5Indexing of reference genome

Reads cannot be mapped on reference genome as it is. First we have to index reference genome using Bowtie. These indexed files are then used as input to TopHat (spliced read aligner) along with fastq files (Adam Roberts *et al.*, 2011).Index files are created using syntax:

bowtie2-build<path to genome.fasta><path to output files>

4.6Mapping short RNA-Seq reads to reference genome

RNA-Seq reads are very short reads therefore we need to align them on reference genome. TopHat is used to align the fastq reads onto genome.

In case of non-strand specific reads TopHatis used with options "-coverage search" and "-microexon search" that enable coverage based search to junction and aids in finding alignment to micro-exons respectively.

Four reads were strand specific, namely, SRR364836, SRR364841, SRR364842 and SRR364846. For these reads one more option is used i.e. "-library-type fr-unstranded".

Alignments are reported in bam format by TopHat. It's a binary format for representing sequence data. Two types of bam files are created, one is accepted_hits.bam that contains reads aligned by TopHat and other is unmapped.bam which contains reads discarded by TopHat. The commands used for mapping reads are mentioned in Appendix II.

4.7 Transcriptome Reconstruction

Transcriptome reconstruction involves assembly of aligned reads into transcripts in the form of overlap graph, provided reads are overlapping and compatible with genome. Cufflink is used for constructing the whole transcriptome map of sample. It employs a genome guided approach for assembly of overlapping aligned fragments.

Cufflinks do not recognize compressed bam file format therefore first bam files are converted into sam files using samtools. The syntax for this convertion is:-

Samtools view accepted_hits.bam > accepted_hits.sam

This will provide accepted_hits.sam as output file which is then used as input to cufflinks. The syntax for cufflinks is:-

cufflinks accepted_hits.sam

Cufflinks will create many files but the most important file is transcripts.gtf that provides information about location of assembled transcripts.

4.8 Calculation of mapping percentage and data filtering

Before going to next step we first calculate number of reads mapped by TopHat from accepted_hits.sam files. Number of reads mapped was calculated using syntax-

awk '{print $1}' accepted_hits.sam |sort |uniq |wc –l

Number of reads mapped obtained is then used to calculate mapping percentage using formula-

Mapping Percentage = (number of reads mapped/ number of reads accepted by TopHat)*100

Reads having mapping percentage less than 60 % and mapped reads less than millions are discarded.We left with only eight reads after filtering.

4.9Reference Annotation Based Transcript (RABT) Assembly

All filtered cufflink assemblies are then merged together using reference annotation based transcripts (RABT) assembly method. RABT assembler employs cuffmerge with –g/--GTF <reference annotation.gtf> option to incorporate reference annotation into assembly. The syntax used for this assembly is:-

cuffmerge -g <path to reference_annotation .gtf><path to text file containing list of paths of all read's transcripts.gtf files>

This will create a new folder 'merged_asm'. This folder will have lots of files but the most important file is merged.gtf which will contain merged information of all transcripts.gtf along with already annotated transcripts information. Cufflink assigns unique cuff_id to each un-annotated transcript.

4.10Extracting transcript sequences using gffread

Cuffmerge gives a merged file of all transcripts. Next step is to extact the fasta sequence of all assembled transcripts in merged.gtf file. gffread has the potential to perform this task. All it need is genome sequence in fasta format that was used to create genome index for mapping reads with TopHat. For making this step quicker it is recommended that index files are placed in the same directory in which genome fasta files is placed. Following command is used to perform the task of retrieval of transcripts sequences -

gffread -w transcripts.fa -g <path to genome.fa><path to merged.gtf>

4.11Estimating coding potential of transcripts

Two softwares are used to calculate coding potential of transcripts in order to identify non coding transcripts and then their results are compared with each other to get more accurate results. But before that transcript having length more than or equal to 200 nucleotides are extracted and rest are discarded using three perl scripts back to back mentioned in Appendix III.
Extracted transcripts are then used as input to getorf software of Jemboss suite. getorf is an ORF predictor tool of an European Molecular Open Software Suite (EMBOSS) package. Jemboss will predict ORF of all transcripts. Jemboss output is then parsed to remove all transcripts greater than 30 amino acids in length because they are likely to be coding transcripts. This is also done using script mentioned in Appendix IV. The output of the perl script is then edited by replacing all \t\n with \t using editor. After that a grep command is used to extract all transcripts less than 30 amino acid-

grep -v -w '[3-9][0-9]\|[1-9][0-9][0-9]' filename

Selected transcripts (less than 30 amino acids) will represent ncRNA molecule. These transcripts are further checked to remove allisoforms of discarded transcripts and other annotated transcripts. This is accomplished by using two perl script back to back (see Appendix V) to extract gene ids of discarded transcript ids and then use these gene ids to extract ids of transcripts encoded by these genes. This will give list off transcript ids which is then compared with ids of transcripts obtained after parsing for less than 30 amino acids and common ids are discarded. The ids obtained after this are then compared with transcript ids of all annotated transcripts in merged.gtf file to remove annotated transcripts. Final transcripts ids obtained after all this parsing steps are then used to extract their sequence which is then used as input to software, coding potential calculator.

Coding potential calculator (CPC) is a support vector machine based classifier, designed to classify transcripts into protein coding RNAs and ncRNAs. It provides specific score to each

transcript. Lower the score higher the chances that transcripts are noncoding.The syntax used for calculating coding potential using CPC is:-

sudo bash run_predict.sh <path to input fasta file><path to output file>

CPC output is then again parsed to get transcripts ids having score less than -0.1 using perl script (see AppendixVI). The transcripts ids obtained after parsing CPC output are considered as the final lncRNA ids. Finally the information about these transcripts is extracted from merged.gtf file using script mentioned in Appendix VII. Also sequence of these RNAs is also extracted from transcripts.fa file obtained after gffread using perl script (Appendix VIII).

4.12Differential expression analysis of lncRNAs

Two softwares are used for differential expression analysis, namely, Cuffdiff and DESeq.

4.12.1 Differential expression analysis using Cuffdiff and CummeRbund

Cuffdiff can be run immediately after cuffmerge done. It will take bam files and merged.gtf file obtained after RABT assembly as input. Cuffdiff will take bam files in space separated form in case of different samples while technical replicates (one with same sample ids) should be provided in comma separated form. The syntax used for differential expression analysis using Cuffdiff is-

cuffdiff -N -o <path to output directory>merged.gtf
accepted_hits_36.bam,accepted_hits_38.bam
accepted_hits_40.bam,accepted_hits_41.bam accepted_hits_42.bam
accepted_hits_46.bam accepted_hits_78.bam,accepted_hits_79.bam

Cuffdiff report outputs in number of files which are then analysed using CummeRbund. CummeRbund is an R environment based programme designed especially for analysis of output of Cuffdiff for plotting and statistical analysis purpose. To carry out analysis with CummeRbundlibrary of cummerbund is loaded in R and directory is set to the path where Cuffdiff output files are stored. Following commands can be executed –

```
>library(cummeRbund)
>cuff<-readCufflinks()
>cuff
```

These commands will create cuffdata.db at the backend and return cuff values. Variousmethods are available in this package for plotting dispersion, density (with or without replicates), box plot

(with or without replicates), scatter plots, dendogram and volcano plots. These plots are created using following commands-

```
>disp<-dispersionPlot(genes(cuff))
>disp
>dens<-csDensity(genes(cuff))
>dens
>densRep<-csDensity(genes(cuff),replicates=T)
>densRep

> b<-csBoxplot(genes(cuff))
>b
>brep<-csBoxplot(genes(cuff),replicates=T)
>brep

> s<-csScatterMatrix(genes(cuff))
>s

>dend<-csDendro(genes(cuff))
> dend.rep<-csDendro(genes(cuff),replicates=T)

> v<-csVolcanoMatrix(genes(cuff))
>v
```

4.12.2 Differential expression analysis using DESeq

DeSeq is also an R based package designed for differential expression analysis. It takes count table as input that contains the count values of each gene in different samples. A cell in a table indicates the count of number of reads mapped to a particular gene i in condition j. these counts can be obtained using HTSeq-count script included in HTSeq python package. Counts for each sample are obtained by using htseq-count script which takes sorted sam files and lncRNA.gtf file as input. lncRNA.gtf file is created b extracting final lncRNAs information from merged.gtf file obtained after RABT assembly. The sam files are sorted using command-

sort –s –k 1,1 accepted_hits.sam > sorted_accepted_hits.sam

These sorted files are then used to find count values through htseq-count-

htseq-count –s no –m intersection-strict –t exon –i gene_id sorted_accepted_hits.sam > count.txt

All these counts are merged in a single text file. Also while preparing count table it should be keep in mind that each column represent one sample and in case of biological or technical replicates their count value is summed in a single column.

Following commands are typed on R in order to load DESeq library and read count table.
library ("DESeq")

>countTable<-read.delim ("lncRNA_counts.txt", header=TRUE, row.names=1)

> head (countTable)

This command will provide snapshot of first few lines of countTable like the one shown below-

	GV_P	**GV**	**GII**	**GII_P**	**SCHIZONT**	**LT_p**	**LT**
XLOC_000001	12	1	2	14	115	2	12
XLOC_000002	2	9	15	15	114	1035	91
XLOC_000003	52	40	13	37	33	28	22
XLOC_000004	62	51	5	3	31	5	21
XLOC_000005	29	48	16	10	15	8	9
XLOC_000006	97	168	38	50	28	40	46

After reading count file and loading DESeq library in R package we need to normalize the counts in each colmn but before that description of table is given using pasillaDesign function, in which each column indicate different type of information and each row indicates different sample like-

> pasillaDesign = data.frame(row.names = colnames(countTable), condition = c("GV_p", "GV", "GII", "GII_p", "Scizont_p", "LT_p", "LT"), libType = c("paired-end","single-end", "single-end", "paired-end", +"paired-end", "paired-end", "single-end"))

>pasillaDesign

This will give following description of data-

condition	**libType**
GV_P	GV_p paired-end
GV	GV single-end
GII	GII single-end
GII_P	GII_p paired-end
SCHIZONT	Scizont_p paired-end
LT_p	LT_p paired-end
LT	LT single-end

The count is normalized by first estimating the size factor that gives the information about effective library size of particular column/sample. The function estimateSizeFactors is used to calculate effective library size for each sample. This size factor is then used to normalize count data by dividing each column with its size factor.

>cdsFull = newCountDataSet(countTable, pasillaDesign)

>cdsFull = estimateSizeFactors(cdsFull)

>sizeFactors(cdsFull)

This will calculate and output size factor for each column like described below-

GV_P	GV	GII	GII_P	SCHIZONT	LT_p	LT
1.1277527	2.1747970	0.5476734	0.5664845	1.2973385	0.8335067	1.1972943

The function estimateDispersions assign dispersion value to each gene by performing three steps, i.e., it first calculate dispersion value for each gene followed by fitting a curve through estimated dispersion and finally assign dispersion value to each gene by choosing between estimated value and the fitted value. Following commands accomplish this purpose-

>cdsFullBlind = estimateDispersions(cdsFull, method = "blind", fitType="local")

>vsdFull = varianceStabilizingTransformation(cdsFullBlind)

After normalization following commands are executed in order to perform statistical analysis on the normalized data. Following command will aid in drawing heatmap of variance stabilized transformed data, heatmap of untransformed data, sample to sample distance plot and principle component analysis plots respectively.

>library("RColorBrewer")

>library("gplots")

> select = order(rowMeans(counts(cdsFull)), decreasing=TRUE)[1:425]

>hmcol = colorRampPalette(brewer.pal(9, "GnBu"))(100)

>heatmap.2(exprs(vsdFull)[select,], col = hmcol, trace="none", margin=c(10, 6))

>heatmap.2(counts(cdsFull)[select,], col = hmcol, trace="none", margin=c(10,6))

>dists = dist(t(exprs(vsdFull)))

>mat = as.matrix(dists)

>rownames(mat) = colnames(mat) = with(pData(cdsFullBlind), paste(condition, libType, sep=" : "))

>heatmap.2(mat, trace="none", col = rev(hmcol), margin=c(13, 13))

>print(plotPCA(vsdFull, intgroup=c("condition", "libType"))

5. RESULTS

5.1 Dataset

RNA-Seq data for *P. falciparum* 3D7 is downloaded from SRA: DNAnexussubmitted by two different groups i.e. University of California- San Francisco and NIHusing Illumina Genome Analyzer II. This data contain both single and paired end runs. Their accession number, read length and kind of run information are provided in table 4.

NIH				University of California			
Single end reads	**Read Length**	**Paired end reads**	**Read Length**	**Single end reads**	**Read Length**	**Paired end reads**	**Read Length**
SRR364834	51	SRR364836	90	SRR066576	48	SRR066581	130
SRR364838	36	SRR364841	90	SRR066577	42	SRR066582	130
SRR364840	36	SRR364842	90	SRR066578	42	SRR066583	130
SRR364843	36	SRR364846	90	SRR066579	42	SRR066584	84
SRR364847	35			SRR066580	42	SRR066585	84
SRR364848	35					SRR066586	84
SRR364849	35					SRR066587	84
						SRR066588	84
						SRR066589	84
						SRR066590	84

Table 4: Dataset information downloaded from SRA: DNAnexus

5.2Trimming of low quality reads

We trimmed all reads in our dataset using DynamicTrim.pl script of SolexaQA software package with phredcutoff value taken as '20'. The trimmed file obtained after this is then sort according to length via LengthSort.pl script. Length is sort with length taken as '36'. LengthSort.pl output create output files, namely, *.discard file and *.single in case of single ended files. Reads smaller than length cutoff are stored in *.discard file and rest are stored in *.single in case of single ended reads. Out of which *.single file is used as input file in TopHat.And for paired ended files it take provide output on the basis of length cutoff in three files – *.paired1>, *.paired2> and

*.single file. Both *.paired1> and *.paired2> will be used as input fastq file while running TopHat.

LengthSort at length parameter as 36 leads to elimination of three single ended reads (SRR364847, SRR364848 and SRR364849) from our dataset as their *.single files obtained were empty because all reads are less than '36' for these RNA-Seq reads.

5.3 Mapping short RNA-Seq reads to reference genome

TopHat align reads onto the reference genome and provide output in bam format. It's a binary format for representing sequence data. Two types of bam files are created, one is accepted_hits.bam that contains reads aligned by TopHat and other is unmapped.bam which contains reads discarded by TopHat. accepted_hits.bam file is converted into sam format using samtools which is then used for calculating percentage of reads mapped by TopHat (Table 5).

Mapping percentage for each gene is reported in table 5 along with information about the stage to which read belong, number of reads mapped by TopHat and total number of reads (calculated by summing number of reads accepted and number of reads discarded by TopHat).

DATASET	STAGE	TOTAL READS	READS MAPPED BY TopHat	MAPPING PERCENTAGE
SRR364834	Ookinete	503005	499625	99.77391265
SRR364838	Gametophyte V	3035261	2814723	92.32543071
SRR364836	Gametophyte V	4602645	4492291	97.6333912
SRR364840	Gametophyte II	2178983	1975785	74.42193216
SRR364841	Gametophyte II	2210220	2152856	97.4631625
SRR066578	Late Trophozoite	1918333	1538581	80.9451433
SRR066579	Late Trophozoite	1317576	1106803	85.1147292
SRR364846	Late Trophozoite	3356106	3171726	94.5274545
SRR066580	Late Trophozoite	999677	854569	86.9269212
SRR066581	Ring	498718	371151	75.7740199
SRR066588	Ring	289957	48262	94.7372554
SRR066589	Ring	127350	41696	83.611061
SRR066590	Ring	37912	41012	83.2629527
SRR066582	Trophozoite	128850	81657	68.7997102
SRR066583	Trophozoite	114456	75579	72.0912265
SRR066584	Trophozoite	60343	46262	93.6231356
SRR066585	Trophozoite	25094	15884	75.3046034
SRR066586	Trophozoite	19495	15814	72.564585
SRR364842	Schizont	3498209	3282585	93.851184
SRR066576	Schizont	665555	518540	79.2302803
Total		25314149	23145401	

Table 5: Percentage of reads mapped by TopHat for each RNA-Seq run

5.4 Data Filtering

Data is again filtered at this step, reads having mapping percentage less than 60% are discarded along with reads that have mapped reads not in range of millions (i.e. have very few reads left after trimming). This filtering leads to elimination of fifteen more reads and we left with only eigth RNA-Seq reads for further analysis which are listed in another table along with one extra column identifier in Cuffdiff output for each read (Table 5). Also we have started our study with RNA-Seq data from seven stages, this elimination limit our study to only four stages, namely, schizont, late trophozoite, gametocyte II and gametocyte V.

Identifier in Cuffdiff	DATASET	STAGE	TOTAL READS	READS MAPPED BY TopHat	MAPPING PERCENTAGE
q1_0	SRR364838	Gametophyte V	10,098,615	9,323,058	92.32543
q1_1	SRR364836	Gametophyte V	10,098,615	9,678,078	95.88614
q2_0	SRR364840	Gametophyte II	10,641,921	7,907,046	74.42193
q2_1	SRR364841	Gametophyte II	4,665,634	4,495,563	96.43096
q3	SRR364842	Schizont	7,180,224	6,592,928	91.84885
q5_0	SRR066578	Late Trophozoite	5,145,942	3,032,308	60.00336
q5_1	SRR066579	Late Trophozoite	5,935,701	3,503,851	60.01743
q4	SRR364846	Late Trophozoite	6,964,029	6,240,092	89.64776

Table 6: Final dataset information obtained after filtering of mapped data

5.5 Transcriptome Reconstruction and RABT assembly

Cufflink is used for constructing the whole transcriptome map of sample. It employs a genome guided approach for assembly of overlapping aligned fragments.Cufflinks will take accepted_hits.sam files of eight reads selected as input file and report assembled transcript information in separate transcripts.gtf file.

Allcufflink assemblies are then merged together using reference annotation based transcripts (RABT) assembly method. This will create a new folder ‘merged_asm’. The most important file

is merged.gtf which will contain merged information of all transcripts.gtf along with already annotated transcripts information. Cufflink assigns unique cuff_id to each un-annotated transcript. Total 7,015 transcripts are obtained for these mapped reads. Of which 2,317 transcripts were already annotated(Figure 18).

We have transcripts information now we want to calculate their coding potential for which we need sequence of these files in fasta format. For this purpose we used gffread which need index files of genome and merged.gtf in which information about transcripts are there. gffread provide fasta sequence of 6,997 transcripts only and 18 transcripts sequence can't be retrieved by gffread. Therefore we check whether these 18 transcripts are protein coding or not and we found that these 18 transcripts are already annotated as protein coding genes. Because our study aim is to detect non coding transcripts, we do not need these 18 transcripts and we discard them.

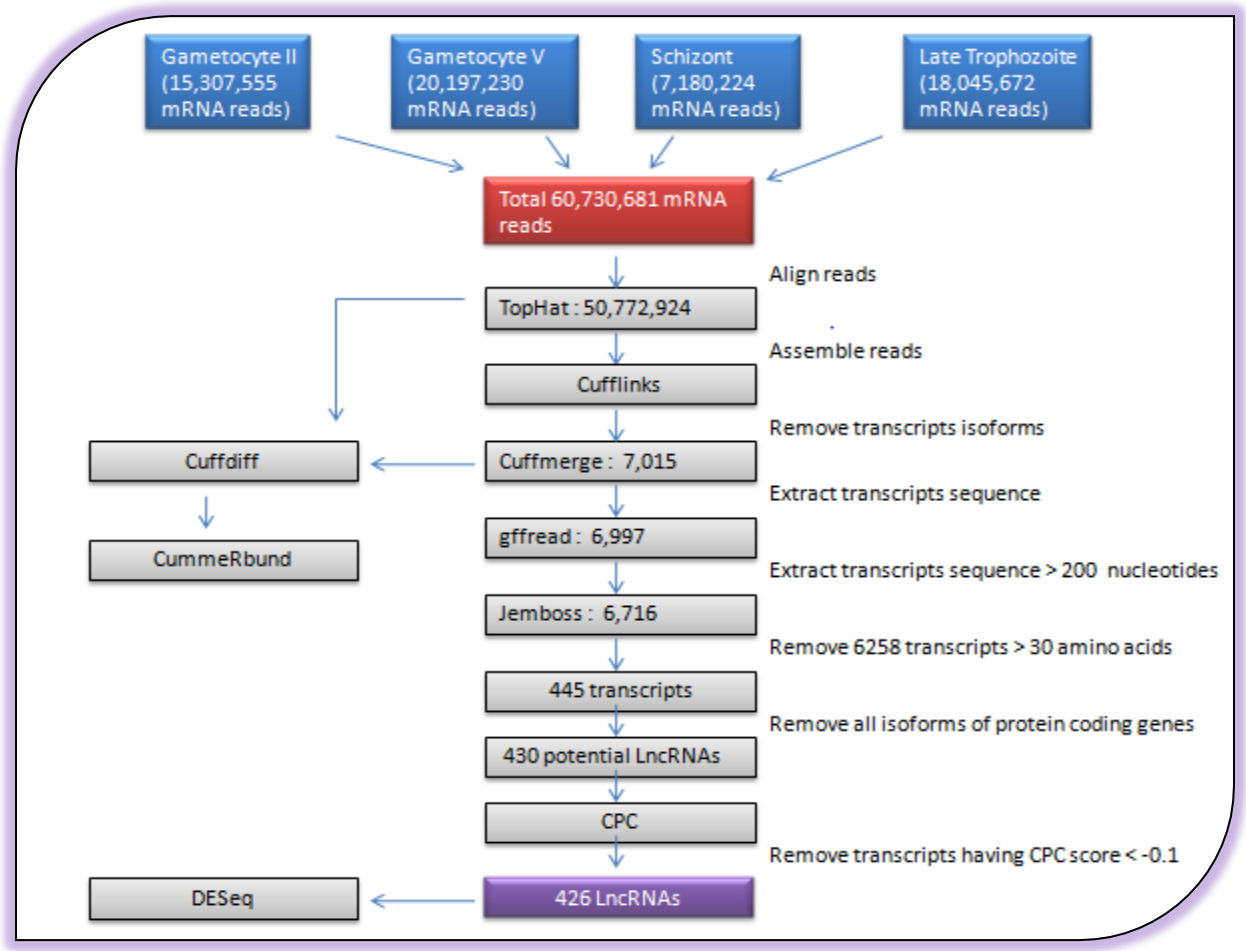


Figure 18: Flowchart representing whole RNA-Seq pipeline adopted for this study

5.6 Estimating coding potential of transcripts

We used two softwares to calculate coding potential of transcripts. But before that transcript having length more than or equal to 200 nucleotides are extracted and rest are discarded using three perl script which discard 281 transcripts. 6716 transcripts were greater than 200 nucleotides; these transcript's seqences are used as input to getorf software of jemboss suite. Jemboss provide coding potential estimates for 6703 transcripts ids rest 13 sequences are not used by jemboss because these sequences have lots of 'N', so jemboss does not predict result for these sequences.

Jemboss output is then parsed to remove all transcripts greater than 30 amino acids in length because they are likely to be coding transcripts. This is also done using script mentioned in Appendix IV. 445 transcripts are selected which have length less than 30 amino acids and rest are discarded. These transcripts are further checked to remove all isoforms of discarded transcripts and other annotated transcripts. This is accomplished by using two perl script back to back (see Appendix V). First perl script is used to extract gene ids of 6258 discarded transcript ids and output 3624 xloc idswhich are then used as input to next perl script to extract transcripts id encoded by these xloc ids. This gives us transcripts ids of 6270 transcripts. These 6270 transcript ids then compared with 445 TCON ids (transcript ids) to remove common ids i.e. 10 common ids because these 10 transcripts are isoforms of protein coding transcripts of transcripts.These 435 transcriptsids obtained are then compared with transcript ids of 2317 annotated transcripts in merged.gtf file to remove 5 common ids and we finally get 430 unannotated noncoding transcript ids. Final 430 transcripts ids obtained after all this parsing steps are then used to extract their sequence which is then used as input to software, coding potential calculator. CPC output is then again parsed to get transcripts ids having score less than -0.1. Finally, 426 transcripts ids obtained after parsing CPC output are considered as the lncRNA ids.

5.7 Differential expression analysis of lncRNAs

Two softwares are used for differential expression analysis, namely, Cuffdiff and DESeq.

5.7.1 Differential expression analysis using Cuffdiff and CummeRbund

Cuffdiff is used for differential expression analysis immediately after cuffmerge which report outputs in number of files which are then analysed using CummeRbund. But before that we manually plot boxplot using Microsoft Excel for differential expression result using isoform.fpkm_tracking file (Figure 19).

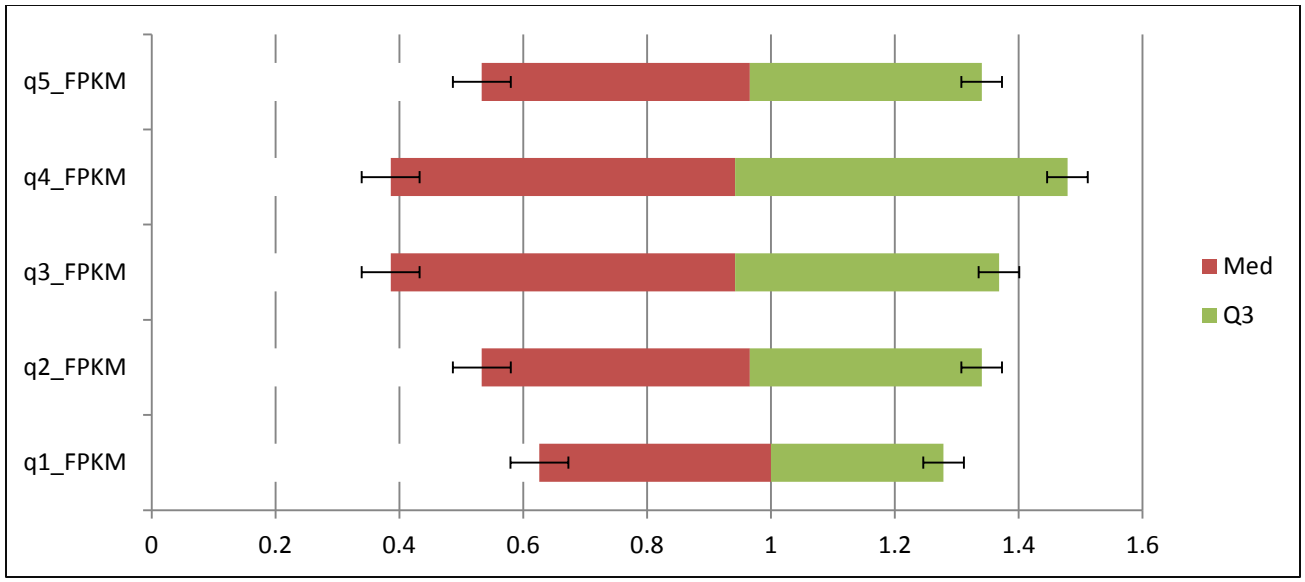


Figure 19: Box plotof five samples

Here q1, q2, q3, q4 and q5 belong to samples from stage gametophyte V, gametophyte II, schizont, late trophozoite (paired-ended reads) and late trophozoite (single-ended reads)respectively.

Differential expression of lncRNA is plotted to draw a venn diagram using Venny. Figure 20 represents expression pattern of 426 lncRNAs in four stages, i.e., schizont, late trophozoite, gametocyte II and gametocyte V.

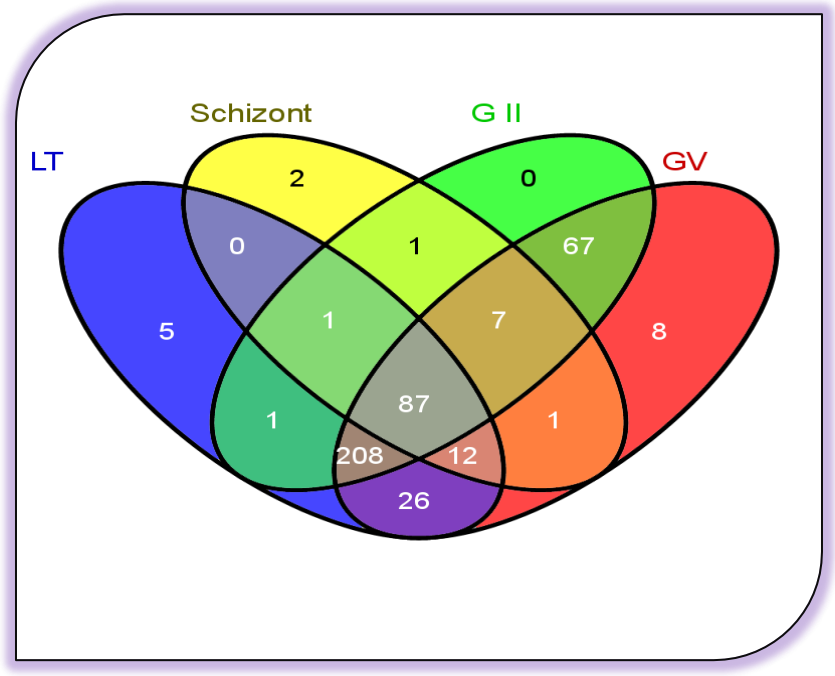


Figure 20: Venn diagram showing differential expression of transcripts in four stages

CummeRbund is then used for statistical analysis and plotting. Various methods are available in this package for plotting dispersion, density (with or without replicates), box plot (with or without replicates), scatter plots, dendogram and volcano plots. Figures of box plots generated using CummeRbund and density plot for replicates are provided in supplementary data.

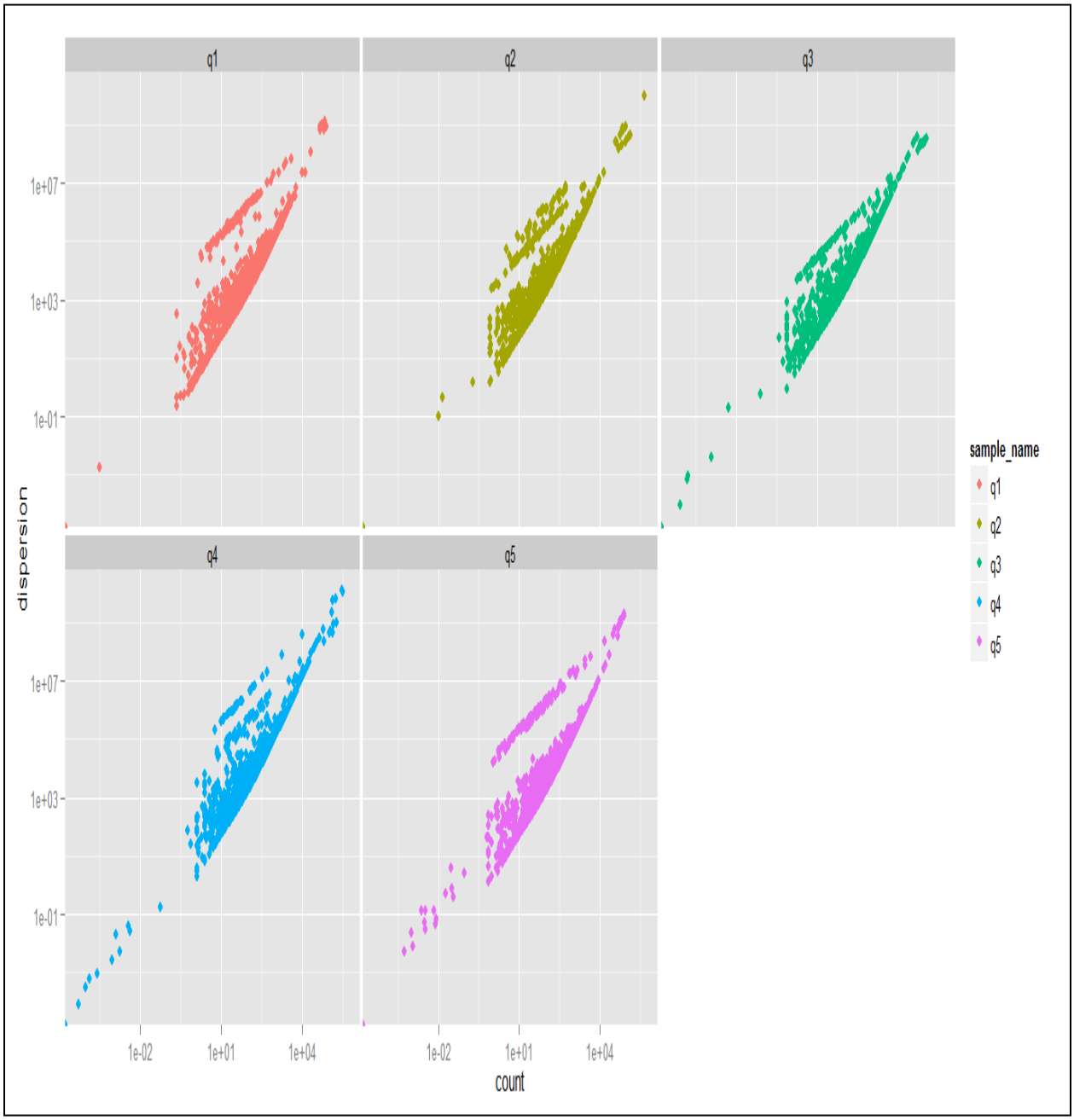


Figure 21: Count vs dispersion plot by condition for all genes

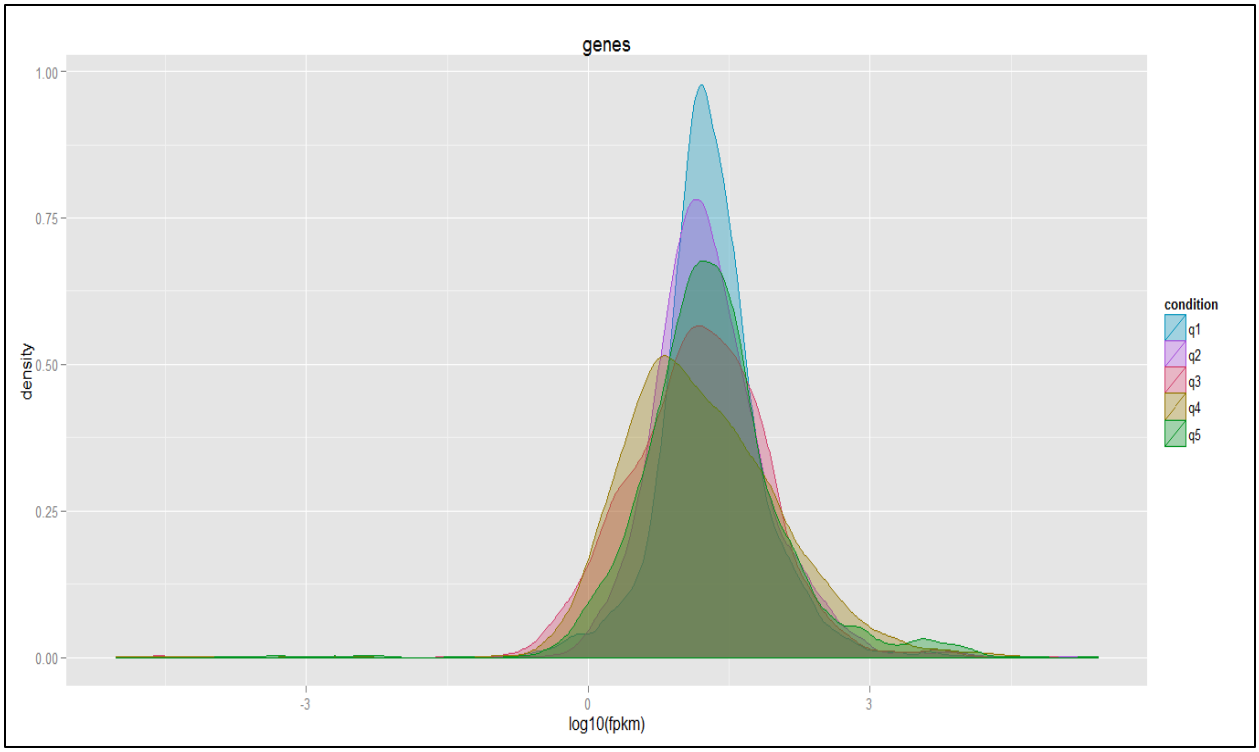


Figure 22: Density plot of individual conditions

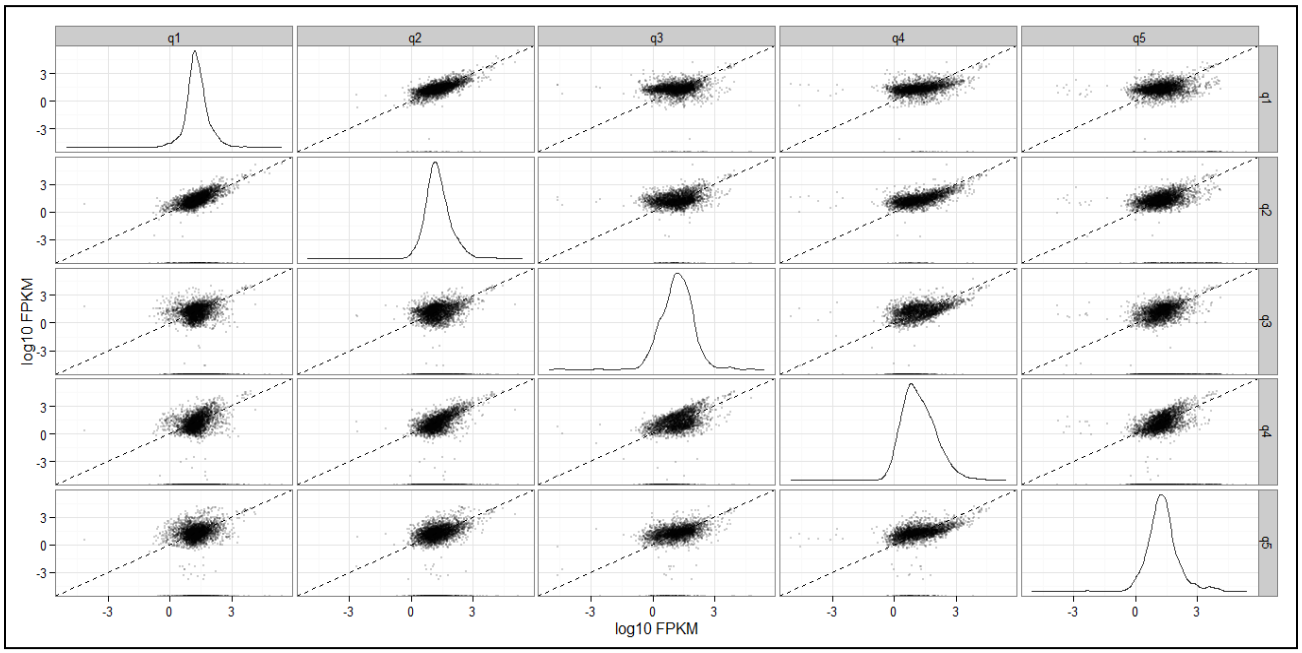


Figure 23: Scatterplots
Useful in reporting global changes and trends in gene expression between pairs of conditions.

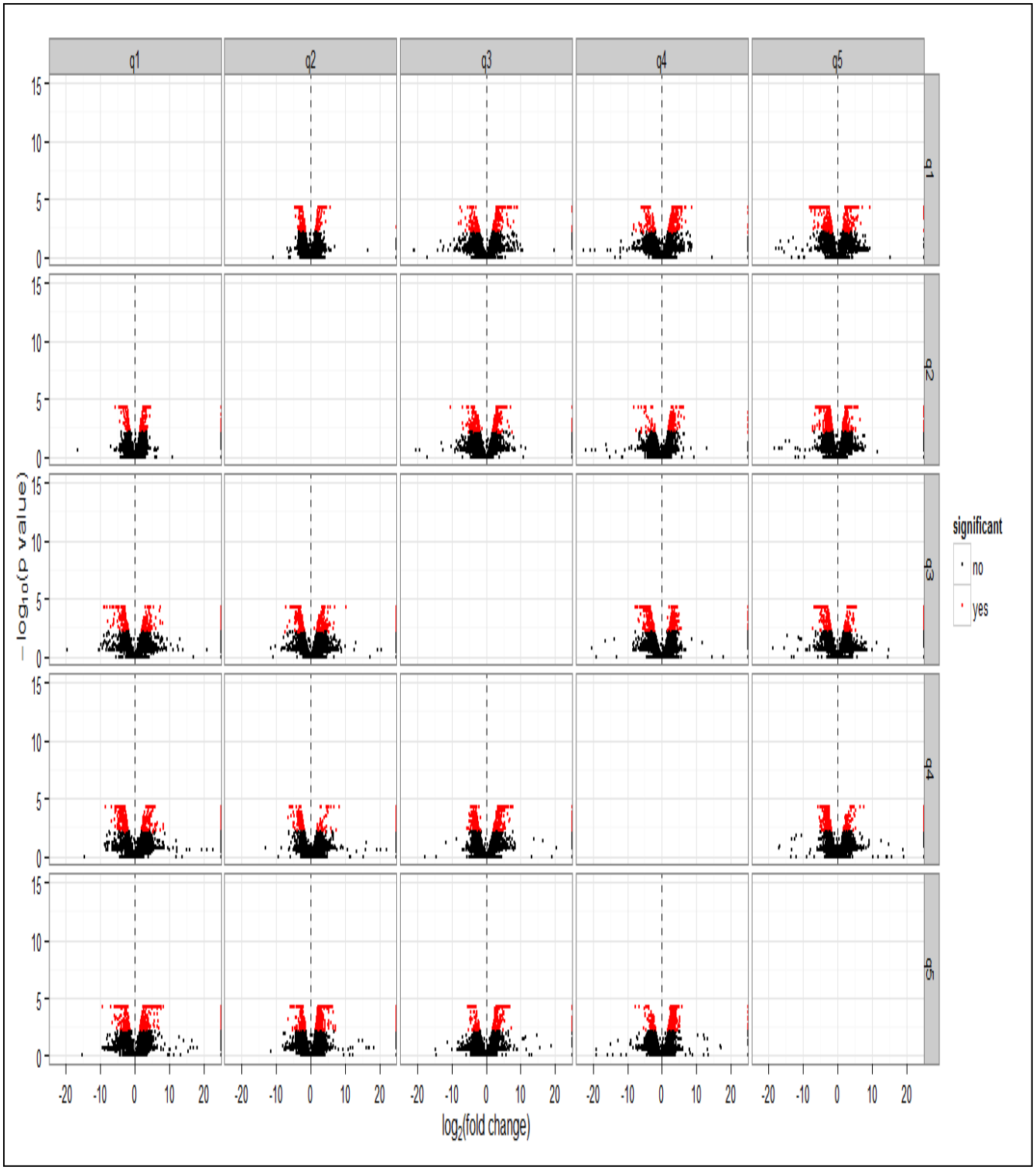


Figure 24: Volcano plots
Help in exploring the relationship between fold-change and significance.

5.7.2 Differential expression analysis using DESeq

DeSeq is also an R based package designed for differential expression analysis. It takes count table as input that contains the count values of each gene in different samples. Counts for each sample are obtained by using htseq-count script which takes sorted sam files and lncRNA.gtf file as input. lncRNA.gtf file is created by extracting final lncRNAs information from merged.gtf file obtained after RABT assembly. All counts file obtained by htseq-count script is then merged in a single text file. Also while preparing count table it should be keep in mind that each column represent one sample and in case of biological or technical replicates their count value is summed in a single column.

The count is normalized by first estimating the size factor that gives the information about effective library size of particular column/sample. The function **estimateSizeFactors** is used to calculate effective library size for each sample. This size factor is then used to normalize count data by dividing each column with its size factor.

The function **estimateDispersions** assign dispersion value to each gene by performing three steps, i.e., it first calculate dispersion value for each gene followed by fitting a curve through estimated dispersion and finally assign dispersion value to each gene by choosing between estimated value and the fitted value. After normalization commands for drawing heat maps are typed that will provide information about expression level of lncRNAs in different samples/stages.

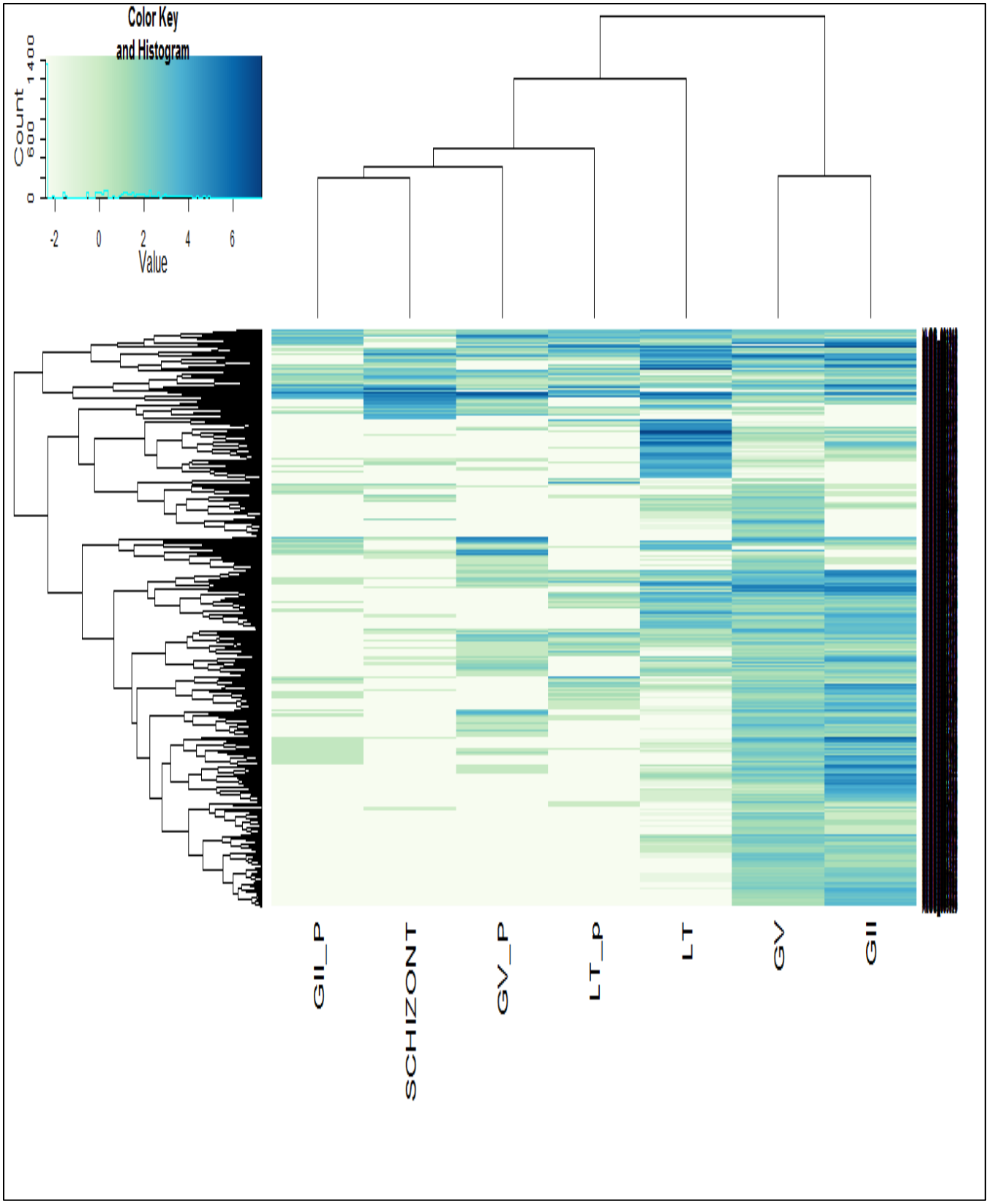


Figure 25: Heatmaps showing the expression data of lncRNAs

Another use of variance stabilized data is sample clustering of sample according to distance. DESeq calculate sample to sample distance by applyingEuclidean distances as calculated from the variance stabilizing transformation of the count data.The clustering reflects that samples belong to same library or same stages are very similar.

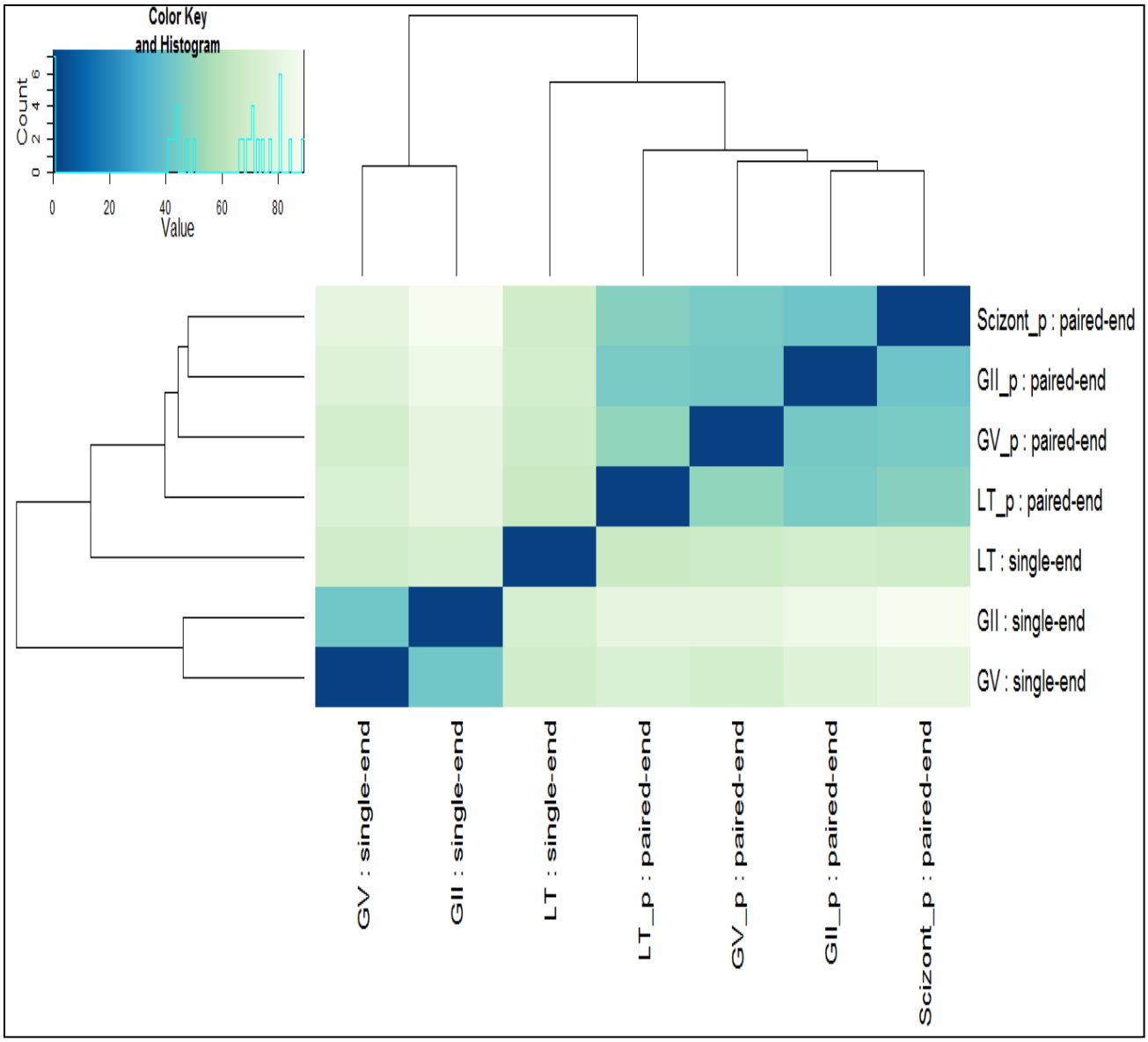


Figure 26: Heatmap showing the Euclidean distances between the samples
Calculated from the variance stabilizing transformation of the count data.

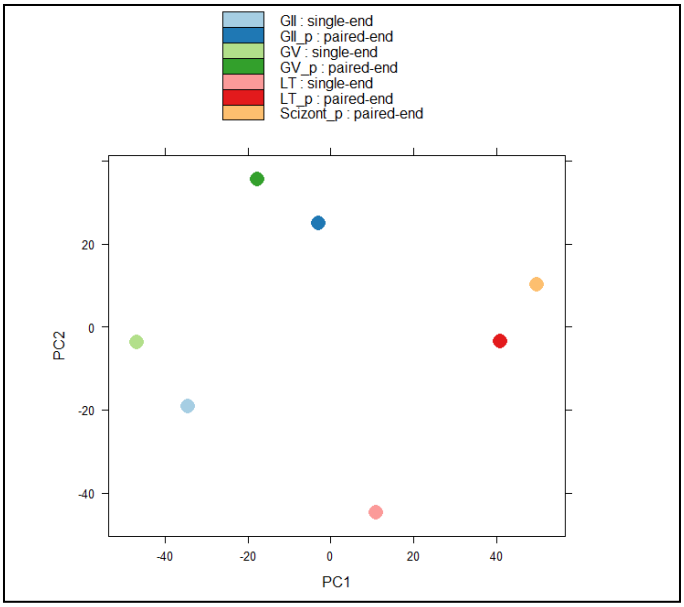


Figure 27: PCA plot

6.DISCUSSION

Plasmodium falciparum's capability of evading host immune system through expression of highly variant erythrocyte surface protein and their cytoadherence capability aiding in escaping clearance by spleen are the issue of concern limiting our desire to get rid of this disease. It has been well established fact that ncRNAs are crucial players of various regulatory pathways involving epigenetic modification, transcription, post transcriptional and translation regulation etc. *Plasmodium falciparum* genome's analysis reveal presence of large number of ncRNAs and various RNA binding proteins and lack of various gene regulatory protein, RNA interference and DNA methylation machinery. This finding reveals the potential role of ncRNAs in integral biology of this parasite. Lots of research in identification of ncRNAs reveals presence of various TAREs ncRNA function as histone chaperones, small ncRNAs and also two *var* genes associated long ncRNA that may have role in mono allelic expression of *var* gene cluster. Still large portion of long ncRNAs are unexplored which inspire us to design our study to reveal these ncRNAs in order to have better understanding of gene regulation networks contributing in virulence of this parasite.

RNA-Seq reads for *P. falciparum* 3D7 are taken and align to latest *Plasmodium falciparum* genome release from Sanger Institute using spliced read aligner TopHat, followed by transcriptome reconstruction using cufflinks. We obtained total 7015 transcripts through our assembly process. These transcripts are then filtered in order to extract non coding transcripts using Jemboss and CPC, both of which together predict 426 lncRNAs in this parasite. These 426 lncRNAs do not include any TARE ncRNAs because the RNA-Seq reads with which we have started this study were synthesized using poly A tail which are lacking in TARE ncRNAs. Then we decided to check differential expression of these 426 lncRNAs in four crucial life cycle stages of this parasite using cuffdiff and DESeq. Differential expression analysis reveals expression of 111, 372, 416 and 340 lncRNAs in schizont, gametocyte II, gametocyte V and late trophozoite stage respectively. Out of which 5, 2 and 8 ncRNA are uniquely expressed in late trophozoite, schizont and gametocyte V stage respectively. While in gametocyte II none of the lncRNA show unique expression. 87 lncRNAs are expressed in all four stages of infection. Since gametocyte II stage has zero uniquely expressed genes, we correlate its lncRNAs with gametocyte V stage and found all its lncRNAs are found to express in gametocyte V stage except 3 lncRNAs, one of which is expressed in schizont and two expressed in late trophozoite stage. In this way we categorize all lncRNA according to their stage specific expression which will provide us insight into their role in virulence at particular stage, also enlighten the potential RNA which should be targeted in order to halter parasite's life cycle.

7. CONCLUSIONAND FUTURE PERSPECTIVE

Plasmodium falciparum, cause of most deadly form of malaria, is endangering life of millions of people annually. Parasite is highly capable of evading human immune system and developed resistance to many antimalarial drugs. Therefore we design our study to reveal these ncRNAs in order to have better understanding of gene regulation networks contributing in virulence of this parasite. ThroughRNA-Seq read alignment to *P. falciparum* 3D7 genome we obtained 7015 transcripts, these transcripts on filtering provide 426 transcripts which do not have coding potential and are part of *P. falciparum* non coding transcriptome. Differential expression estimation of these 426 lncRNA in four stages i.e. schizont, late trophozoite, gametocyte V and gametocyte II reveals expression of total 111, 340, 416 and 372 lncRNAs respectively. Unique expression of lncRNAs is checked in order to find lncRNA that are crucial in that stage and we find 5, 2 and 8 ncRNA showing unique expression in late trophozoite, schizont and gametocyte V stage respectively.These lncRNAs can be categorized according to their stage specific expression and this might provide us insight into their role at particular stages.The potentiallncRNA can be used as targetsfor designing antimalarial drug.

Also by taking advantage of high throughput sequencing technologies same analysis on hyper-virulent clinical isolates, drug resistant strains and hypo-virulent parasites (due to some mis-regulation in virulence genes) may provide best solution to eliminate this disease from root.

8. REFERENCES

Broadbent, KM; Park**,** D;Wolf,AR;Tyne**,** DV; Sims, JS; Ribacke, U; Volkman, S;Duraisingh, M; Wirth, D; Sabeti, PC and Rinn, JL(2011).Aglobal transcriptional analysis of *Plasmodium falciparum* malaria reveals a novel family of telomere-associated lncRNAs.Genome Biology.*12*, R56.

Centonze, D; Rossi, S; Napoli, I; Mercaldo, V; Lacoux, C; Ferrari, F; Ciotti, MT; Chiara, VD; Prosperetti, C; Maccarrone, M; Fezza, F; Calabresi, P; Bernardi, G and Bagni, C (2007). The brain cytoplasmic RNA BC1 regulates dopamine D2 receptor-mediated transmission in the striatum. The Journal of Neuroscience. *27*, 8885–92.

Cheetham, SW; Gruhl, F; Mattick, JS. and Dinger, ME (2013).Long ncRNAs and the genetics of cancer. British Journal of Cancer.*233*, 1-7.

Chen, K & Rajewsky, N (2007).The evolution of gene regulation by transcription factors and microRNAs. Nature Reviews Genetics. *8*, 93–103.

Clark, IA; Budd, AC; Alleva1, LM and Cowden, WB (2006).Human malarial disease: a consequence of inflammatory cytokine release. Malaria Journal.*5*, 85.

Clark, IA;Alleva, LM; Mills, AC and Cowden, WB (2004).Disease pathogenesis in malaria and clinically similar conditions.Clin Microbiol Rev. *17*,509-539.

Cox, MP; Peterson, DA; Biggs, PJ. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics.*11*, 485.

Davy, CP; Sicuri, E; Ome, M; Lawrence-Wood, E; Siba, P; Warvi, G; Mueller, I and Conteh, L (2010). Seeking treatment for symptomatic malaria in Papua New Guinea.Malaria Journal.*9*, 268.

Epp, C; Li, F; Howitt, CA; Chookajorn, T and Deitsch, KW (2009).Chromatin associated sense and antisense ncRNAs are transcribed from the *var* gene family of virulence genes of the malaria parasite *Plasmodium falciparum.* RNA.*15*, 116-127.

Feng, J; Bi, C; Clark, BS; Mady, R; Shah, P and Kohtz, JD (2006). The Evf-2 ncRNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev. *20*, 1470–1484.

Florens, L; Washburn, MP; Raine, JD; Anthony, RM; Grainger, M; Haynes, JD; Moch, JK; Muster, N; Sacci, JB; Tabb, DL; Witney, AA; Wolters, D; Wu, Y; Gardner, MJ; Holder, AA; Sinden, RE; Yates, JR & Carucci, DJ (2002). A proteomic view of the *Plasmodium falciparum* life cycle.Nature.*419*, 520-526.

Garber, M; Grabherr, MG; Guttman, M and Trapnell, C (2011).Computational methods for transcriptome annotation and quantification using RNA-Seq. Nature Methods**.***8*, 469-477.
Gibb, EA; Brown, CJ and Lam, WL (2011).The functional role of long ncRNA in human carcinomas.Molecular Cancer.*10*, 38.

Grabherr, MG; Haas, BJ; Yassour, M; Levin, JZ; Thompson, DA; Amit, I; Adiconis, X; Fan, L; Raychowdhury, R; Zeng, Q; Chen, Z; Mauceli, E; Hacohen, N; Gnirke, A; Rhind, N; Palma, FD; Birren, BW; Nusbaum, C; Lindblad-Toh, K; Friedman, N and Regev, A (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome.Nature Biotechnology.*29*, 644-652.

Kong, L; Zhang, Y; Ye, ZQ; Liu, XQ; Zhao, SQ; Wei, L and Gao, G (2007). CPC: assess the protein-coding potential of transcripts using sequences features and support vector machine. Nucleic Acids Research.*35*, 345-349.

Langmead, B; Trapnell, C; Pop, M and Salzberg, SL (2009).Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.GenomeBiology.*10*, R25.

Mercer, TR; Dinger, ME and Mattick, JS (2009).Long ncRNAs: Insights into functions. Nature Reviews. *10*, 155-159.

Mercer, TR; Mattick, JS (2013). Structure and function of long ncRNAs in epigenetic regulation. Nature Structural & Molecular Biology. *20*,300–307.

Metzker, ML (2005). Emerging technologies in DNA sequencing.Genome Res.*15*, 1767–1776.

Metzker, ML (2010). Sequencing technologies -the next generation. Nature Reviews. Genetics.*11*, 31-46.

Ndyomugyenyi, R; Magnussen, P; and Clarke, S (2007).Diagnosis and treatment of malaria in peripheral health facilities in Uganda: findings from an area of low transmission in south-western Uganda.Malaria Journal.*6*, 39.

Nie, L; Wu, HJ; Hsu, JM; Chang, SS; LaBaff, AM; Li, CW; Wang, Y; Hsu, JL; Hung, MC (2012). Long ncRNAs: versatile master regulators of gene expression and crucial players in cancer. Am J Transl Res. *4*, 127-150.

Ponting, CP; Oliver, PL and Reik, W (2009).Evolution and Functions of Long NcRNAs.Cell.*136,* 629–641.

Reis-Filho, JS (2009).Next-generation sequencing.Breast Cancer Research**.***11*, S3.

Roberts, A; Pimentel, H; Trapnell, C and Pachter, L (2011).Identification of novel transcripts in annotated genomes.Bioinformatics. *27*, 2325-2329.

Sana, J; Faltejskova, P; Svoboda, M and Slaby, O (2012).Novel classes of ncRNAs and cancer**.**Journal of Translational Medicine.*10,* 103.

Scherf, A; Lopez-Rubio, JJ and Riviere, L (2008).Antigenic variation in *Plasmodium falciparum.*Annu. Rev. Microbiol. *62*, 445-470.

Shendure, J and Ji, H (2008). Next-generation DNA sequencing. Nature Biotechnology.*26*, 1135-1145.

Sierra-Miranda, M; Delgadillo, DM; Mancio-Silva, L; Vargas, M; Villegas-Sepulveda, N; Martínez-Calvillo, S; Scherf, A; Hernandez-Rivas, R (2012). Two long ncRNAs generated from subtelomeric regions accumulate in a novel perinuclear compartment in *Plasmodium falciparum*. Molecular and Biochemical Parasitology.*185*, 36-47.

Sundquist, A; Ronaghi, M; Tang, H; Pevzner, P. and Batzoglou, S (2007).Whole-genome sequencing and assembly with high-throughput, short-read technologies.PLoS ONE.*2*, e484.
Tarimo, DS; Minjas, JN; Bygbjerg, IC (2001).Malaria diagnosis and treatment under the strategy of the integrated management of childhood illness (IMCI): relevance of laboratory support from the rapid immunochromatographic tests of ICT Malaria P.f/P.v and OptiMal. Annals of Tropical Medicine and Parasitology.*95*, 437-444.

Trapnell, C; Pachter, L and Salzberg, SL (2009).TopHat: discovering splice junctions with RNA-Seq.Bioinformatics. *25*, 1105-1111.

Trapnell, C; Williams, BA; Pertea, G; Mortazavi, A; Kwan, G; Baren, MJV; Salzberg, SL; Wold, BJ and Pachter, L(2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology.*28*, 511–515.

Trapnell, C; Roberts, A; Goff, L; Pertea, G; Kim, D; Kelley, DR; Pimentel, H; Salzberg, SL; Rinn, JL and Pachter, L(2012). Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks.Nature.*7*, 562–578.

Trapnell, C; Hendrickson, DG; Sauvageau, M; Goff, L; Rinn, JL & Pachter, L (2013).Differential analysis of gene regulation at transcript resolution with RNA-Seq. Nature Biotechnology.*31*, 46-53.

Wang, X; Arai, S; Song, X; \Reichart, D; Du, K; Pascual, G; Tempst, P; Rosenfeld, MG; Glass, CK & Kurokawa, R (2008). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature.*454*, 126-130.

Wu, TD and Watanabe, CK (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics.*21*, 1859–1875.

Yu, H; Lindsay, J; Feng, ZP; Frankenberg, S; Hu, Y; Carone, D; Shaw, G; Pask, AJ; O'Neill, R; Papenfuss, AT and Renfree, MB (2012). Evolution of coding and non-coding genes in HOX clusters of a marsupial. BMC Genomics.*13*,251.

9. APPENDIX

APPENDIX I

Table representing related experiment for studySRP009370 on *Plasmodium falciparum* 3D7 submitted by NIH

ACCESSION	TITLE	INSTRUMENT	STUDY	SUBMISSION
SRX105936	Sequence of the ring bidirectional libraries for the "Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*	Illumina Genome Analyzer	SRP009370	SRA048120
SRX105937	Sequence of the schizont bidirectional libraries for the "Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*	Illumina Genome Analyzer	SRP009370	SRA048120
SRX105938	Sequence of the ET bidirectional libraries for the "Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*	Illumina Genome Analyzer	SRP009370	SRA048120
SRX105939	Sequence of the LT bidirectional libraries for the "Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*	Illumina Genome Analyzer	SRP009370	SRA048120
SRX105940	Sequence of the ookinete bidirectional libraries for the "Directional gene expression and antisense transcripts in sexual and asexual stages of	Illumina Genome Analyzer II	SRP009370	SRA048120

	Plasmodium falciparum			
SRX105941	Sequence of the Gametocyte II bidirectional libraries for the "Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*	Illumina Genome Analyzer	SRP009370	SRA048120
SRX105942	Sequence of the GV bidirectional libraries for the "Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*	Illumina Genome Analyzer	SRP009370	SRA048120
SRX106027	Sequence of the GII strand specific library for the "Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*	Illumina Genome Analyzer II	SRP009370	SRA048120
SRX106028	Sequence of the GV strand specific library for the "Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*	Illumina Genome Analyzer II	SRP009370	SRA048120
SRX106029	Sequence of the schizont strand specific library for the "Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*	Illumina Genome Analyzer II	SRP009370	SRA048120
SRX106030	Sequence of the LT strand specific library for the "Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*	Illumina Genome Analyzer II	SRP009370	SRA048120

Table representing related samples for study SRP009370 submitted by NIH

ACCESSION	TITLE	ORGANISM	STUDIES	SUBMISSION
SRS271077	*P. falciparum* ring-stage (R)	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRS271078	*P. falciparum* early trophozoite stage (ET)	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRS271079	*P. falciparum* late trophozoite stage (LT)	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRS271085	*P. falciparum* schizont	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRS271086	*P. falciparum* gametocytes stage (GII)	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRS271089	*P. falciparum* gametocytes V stage (GV)	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRS271090	*P. falciparum* ookinete stage (Oo)	*Plasmodium falciparum* 3D7	SRP009370	SRA048120

Table representing related runs for study SRP009370 submitted by NIH

ACCESSION	EXPERIMENT	ORGANISM	STUDY	SUBMISSION
SRR364834	SRX105940	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRR364836	SRX106028	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRR364838	SRX105942	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRR364840	SRX105941	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRR364841	SRX106027	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRR364842	SRX106029	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRR364843	SRX105937	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRR364846	SRX106030	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRR364847	SRX105939	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRR364848	SRX105938	*Plasmodium falciparum* 3D7	SRP009370	SRA048120
SRR364849	SRX105936	*Plasmodium falciparum* 3D7	SRP009370	SRA048120

Table representing related experiments for study SRP003615 submitted by University of California – San Francisco

ACCESSION	TITLE	ORGANISM	INSTRUMENT	STUDY	SUBMISSION
SRX027155	RNA-Seq of TP4, 08/08/21 run	*Plasmodium falciparum* 3D7	Illumina Genome Analyzer II	SRP003615	SRA024324
SRX027201	RNA-Seq of TP3, 08/12/05 run	*Plasmodium falciparum* 3D7	Illumina Genome Analyzer II	SRP003615	SRA024324
SRX027202	RNA-Seq of TP1 and TP2, 09/04/20 run	*Plasmodium falciparum* 3D7	Illumina Genome Analyzer II	SRP003615	SRA024324
SRX027203	RNA-Seq of TP1 and TP2, 09/04/20 run	*Plasmodium falciparum* 3D7	Illumina Genome Analyzer II	SRP003615	SRA024324
SRX027205	RNA-Seq of TP1 and TP2, 10/02/03 run	*Plasmodium falciparum* 3D7	Illumina Genome Analyzer II	SRP003615	SRA024324

Table representing related samples for study SRP003615 submitted by University of California – San Francisco

ACCESSION	TITLE	ORGANISM(S)	STUDIES	SUBMISSION
SRS115098	TP1, 11hr post-invasion, ring stage, *Plasmodium falciparum* 3D7 Oxford	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRS115099	TP2, 22hrs post-invasion, trophozoites, *Plasmodium falciparum* 3D7 Oxford	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRS115100	TP3, 33hrs post-invasion, late trophozoite/early schizont, *Plasmodium falciparum* 3D7 Oxford	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRS115101	TP4, 44hrs post-invasion, late schizont *Plasmodium falciparum* 3D7 Oxford	*Plasmodium falciparum* 3D7	SRP003615	SRA024324

Table representing related runs for study SRP003615 submitted by University of California – San Francisco

ACCESSION	EXPERIMENT	ORGANISM	STUDY	SUBMISSION
SRR066576	SRX027155	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066577	SRX027201	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066578	SRX027201	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066579	SRX027201	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066580	SRX027201	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066581	SRX027203	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066582	SRX027205	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066583	SRX027205	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066584	SRX027202	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066585	SRX027202	*Plasmodium falciparum* 3D7	SRP003615	SRA024324

SRR066586	SRX027202	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066587	SRX027202	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066588	SRX027203	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066589	SRX027203	*Plasmodium falciparum* 3D7	SRP003615	SRA024324
SRR066590	SRX027203	*Plasmodium falciparum* 3D7	SRP003615	SRA024324

APPENDIX II

Commands used to run TopHat for mapping RNA-Seq reads onto reference genome –

tophat --coverage-search --microexon-search -r 120 -o <path to output directory><path to reference genome index><path to trimmed fastq file/SRR066581_1.fastq.trimmed.paired1><path to trimmed fastq file/SRR066581_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search -r 120 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066582_1.fastq.trimmed.paired1> <path to trimmed fastq file/SRR066582_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search -r 120 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066583_1.fastq.trimmed.paired1> <path to trimmed fastq file/SRR066583_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search -r 166 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066584_1.fastq.trimmed.paired1> <path to trimmed fastq file/SRR066584_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search -r 166 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066585_1.fastq.trimmed.paired1> <path to trimmed fastq file/SRR066585_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search -r 166 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066586_1.fastq.trimmed.paired1> <path to trimmed fastq file/SRR066586_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search -r 166 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066587_1.fastq.trimmed.paired1> <path to trimmed fastq file/SRR066587_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search -r 166 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066588_1.fastq.trimmed.paired1> <path to trimmed fastq file/SRR066588_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search -r 166 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066589_1.fastq.trimmed.paired1> <path to trimmed fastq file/SRR066589_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search -r 166 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066590_1.fastq.trimmed.paired1> <path to trimmed fastq file/SRR066590_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search --library-type fr-unstranded -r 160 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR364836_1.fastq.trimmed.paired1><path to trimmed fastq file/SRR364836_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search --library-type fr-unstranded -r 160 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR364841_1.fastq.trimmed.paired1><path to trimmed fastq file/SRR364841_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search --library-type fr-unstranded -r 160 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR364842_1.fastq.trimmed.paired1><path to trimmed fastq file/SRR364842_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search --library-type fr-unstranded -r 160 -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR364846_1.fastq.trimmed.paired1><path to trimmed fastq file/SRR364846_1.fastq.trimmed.paired2>

tophat --coverage-search --microexon-search -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR364834.fastq.trimmed.single

tophat --coverage-search --microexon-search -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR364838.fastq.trimmed.single

tophat --coverage-search --microexon-search -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR364840.fastq.trimmed.single

tophat --coverage-search --microexon-search -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066576.fastq.trimmed.single>

tophat --coverage-search --microexon-search -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066577.fastq.trimmed.single>

tophat --coverage-search --microexon-search -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066578.fastq.trimmed.single>
tophat --coverage-search --microexon-search -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066579.fastq.trimmed.single>

tophat --coverage-search --microexon-search -o <path to output directory><path to reference genome index> <path to trimmed fastq file/SRR066580.fastq.trimmed.single>

APPENDIX III

Perl scripts used for extracting transcripts having length > 200 are written below-

I.
```
#!/usr/bin/perl -w
use strict;
open (SEQFILE, "transcripts.fa");
@seq=<SEQFILE>;
closeSEQFILE;
@seq = split(" ", $_ );

for ($i=1;$i<scalar @seq;$i++)
        {

                print "$seq[0],\n";
        }

exit;
```

II.
```
#!/usr/bin/perl –w
use strict;
open (MYFILE, "script1_out");

while ($line = <MYFILE>)
{
   @a=split('\s', $line);

if (@a[2] > 200)
{
print $a[0] . "\n";
}

}
exit;
```

III.
```
#!/usr/bin/perl -w
use strict;
```

```
my $idsfile = "script2_out";
my $seqfile = "seq_extracted.fa";
my %ids  = ();

openMYFILE, $idsfile;
while (<MYFILE>)
{
chomp;
  $ids{$_} += 1;
}
closeMYFILE;

local $/ = "\n>";  # read by FASTA record

open FASTA, $seqfile or die $!;
while (<FASTA>) {
chomp;
my $seq = $_;
my ($id) = $seq =~ /^>*(\S+)/;  # parse ID as first word in FASTA header
if (exists($ids{$id}))
    {
print "$seq" ."\n";
    }
}
close FASTA;
exit;
```

APPENDIX IV

BioPerl script used for calculating length of ORF in transcripts from output of Jemboss in order to extract transcripts ids having ORF length < 30 amino acid:-

```
#!/usr/bin/perl -w
use Bio::SeqIO;

my $seqio = Bio::SeqIO->new(-file => "jemboss_out", '-format' => 'Fasta');
                                                        while(my $seq = $seqio->next_seq)
{
                                                        chomp $seq;
                                                        my $string = $seq->seq;
                                                        my $id = $seq->display_id;
                                                        @string_seq= split (",$string);
                                                         $length_seq= scalar @string_seq;

                                                        $id=~ /(.*)_(\d+)/;
                                                        $prefix=$1;
                                                        push @all_prefix, $prefix;
                                                                        push @id_list, $id;
                                                                push @seq_list, $string;
#print "$id\t$prefix\n";
                                                                push @length_list,
$length_seq;

#print "@id_list\n";
                                                         }
@same_id=";
@length_same_ids=";

        for ($i=0; $i< scalar @id_list ; $i++)
                {
print "\n";
                        if ($all_prefix[$i]=~ /^$all_prefix[$i+1]$/)
                        {
                                print "$id_list[$i]\t$length_list[$i]\t";
                         }
                        }
exit;
```

APPENDIX V

I. Perl script for extracting gene ids for given transcript ids from merged.gtf file

```
#!/usr/bin/perl -w

use strict;

open (MYFILE1,"<discaded_TCONS") or die "Cant open File1";
my @discarded=<MYFILE1>;

closeMYFILE1;

open (MYFILE2, "<tcons_xloc_uniq" ) or die "Can't open File2";
my @all =  <MYFILE2>;
closeMYFILE2;

for(my $i=0;$i<@discarded;$i++)
{
        chomp $discarded[$i];
        for(my $j=0;$j<@all;$j++)
        {
                chomp $all[$j];
                my @allsplit=split ('\s', $all[$j]);
                if ($discarded[$i]=~ /^$allsplit[0]$/)
                {

                        print "$allsplit[1]\n";
                }
}
}

exit;
```

II. Perl script for extracting transcript ids encoded by given gene ids from merged.gtf file

```
#!/usr/bin/perl -w

use strict;

open (MYFILE1,"<discarded_xloc.txt") or die "Cant open File1";
my @discarded=<MYFILE1>;

closeMYFILE1;

open (MYFILE2, "<anno_tcons_xloc_uniq.txt" ) or die "Cant open File2";
my @all =  <MYFILE2>;
closeMYFILE2;

for(my $i=0;$i<@discarded;$i++)
{
        chomp $discarded[$i];
        for(my $j=0;$j<@all;$j++)
        {
                chomp $all[$j];
                my @line=split ('\t', $all[$j]);
                if ($discarded[$i]=~ /^$line[1]$/)
                {

                        print "$line[0]\n";
                }
}
}

exit;
```

APPENDIX VI

Perl script for parsing CPC output in order to extract transcripts ids having score less than -0.1

```
#!/usr/bin/perl
open (MYFILE, "cpcout2.txt");
# @cpc=<MYFILE>;
while ($cpc_out = <MYFILE>)
{
   @score=split('\t', $cpc_out);

if (@score[2] < -0.1)
{
print $score[0] . "\n";
}

}
exit;
```

APPENDIX VII

Perl script for extracting information of lncRNA from merged.gtf file

```
#!/usr/bin/perl -w

use strict;

open (MYFILE1,"<cpc_parse_uniq") or die "Cant open File1";
my @lncRNA=<MYFILE1>;

closeMYFILE1;

open (MYFILE2, "<merged.gtf" ) or die "Cant open File2";
my @pos =  <MYFILE2>;
closeMYFILE2;

for(my $i=0;$i<@lncRNA;$i++)
{
        chomp $lncRNA[$i];
        for(my $j=0;$j< @pos;$j++)
        {
                chomp $pos[$j];
                my @pos_split=split ('\s', $pos[$j]);
                if ($lncRNA[$i]=~ /^$pos_split[11]$/)
                {

                        print "$pos[$j]\n";
                }
}
}

exit;
```

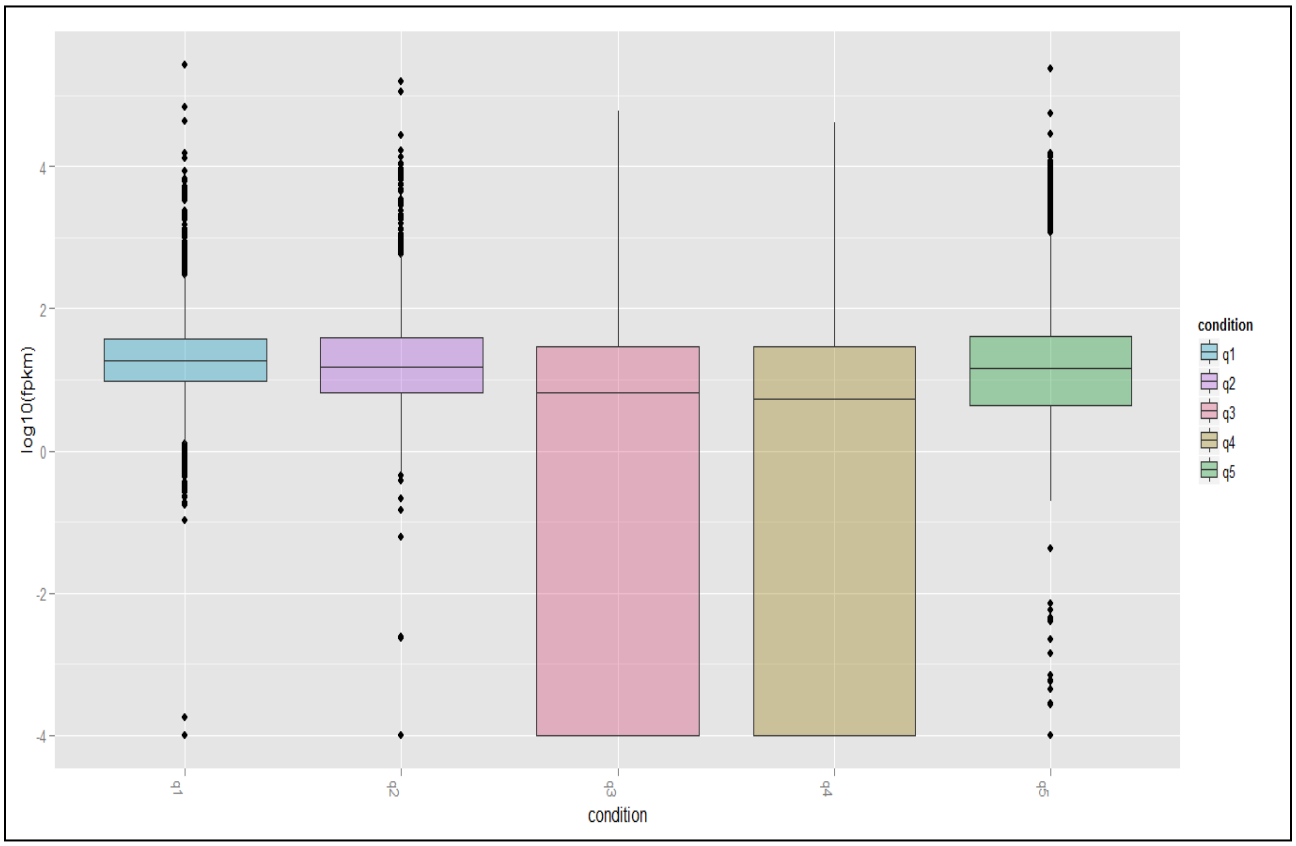
APPENDIX VIII

Perl script for extracting fasta sequence of final lncRNAs:-

```
#!/usr/bin/perl -w

use strict;

my $idsfile = "final_lncRNA_ids";
my $seqfile = "transcripts.fa";
my %ids  = ();

openMYFILE, $idsfile;
while (<MYFILE>)
{
chomp;
  $ids{$_} += 1;
}
closeMYFILE;

local $/ = "\n>";  # read by FASTA record

open FASTA, $seqfile or die $!;
while (<FASTA>) {
chomp;
my $seq = $_;
my ($id) = $seq =~ /^>*(\S+)/;  # parse ID as first word in FASTA header
if (exists($ids{$id}))
   {
print "$seq" ."\n";
   }
}
close FASTA;
exit
```

APPENDIX IX

Supplementary Figures and Tables



Figurerepresenting box plot of FPKM distributions for individual conditions

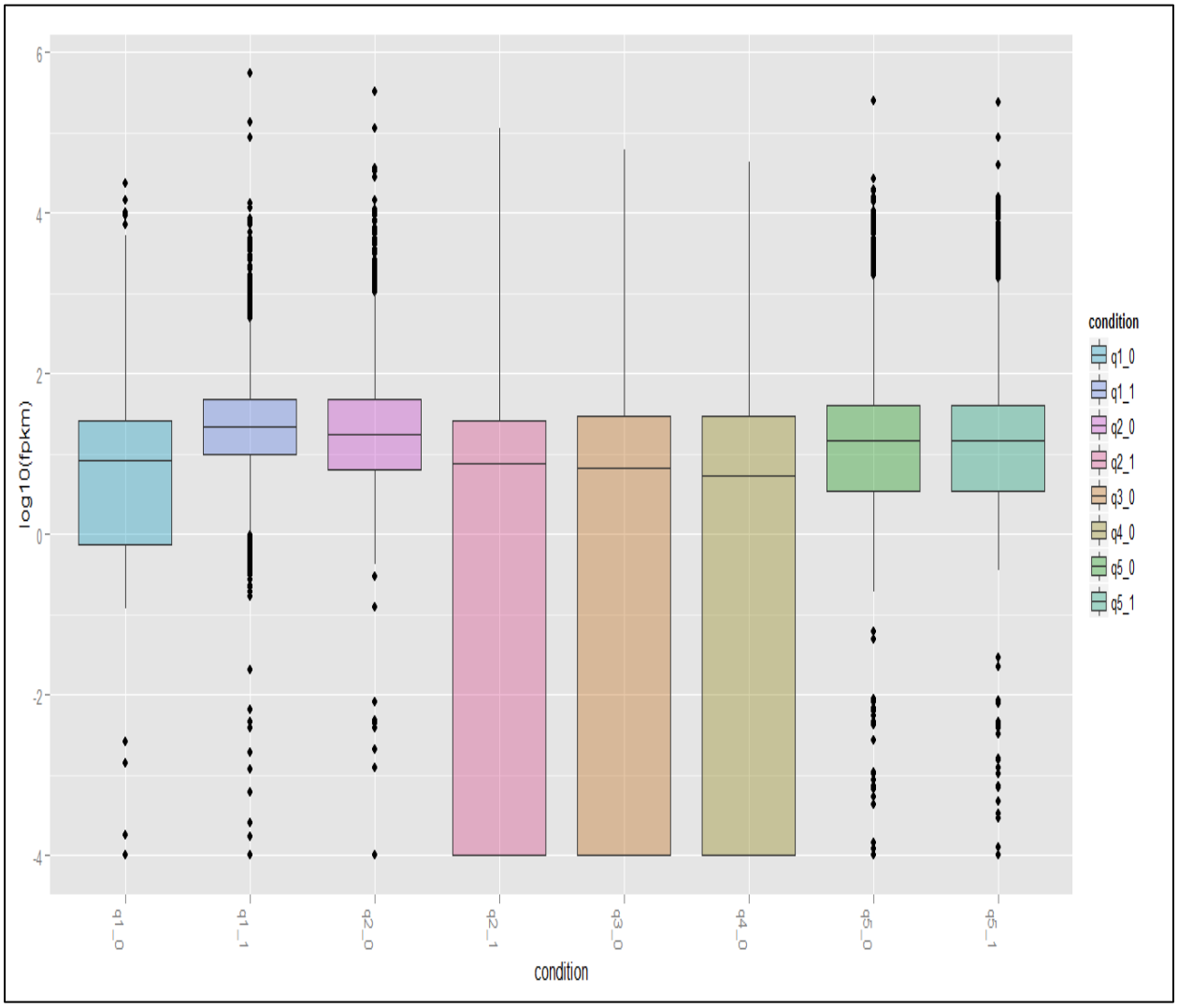


Figure representing box plot with replicates=TRUE exposes individual replicate

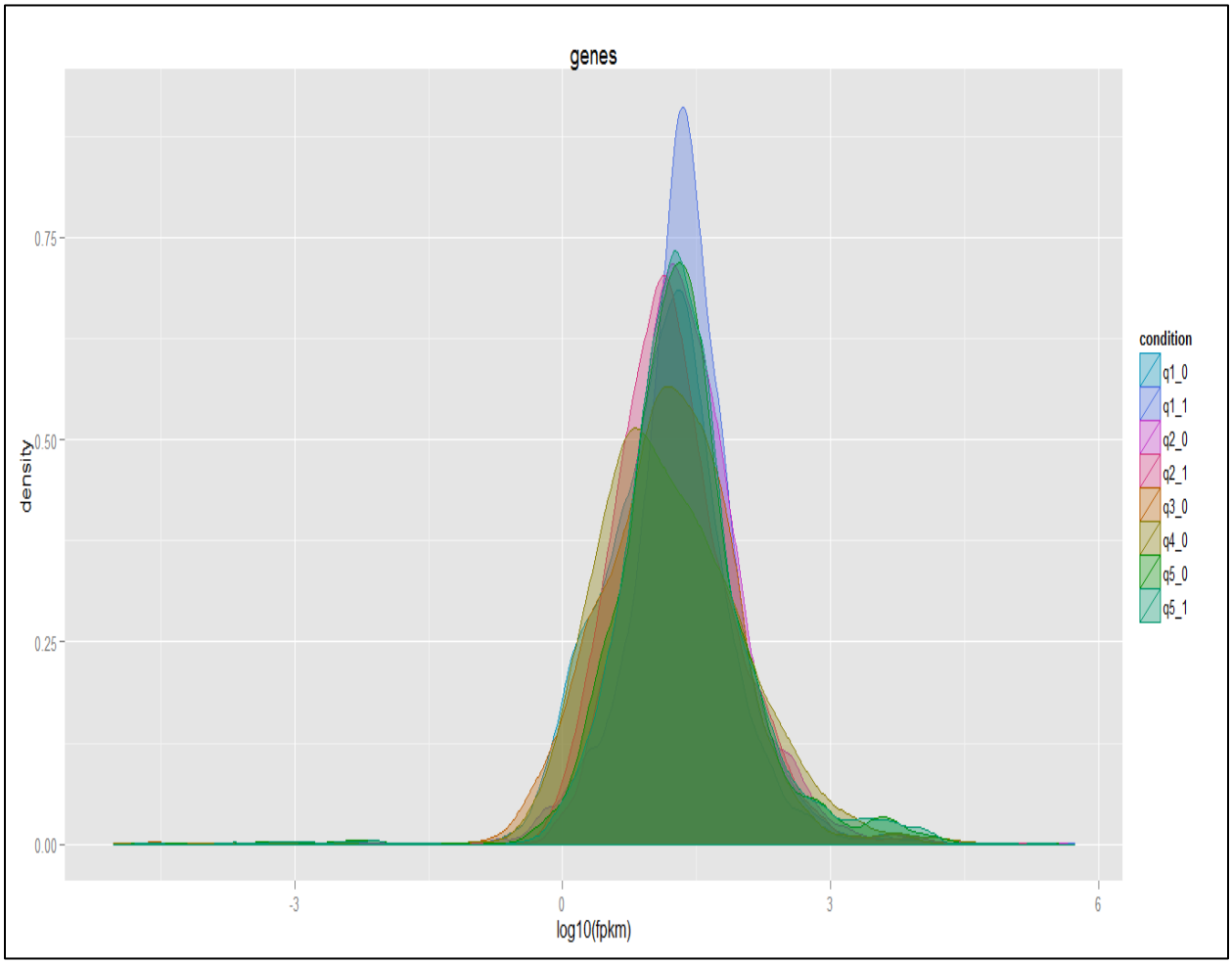


Figure representing density plot with replicates=TRUE exposes individual replicate FPKM distributions

Table indicating information about location of lncRNAs

S. No.	Transcript Id	Gene Id	Chromosome No.	Exon No.	Start	Stop
1	TCONS_00002540	XLOC_002233	6	1	608378	608701
2	TCONS_00004531	XLOC_004060	9	1	921716	921961
3	TCONS_00000115	XLOC_000092	1	1	187821	188045
4	TCONS_00003115	XLOC_002761	7	1	682004	682232
5	TCONS_00003715	XLOC_003311	8	1	566555	566883
6	TCONS_00003974	XLOC_003569	8	1	1338726	1338999
7	TCONS_00004525	XLOC_004054	9	1	901877	902100
8	TCONS_00001378	XLOC_001210	4	1	549453	549805
9	TCONS_00000153	XLOC_000130	1	1	312630	312870
10	TCONS_00002437	XLOC_002130	6	1	187276	187603
11	TCONS_00002699	XLOC_002388	6	1	1087689	1087975
12	TCONS_00001953	XLOC_001708	5	1	636226	636454
13	TCONS_00004379	XLOC_003909	9	1	461555	461760
14	TCONS_00000877	XLOC_000754	3	1	398825	399081
15	TCONS_00002150	XLOC_001901	5	1	1278749	1279011
16	TCONS_00003320	XLOC_002966	7	1	1351528	1351782
17	TCONS_00000806	XLOC_000683	3	1	164232	164436
18	TCONS_00003804	XLOC_003399	8	1	792291	792565
19	TCONS_00003837	XLOC_003432	8	1	919119	919689
20	TCONS_00002421	XLOC_002114	6	1	137505	138051
21	TCONS_00001249	XLOC_001081	4	1	179440	179642
22	TCONS_00000190	XLOC_000167	1	1	444764	445205
23	TCONS_00000780	XLOC_000657	3	1	84229	84449
24	TCONS_00003836	XLOC_003431	8	1	917917	919040
25	TCONS_00000799	XLOC_000676	3	1	155024	155227
26	TCONS_00002659	XLOC_002350	6	1	975460	975771
27	TCONS_00003960	XLOC_003555	8	1	1303153	1303649
28	TCONS_00004272	XLOC_003802	9	1	196658	196944
29	TCONS_00000927	XLOC_000804	3	1	561306	561528
30	TCONS_00004363	XLOC_003893	9	1	429959	430243
31	TCONS_00000874	XLOC_000751	3	1	393230	393598
32	TCONS_00003847	XLOC_003442	8	1	949113	949339
33	TCONS_00002698	XLOC_002387	6	1	1087126	1087627
34	TCONS_00002395	XLOC_002088	6	1	81606	81809
35	TCONS_00003706	XLOC_003302	8	1	525629	525868
36	TCONS_00003247	XLOC_002893	7	1	1154083	1154297
37	TCONS_00003095	XLOC_002741	7	1	579735	580045
38	TCONS_00002952	XLOC_002598	7	1	143084	143375
39	TCONS_00001460	XLOC_001292	4	1	913410	913656
40	TCONS_00001828	XLOC_001583	5	1	213266	213665

41	TCONS_00000929	XLOC_000806	3	1	568838	569301
42	TCONS_00000846	XLOC_000723	3	1	284678	284957
43	TCONS_00001801	XLOC_001556	5	1	121123	121666
44	TCONS_00004263	XLOC_003793	9	1	159276	159636
45	TCONS_00002648	XLOC_002340	6	1	944308	944596
46	TCONS_00003637	XLOC_003233	8	1	321280	321506
47	TCONS_00003806	XLOC_003401	8	1	819822	820309
48	TCONS_00002113	XLOC_001864	5	1	1114848	1115790
49	TCONS_00004320	XLOC_003850	9	1	303902	304263
50	TCONS_00001288	XLOC_001120	4	1	321076	321339
51	TCONS_00004350	XLOC_003880	9	1	372848	373079
52	TCONS_00002553	XLOC_002246	6	1	673191	673448
53	TCONS_00003096	XLOC_002742	7	1	582003	582429
54	TCONS_00004547	XLOC_004076	9	1	955347	955796
55	TCONS_00001355	XLOC_001187	4	1	502852	503198
56	TCONS_00002559	XLOC_002252	6	1	683702	684028
57	TCONS_00004477	XLOC_004006	9	1	720128	720407
58	TCONS_00004587	XLOC_004115	9	1	1103092	1103406
59	TCONS_00003600	XLOC_003196	8	1	164623	164901
60	TCONS_00004332	XLOC_003862	9	1	326980	327521
61	TCONS_00003581	XLOC_003177	8	1	90795	91090
62	TCONS_00004449	XLOC_003978	9	1	646951	647452
63	TCONS_00002491	XLOC_002184	6	1	394839	395059
64	TCONS_00001337	XLOC_001169	4	1	448281	448718
65	TCONS_00003258	XLOC_002904	7	1	1192999	1193545
66	TCONS_00004561	XLOC_004090	9	1	1024515	1024824
67	TCONS_00004411	XLOC_003940	9	1	554399	554614
68	TCONS_00001390	XLOC_001222	4	1	641888	642224
69	TCONS_00001841	XLOC_001596	5	1	247145	247352
70	TCONS_00002387	XLOC_002080	6	1	75346	75694
71	TCONS_00003238	XLOC_002884	7	1	1136020	1136503
72	TCONS_00001986	XLOC_001739	5	1	718455	719042
73	TCONS_00003902	XLOC_003497	8	1	1139071	1139590
74	TCONS_00003577	XLOC_003173	8	1	75460	75815
75	TCONS_00003051	XLOC_002697	7	1	431459	431875
76	TCONS_00004390	XLOC_003920	9	1	482392	482900
77	TCONS_00004606	XLOC_004134	9	1	1184054	1184393
78	TCONS_00003798	XLOC_003393	8	1	781393	781699
79	TCONS_00001046	XLOC_000923	3	1	959328	959579
80	TCONS_00001905	XLOC_001660	5	1	483710	484225
81	TCONS_00002101	XLOC_001852	5	1	1069792	1070050
82	TCONS_00003852	XLOC_003447	8	1	964943	965799
83	TCONS_00003237	XLOC_002883	7	1	1132290	1132772

84	TCONS_00000144	XLOC_000121	1	1	271376	271857
85	TCONS_00003705	XLOC_003301	8	1	525345	525559
86	TCONS_00002162	XLOC_001913	5	1	1304785	1305052
87	TCONS_00003721	XLOC_003317	8	1	589549	590033
88	TCONS_00003822	XLOC_003417	8	1	875092	875505
89	TCONS_00001321	XLOC_001153	4	1	401867	402153
90	TCONS_00003171	XLOC_002817	7	1	881969	882205
91	TCONS_00003255	XLOC_002901	7	1	1172694	1173249
92	TCONS_00003087	XLOC_002733	7	1	549591	549858
93	TCONS_00002589	XLOC_002281	6	1	794952	795168
94	TCONS_00001446	XLOC_001278	4	1	847473	847688
95	TCONS_00001016	XLOC_000893	3	1	833089	833412
96	TCONS_00003872	XLOC_003467	8	1	1027634	1027957
97	TCONS_00002485	XLOC_002178	6	1	361460	361765
98	TCONS_00001784	XLOC_001539	5	1	86625	87389
99	TCONS_00004679	XLOC_004207	9	1	1410335	1410665
100	TCONS_00004575	XLOC_004104	9	1	1063858	1064130
101	TCONS_00004264	XLOC_003794	9	1	161191	161473
102	TCONS_00004532	XLOC_004061	9	1	922016	922248
103	TCONS_00003020	XLOC_002666	7	1	336494	336815
104	TCONS_00001503	XLOC_001335	4	1	1042580	1042884
105	TCONS_00001396	XLOC_001228	4	1	653905	654134
106	TCONS_00001497	XLOC_001329	4	1	1023897	1024219
107	TCONS_00003018	XLOC_002664	7	1	324496	324847
108	TCONS_00004489	XLOC_004018	9	1	752677	752966
109	TCONS_00001479	XLOC_001311	4	1	993752	993990
110	TCONS_00004593	XLOC_004121	9	1	1117552	1117762
111	TCONS_00003752	XLOC_003348	8	1	664547	665188
112	TCONS_00003331	XLOC_002977	7	1	1389656	1390107
113	TCONS_00004543	XLOC_004072	9	1	950149	950435
114	TCONS_00002512	XLOC_002205	6	1	456012	456352
115	TCONS_00001489	XLOC_001321	4	1	1001068	1001283
116	TCONS_00002978	XLOC_002624	7	1	211013	211218
117	TCONS_00004281	XLOC_003811	9	1	233330	233636
118	TCONS_00000450	XLOC_000383	2	1	435016	435579
119	TCONS_00000179	XLOC_000156	1	1	413870	414142
120	TCONS_00004447	XLOC_003976	9	1	644965	645523
121	TCONS_00000842	XLOC_000719	3	1	278176	278481
122	TCONS_00002028	XLOC_001779	5	1	861655	862053
123	TCONS_00000962	XLOC_000839	3	1	661900	662368
124	TCONS_00004259	XLOC_003789	9	1	138769	139085
125	TCONS_00002430	XLOC_002123	6	1	165693	166024
126	TCONS_00003829	XLOC_003424	8	1	898296	898624

127	TCONS_00003579	XLOC_003175	8	1	85699	85994
128	TCONS_00002676	XLOC_002365	6	1	1037765	1038146
129	TCONS_00001916	XLOC_001671	5	1	524143	524355
130	TCONS_00002402	XLOC_002095	6	1	99128	99410
131	TCONS_00001305	XLOC_001137	4	1	362073	362418
132	TCONS_00004354	XLOC_003884	9	1	400952	401219
133	TCONS_00000141	XLOC_000118	1	1	262750	263151
134	TCONS_00003656	XLOC_003252	8	1	389973	390382
135	TCONS_00003234	XLOC_002880	7	1	1119949	1120273
136	TCONS_00003886	XLOC_003481	8	1	1066029	1066603
137	TCONS_00004684	XLOC_004212	9	1	1431543	1431755
138	TCONS_00002117	XLOC_001868	5	1	1131741	1132129
139	TCONS_00003745	XLOC_003341	8	1	643053	643357
140	TCONS_00001413	XLOC_001245	4	1	717320	717620
141	TCONS_00002947	XLOC_002593	7	1	105752	106166
142	TCONS_00003889	XLOC_003484	8	1	1083458	1083711
143	TCONS_00004453	XLOC_003982	9	1	656629	657029
144	TCONS_00002067	XLOC_001818	5	1	992406	992642
145	TCONS_00004641	XLOC_004169	9	1	1308691	1308958
146	TCONS_00003248	XLOC_002894	7	1	1154528	1154973
147	TCONS_00003858	XLOC_003453	8	1	995598	996101
148	TCONS_00002077	XLOC_001828	5	1	1012829	1013173
149	TCONS_00000555	XLOC_000488	2	1	737399	737603
150	TCONS_00003863	XLOC_003458	8	1	1004639	1005123
151	TCONS_00002706	XLOC_002395	6	1	1118399	1119331
152	TCONS_00003667	XLOC_003263	8	1	426819	427051
153	TCONS_00000833	XLOC_000710	3	1	244643	244882
154	TCONS_00003892	XLOC_003487	8	1	1094164	1094562
155	TCONS_00001423	XLOC_001255	4	1	762580	762974
156	TCONS_00003008	XLOC_002654	7	1	305385	305595
157	TCONS_00003206	XLOC_002852	7	1	1038827	1039337
158	TCONS_00002651	XLOC_002343	6	1	952489	952719
159	TCONS_00003268	XLOC_002914	7	1	1217430	1217711
160	TCONS_00000878	XLOC_000755	3	1	405925	406339
161	TCONS_00001799	XLOC_001554	5	1	118219	118871
162	TCONS_00000841	XLOC_000718	3	1	277102	277331
163	TCONS_00000779	XLOC_000656	3	1	83445	83700
164	TCONS_00003186	XLOC_002832	7	1	976042	976352
165	TCONS_00002423	XLOC_002116	6	1	147243	147745
166	TCONS_00002399	XLOC_002092	6	1	90962	91318
167	TCONS_00003319	XLOC_002965	7	1	1351007	1351301
168	TCONS_00004254	XLOC_003784	9	1	116744	117190
169	TCONS_00000154	XLOC_000131	1	1	313826	314233

170	TCONS_00000510	XLOC_000443	2	1	625045	625377
171	TCONS_00002741	XLOC_002430	6	1	1279154	1279421
172	TCONS_00004484	XLOC_004013	9	1	732044	732309
173	TCONS_00004580	XLOC_004109	9	1	1083603	1083981
174	TCONS_00002558	XLOC_002251	6	1	683383	683599
175	TCONS_00003069	XLOC_002715	7	1	489083	489284
176	TCONS_00003222	XLOC_002868	7	1	1078193	1078486
177	TCONS_00001367	XLOC_001199	4	1	521822	522162
178	TCONS_00002021	XLOC_001772	5	1	845882	846098
179	TCONS_00003864	XLOC_003459	8	1	1006098	1006360
180	TCONS_00001328	XLOC_001160	4	1	419058	419401
181	TCONS_00003180	XLOC_002826	7	1	946455	946915
182	TCONS_00002484	XLOC_002177	6	1	360482	361192
183	TCONS_00001926	XLOC_001681	5	1	549575	549878
184	TCONS_00001368	XLOC_001200	4	1	522267	522761
185	TCONS_00001442	XLOC_001274	4	1	839878	840480
186	TCONS_00002991	XLOC_002637	7	1	228484	228821
187	TCONS_00004681	XLOC_004209	9	1	1411470	1411711
188	TCONS_00003646	XLOC_003242	8	1	351021	351344
189	TCONS_00002711	XLOC_002400	6	1	1158579	1159340
190	TCONS_00002005	XLOC_001756	5	1	782341	782741
191	TCONS_00003813	XLOC_003408	8	1	839759	840235
192	TCONS_00003192	XLOC_002838	7	1	989569	989902
193	TCONS_00002946	XLOC_002592	7	1	103505	103831
194	TCONS_00003585	XLOC_003181	8	1	119286	119724
195	TCONS_00002449	XLOC_002142	6	1	238818	239079
196	TCONS_00002645	XLOC_002337	6	1	931241	931459
197	TCONS_00003784	XLOC_003379	8	1	752924	753289
198	TCONS_00003952	XLOC_003547	8	1	1271684	1272057
199	TCONS_00002088	XLOC_001839	5	1	1038863	1039073
200	TCONS_00003151	XLOC_002797	7	1	827759	828060
201	TCONS_00003920	XLOC_003515	8	1	1195234	1195469
202	TCONS_00004315	XLOC_003845	9	1	298172	298428
203	TCONS_00001800	XLOC_001555	5	1	119066	119473
204	TCONS_00002070	XLOC_001821	5	1	995626	995868
205	TCONS_00003773	XLOC_003369	8	1	715654	716214
206	TCONS_00002098	XLOC_001849	5	1	1059530	1059872
207	TCONS_00002943	XLOC_002590	7	1	100766	102030
208	TCONS_00001901	XLOC_001656	5	1	465673	466023
209	TCONS_00002114	XLOC_001865	5	1	1115898	1116339
210	TCONS_00000209	XLOC_000186	1	1	543007	543320
211	TCONS_00000122	XLOC_000099	1	1	206130	206352
212	TCONS_00003249	XLOC_002895	7	1	1155153	1155503

213	TCONS_00001978	XLOC_001731	5	1	707652	708237
214	TCONS_00002596	XLOC_002288	6	1	813280	813555
215	TCONS_00002945	XLOC_002591	7	1	103054	103292
216	TCONS_00004251	XLOC_003781	9	1	97746	97973
217	TCONS_00003767	XLOC_003363	8	1	706705	707264
218	TCONS_00004560	XLOC_004089	9	1	1016654	1016859
219	TCONS_00000116	XLOC_000093	1	1	188301	188822
220	TCONS_00002724	XLOC_002413	6	1	1209361	1210154
221	TCONS_00004435	XLOC_003964	9	1	606348	606630
222	TCONS_00003269	XLOC_002915	7	1	1221902	1222171
223	TCONS_00004550	XLOC_004079	9	1	965065	965408
224	TCONS_00001831	XLOC_001586	5	1	215545	215768
225	TCONS_00002538	XLOC_002231	6	1	603225	603506
226	TCONS_00004662	XLOC_004190	9	1	1361131	1361360
227	TCONS_00001920	XLOC_001675	5	1	531052	531335
228	TCONS_00002383	XLOC_002076	6	1	72256	72571
229	TCONS_00001863	XLOC_001618	5	1	321342	321712
230	TCONS_00003898	XLOC_003493	8	1	1123516	1123726
231	TCONS_00000554	XLOC_000487	2	1	730600	730932
232	TCONS_00004604	XLOC_004132	9	1	1170491	1170691
233	TCONS_00002142	XLOC_001893	5	1	1241140	1241546
234	TCONS_00003615	XLOC_003211	8	1	208546	208920
235	TCONS_00001886	XLOC_001641	5	1	417849	418221
236	TCONS_00000396	XLOC_000331	2	1	181565	181961
237	TCONS_00003286	XLOC_002932	7	1	1259892	1260192
238	TCONS_00001798	XLOC_001553	5	1	117876	118145
239	TCONS_00004546	XLOC_004075	9	1	953977	954210
240	TCONS_00004443	XLOC_003972	9	1	629194	629918
241	TCONS_00000539	XLOC_000472	2	1	688761	689028
242	TCONS_00001900	XLOC_001655	5	1	465129	465402
243	TCONS_00002091	XLOC_001842	5	1	1044394	1044657
244	TCONS_00003911	XLOC_003506	8	1	1162250	1162486
245	TCONS_00002382	XLOC_002075	6	1	70378	70838
246	TCONS_00003694	XLOC_003290	8	1	499626	500013
247	TCONS_00003318	XLOC_002964	7	1	1347856	1348144
248	TCONS_00004438	XLOC_003967	9	1	613815	614091
249	TCONS_00003904	XLOC_003499	8	1	1141969	1142214
250	TCONS_00000956	XLOC_000833	3	1	644336	644636
251	TCONS_00002517	XLOC_002210	6	1	474928	475491
252	TCONS_00004397	XLOC_003927	9	1	499076	499461
253	TCONS_00004269	XLOC_003799	9	1	189012	189569
254	TCONS_00000472	XLOC_000405	2	1	507368	507897
255	TCONS_00001056	XLOC_000933	3	1	990786	991200

256	TCONS_00004274	XLOC_003804	9	1	201336	201671
257	TCONS_00001006	XLOC_000883	3	1	817706	818003
258	TCONS_00001872	XLOC_001627	5	1	363150	363468
259	TCONS_00001874	XLOC_001629	5	1	364912	365973
260	TCONS_00003058	XLOC_002704	7	1	457996	458251
261	TCONS_00001369	XLOC_001201	4	1	522879	523672
262	TCONS_00002617	XLOC_002309	6	1	847189	847536
263	TCONS_00004310	XLOC_003840	9	1	288281	288603
264	TCONS_00003751	XLOC_003347	8	1	656044	656420
265	TCONS_00002133	XLOC_001884	5	1	1208654	1209062
266	TCONS_00003065	XLOC_002711	7	1	476909	477170
267	TCONS_00003692	XLOC_003288	8	1	497788	498078
268	TCONS_00002944	XLOC_002590	7	1	101517	102030
269	TCONS_00003159	XLOC_002805	7	1	838877	839122
270	TCONS_00001790	XLOC_001545	5	1	99868	100070
271	TCONS_00004328	XLOC_003858	9	1	317040	317698
272	TCONS_00001269	XLOC_001101	4	1	231832	232158
273	TCONS_00001345	XLOC_001177	4	1	474318	475065
274	TCONS_00000402	XLOC_000337	2	1	196637	196974
275	TCONS_00003578	XLOC_003174	8	1	77550	77885
276	TCONS_00002122	XLOC_001873	5	1	1161976	1162410
277	TCONS_00003124	XLOC_002770	7	1	699710	699933
278	TCONS_00003728	XLOC_003324	8	1	596523	596784
279	TCONS_00000791	XLOC_000668	3	1	125860	126067
280	TCONS_00001777	XLOC_001532	5	1	68020	68234
281	TCONS_00003765	XLOC_003361	8	1	701712	701977
282	TCONS_00001884	XLOC_001639	5	1	408560	409037
283	TCONS_00001981	XLOC_001734	5	1	710428	710839
284	TCONS_00002446	XLOC_002139	6	1	232521	232829
285	TCONS_00004680	XLOC_004208	9	1	1410718	1411121
286	TCONS_00004591	XLOC_004119	9	1	1112874	1113449
287	TCONS_00003219	XLOC_002865	7	1	1074137	1074551
288	TCONS_00003334	XLOC_002980	7	1	1404702	1405034
289	TCONS_00003288	XLOC_002934	7	1	1260889	1261151
290	TCONS_00002454	XLOC_002147	6	1	246654	246862
291	TCONS_00000921	XLOC_000798	3	1	527897	528155
292	TCONS_00004630	XLOC_004158	9	1	1262059	1262348
293	TCONS_00001266	XLOC_001098	4	1	224007	224247
294	TCONS_00003036	XLOC_002682	7	1	384471	384725
295	TCONS_00001505	XLOC_001337	4	1	1043623	1044023
296	TCONS_00004677	XLOC_004205	9	1	1407412	1407834
297	TCONS_00003047	XLOC_002693	7	1	421309	421813
298	TCONS_00000572	XLOC_000504	2	1	808406	808937

299	TCONS_00003953	XLOC_003548	8	1	1272391	1272760
300	TCONS_00002068	XLOC_001819	5	1	992756	992990
301	TCONS_00001485	XLOC_001317	4	1	997828	998247
302	TCONS_00003790	XLOC_003385	8	1	772797	773011
303	TCONS_00002167	XLOC_001918	5	1	1342733	1342942
304	TCONS_00004664	XLOC_004192	9	1	1383652	1384161
305	TCONS_00001027	XLOC_000904	3	1	883229	883433
306	TCONS_00003720	XLOC_003316	8	1	589091	589471
307	TCONS_00000084	XLOC_000061	1	1	1572	2001
308	TCONS_00002107	XLOC_001858	5	1	1097965	1098338
309	TCONS_00003811	XLOC_003406	8	1	838504	838866
310	TCONS_00003659	XLOC_003255	8	1	407301	407515
311	TCONS_00004367	XLOC_003897	9	1	437249	437595
312	TCONS_00001860	XLOC_001615	5	1	315473	315753
313	TCONS_00004437	XLOC_003966	9	1	607798	608119
314	TCONS_00003015	XLOC_002661	7	1	316615	316896
315	TCONS_00001268	XLOC_001100	4	1	231314	231553
316	TCONS_00003168	XLOC_002814	7	1	855312	856091
317	TCONS_00002962	XLOC_002608	7	1	164172	164380
318	TCONS_00003814	XLOC_003409	8	1	840355	840650
319	TCONS_00000143	XLOC_000120	1	1	270842	271174
320	TCONS_00004252	XLOC_003782	9	1	100257	100610
321	TCONS_00003818	XLOC_003413	8	1	856093	856316
322	TCONS_00003075	XLOC_002721	7	1	504952	505697
323	TCONS_00003287	XLOC_002933	7	1	1260259	1260495
324	TCONS_00000910	XLOC_000787	3	1	484579	484838
325	TCONS_00002066	XLOC_001817	5	1	989391	989673
326	TCONS_00002127	XLOC_001878	5	1	1179634	1179866
327	TCONS_00000782	XLOC_000659	3	1	85782	86001
328	TCONS_00001480	XLOC_001312	4	1	994917	995346
329	TCONS_00004422	XLOC_003951	9	1	573465	573855
330	TCONS_00001482	XLOC_001314	4	1	995824	996068
331	TCONS_00003068	XLOC_002714	7	1	487566	488588
332	TCONS_00000886	XLOC_000763	3	1	415041	415245
333	TCONS_00002990	XLOC_002636	7	1	225960	226548
334	TCONS_00000961	XLOC_000838	3	1	654135	654369
335	TCONS_00000372	XLOC_000307	2	1	77667	77875
336	TCONS_00003950	XLOC_003545	8	1	1269479	1269703
337	TCONS_00001370	XLOC_001202	4	1	523742	523958
338	TCONS_00001304	XLOC_001136	4	1	354216	354633
339	TCONS_00001258	XLOC_001090	4	1	201323	201524
340	TCONS_00003912	XLOC_003507	8	1	1164171	1165341
341	TCONS_00002985	XLOC_002631	7	1	218226	218609

342	TCONS_00004509	XLOC_004038	9	1	848569	848951
343	TCONS_00003851	XLOC_003446	8	1	963573	964152
344	TCONS_00000537	XLOC_000470	2	1	685483	685862
345	TCONS_00003871	XLOC_003466	8	1	1025659	1027371
346	TCONS_00004627	XLOC_004155	9	1	1253924	1254281
347	TCONS_00003907	XLOC_003502	8	1	1144322	1144615
348	TCONS_00000970	XLOC_000847	3	1	694802	695013
349	TCONS_00003609	XLOC_003205	8	1	190932	191153
350	TCONS_00003859	XLOC_003454	8	1	1001538	1001969
351	TCONS_00002592	XLOC_002284	6	1	809693	810015
352	TCONS_00002161	XLOC_001912	5	1	1303121	1303440
353	TCONS_00003285	XLOC_002931	7	1	1259387	1259603
354	TCONS_00000550	XLOC_000483	2	1	718930	719205
355	TCONS_00003830	XLOC_003425	8	1	898700	898988
356	TCONS_00002407	XLOC_002100	6	1	110737	111112
357	TCONS_00002555	XLOC_002248	6	1	678942	679200
358	TCONS_00002000	XLOC_001751	5	1	768702	769122
359	TCONS_00001402	XLOC_001234	4	1	678135	678374
360	TCONS_00000371	XLOC_000306	2	1	77411	77613
361	TCONS_00003878	XLOC_003473	8	1	1042444	1042671
362	TCONS_00003693	XLOC_003289	8	1	499112	499354
363	TCONS_00003719	XLOC_003315	8	1	575664	576019
364	TCONS_00000978	XLOC_000855	3	1	716312	717374
365	TCONS_00001483	XLOC_001315	4	1	996425	996635
366	TCONS_00001433	XLOC_001265	4	1	799711	800051
367	TCONS_00002533	XLOC_002226	6	1	553662	554090
368	TCONS_00002112	XLOC_001863	5	1	1114462	1114754
369	TCONS_00004303	XLOC_003833	9	1	268469	268969
370	TCONS_00002413	XLOC_002106	6	1	118184	118411
371	TCONS_00001454	XLOC_001286	4	1	883322	883821
372	TCONS_00001917	XLOC_001672	5	1	524922	525180
373	TCONS_00002740	XLOC_002429	6	1	1278827	1279100
374	TCONS_00003293	XLOC_002939	7	1	1292374	1292833
375	TCONS_00003875	XLOC_003470	8	1	1035688	1036052
376	TCONS_00001853	XLOC_001608	5	1	296181	296660
377	TCONS_00000865	XLOC_000742	3	1	347752	348104
378	TCONS_00000204	XLOC_000181	1	1	518418	518635
379	TCONS_00004418	XLOC_003947	9	1	570784	571014
380	TCONS_00004511	XLOC_004040	9	1	851590	851834
381	TCONS_00003203	XLOC_002849	7	1	1030714	1030922
382	TCONS_00000465	XLOC_000398	2	1	491505	491729
383	TCONS_00003250	XLOC_002896	7	1	1158635	1158839
384	TCONS_00002392	XLOC_002085	6	1	79894	80438

385	TCONS_00002025	XLOC_001776	5	1	852063	852896
386	TCONS_00003114	XLOC_002760	7	1	681659	681943
387	TCONS_00003240	XLOC_002886	7	1	1146230	1146552
388	TCONS_00003855	XLOC_003450	8	1	983666	983924
389	TCONS_00001297	XLOC_001129	4	1	345374	345690
390	TCONS_00001353	XLOC_001185	4	1	499634	500594
391	TCONS_00003132	XLOC_002778	7	1	753410	753735
392	TCONS_00004294	XLOC_003824	9	1	254666	254973
393	TCONS_00000417	XLOC_000352	2	1	248153	248467
394	TCONS_00000139	XLOC_000116	1	1	248268	248633
395	TCONS_00003164	XLOC_002810	7	1	844364	844709
396	TCONS_00001296	XLOC_001128	4	1	344204	344557
397	TCONS_00001513	XLOC_001345	4	1	1080071	1080589
398	TCONS_00003748	XLOC_003344	8	1	647670	647998
399	TCONS_00003300	XLOC_002946	7	1	1316768	1316986
400	TCONS_00003162	XLOC_002808	7	1	841559	842610
401	TCONS_00002125	XLOC_001876	5	1	1167713	1168090
402	TCONS_00001904	XLOC_001659	5	1	479480	479943
403	TCONS_00000533	XLOC_000466	2	1	680450	680727
404	TCONS_00000400	XLOC_000335	2	1	195363	195753
405	TCONS_00000784	XLOC_000661	3	1	95267	95651
406	TCONS_00002993	XLOC_002639	7	1	230692	230960
407	TCONS_00001847	XLOC_001602	5	1	266345	266633
408	TCONS_00000817	XLOC_000694	3	1	209652	209929
409	TCONS_00003825	XLOC_003420	8	1	880312	880532
410	TCONS_00002779	XLOC_002460	7	1	883094	883157
411	TCONS_00004618	XLOC_004146	9	1	1230217	1230483
412	TCONS_00002249	XLOC_001977	6	1	908015	908110
413	TCONS_00003947	XLOC_003542	8	1	1266359	1266602
414	TCONS_00003239	XLOC_002885	7	1	1137991	1138492
415	TCONS_00002445	XLOC_002138	6	1	231400	232318
416	TCONS_00001797	XLOC_001552	5	1	115529	115970
417	TCONS_00004462	XLOC_003991	9	1	688182	688480
418	TCONS_00001324	XLOC_001156	4	1	404587	404909
419	TCONS_00001011	XLOC_000888	3	1	822542	822793
420	TCONS_00003183	XLOC_002829	7	1	960999	961335
421	TCONS_00003223	XLOC_002869	7	1	1081092	1081313
422	TCONS_00002081	XLOC_001832	5	1	1024246	1024527
423	TCONS_00000428	XLOC_000363	2	1	337793	338170
424	TCONS_00000575	XLOC_000507	2	1	834896	835253
425	TCONS_00003315	XLOC_002961	7	1	1338221	1338690
426	TCONS_00000560	XLOC_000493	2	1	755576	755898

Table indicating isoforms FPKM values and count values obtained from cuffdiff and htseq-count respectively

Transcript Ids	Gametocyte V			Gametocyte II			SCHIZONT		Late Trophozoite			
	Single-ended		Paired ended	Single-ended		Paired ended	Single-ended		Paired-ended		Single-ended	
	FPKM	Count		FPKM	Count		FPKM	Count	FPKM	Count	FPKM	Count
TCONS_00000084	0	0	0	0	0	0	13.98	10	0	0	0	0
TCONS_00000115	93.28	32	0	333.3	30	0	0	0	0	0	72.55	7
TCONS_00000116	5.14	13	0	50.9	32	2	0	0	1.9	1	4.69	3
TCONS_00000122	35.29	11	1	57.18	5	0	0	0	0	0	100.46	9
TCONS_00000139	14.63	19	0	16.98	5	1	0	0	0	0	8.33	3
TCONS_00000141	8.22	13	0	26.09	10	1	0	0	5.5	2	0	0
TCONS_00000143	14.35	15	0	28.21	7	1	0	0	0	0	8.22	2
TCONS_00000144	7.68	17	0	13.77	8	0	0	0	2.11	1	1.48	1
TCONS_00000153	23.45	10	0	53.62	6	0	0	0	13.92	2	177.14	20
TCONS_00000154	7.98	13	0	14.04	6	0	0	0	0	0	4	2
TCONS_00000179	49.09	18	13	0	0	0	3.44	1	0	0	10.52	2
TCONS_00000190	9.49	18	0	30.12	15	0	0	0	11.94	5	6.8	3
TCONS_00000204	0	0	0	0	0	0	0	0	0	0	223.04	18
TCONS_00000209	24.33	22	0	83.07	19	1	0	0	0	0	0	0
TCONS_00000371	9.7	1	2	23.63	1	1	0	0	0	0	376	25
TCONS_00000372	2.63	0	1	28.94	1	2	0	0	9.88	1	268.98	20
TCONS_00000396	9.08	14	0	9.89	4	0	0	0	0	0	0	0
TCONS_00000400	6.9	9	1	2.08	0	1	21.58	12	8.6	3	50.54	20
TCONS_00000402	4.01	2	2	8.78	1	2	22.89	10	10.86	3	14.91	4
TCONS_00000417	15.47	13	1	12.55	3	0	0	0	0	0	222.32	52
TCONS_00000428	25.13	35	0	5.48	2	0	0	0	3.02	1	3.09	1
TCONS_00000450	44.36	124	2	15.97	12	0	0	0	1.71	1	41.89	33
TCONS_00000465	32.07	11	0	88.88	8	0	0	0	0	0	0	0
TCONS_00000472	16.91	11	22	12.16	1	8	35.66	31	48.17	26	62.4	43
TCONS_00000510	9.57	10	0	25.51	7	0	0	0	0	0	0	0
TCONS_00000533	9.08	3	3	3.78	0	1	36.42	11	5.16	1	0	0
TCONS_00000537	5.68	8	0	29.76	11	0	0	0	0	0	5.37	2
TCONS_00000539	32.36	19	0	6.5	1	0	3.58	1	0	0	57.02	9
TCONS_00000550	28.17	18	0	47.72	8	0	0	0	0	0	5.1	1
TCONS_00000554	16.56	15	2	18.23	5	0	0	0	3.71	1	11.34	3
TCONS_00000555	43.58	11	0	45.29	3	0	0	0	41.62	4	209.51	14
TCONS_00000560	15.45	15	0	3.93	1	0	0	0	0	0	11.14	3
TCONS_00000572	0	0	0	0	0	0	0	0	16.59	9	206.42	145

TCONS_00000575	0.81	1	0	0	0	0	0	0	0	0	339.24	110
TCONS_00000779	19.5	10	0	7.43	1	0	0	0	0	0	0	0
TCONS_00000780	33.22	10	1	47.1	4	0	0	0	17.11	2	0	0
TCONS_00000782	37.63	12	0	0	0	0	0	0	17.3	2	176.06	15
TCONS_00000784	0	0	0	0	0	0	0	0	38.14	13	2.97	1
TCONS_00000791	9.08	1	2	209.41	13	3	0	0	0	0	12.3	1
TCONS_00000799	48.32	12	0	76.74	5	0	0	0	0	0	34.62	2
TCONS_00000806	67.34	17	0	45.29	3	0	0	0	0	0	12.91	1
TCONS_00000817	36.8	24	0	23.38	4	0	0	0	0	0	23.17	4
TCONS_00000833	33.25	14	0	45.26	5	0	0	0	0	0	10.21	1
TCONS_00000841	40.75	15	0	51.77	5	0	0	0	7.76	1	8.86	1
TCONS_00000842	13	11	0	31.53	7	0	0	0	0	0	0	0
TCONS_00000846	25.43	13	4	17.19	3	0	0	0	15.24	3	30.73	5
TCONS_00000865	16.21	17	2	0	0	0	0	0	0	0	6.31	2
TCONS_00000874	9.83	13	0	11.53	4	0	0	0	0	0	8.97	3
TCONS_00000877	9.64	5	0	95.47	13	0	0	0	0	0	22.84	3
TCONS_00000878	29.56	39	8	13.21	5	1	8.2	5	5.24	2	3.87	2
TCONS_00000886	43.58	11	0	0	0	0	0	0	0	0	34.05	2
TCONS_00000910	27.92	15	0	28.38	4	0	0	0	0	0	0	0
TCONS_00000921	3.58	1	1	231.19	31	2	0	0	101.09	17	79.22	11
TCONS_00000927	44.29	14	1	22.88	2	0	0	0	0	0	0	0
TCONS_00000929	16.61	33	1	42.3	23	0	0	0	6.68	3	1.58	1
TCONS_00000956	22.21	18	0	0	0	0	0	0	0	0	4.02	1
TCONS_00000961	45.67	18	0	0	0	0	4.85	1	0	0	82.32	9
TCONS_00000962	20.57	42	1	30.68	17	0	4.1	3	0	0	64.34	35
TCONS_00000970	42.5	12	0	54	4	0	0	0	0	0	0	0
TCONS_00000978	18.71	20	63	26.41	22	25	44.43	92	7.87	10	40.23	73
TCONS_00001006	12.67	10	0	19.32	4	0	5.77	2	0	0	119.01	25
TCONS_00001011	28.62	14	0	15.59	2	0	0	0	0	0	50.87	7
TCONS_00001016	11.25	11	0	13.46	2	2	0	0	7.78	2	8.79	2
TCONS_00001027	51.5	13	0	15.1	1	0	0	0	10.41	1	0	0
TCONS_00001046	0	0	0	0	0	0	0	0	0	0	152.33	20
TCONS_00001056	7.13	12	0	2.27	1	0	0	0	2.62	1	0	0
TCONS_00001249	53.22	13	0	78.02	5	0	0	0	0	0	0	0
TCONS_00001258	66.6	16	0	39.87	2	1	0	0	0	0	35.78	2
TCONS_00001266	25.79	11	0	35.75	4	0	4.56	1	0	0	0	0
TCONS_00001268	28.5	12	0	0	0	0	4.61	1	0	0	0	0
TCONS_00001269	13	13	0	7.62	2	0	0	0	0	0	3.26	1
TCONS_00001288	17.8	10	0	38.11	5	1	0	0	0	0	0	0

TCONS_00001296	18.96	18	4	45.48	9	7	152.22	72	3.36	1	0	0
TCONS_00001297	9.95	7	2	4.12	1	0	0	0	0	0	220.36	55
TCONS_00001304	11.95	15	4	8.57	3	1	6.49	4	7.77	3	5.04	2
TCONS_00001305	27.96	32	0	6.66	2	0	0	0	0	0	0	0
TCONS_00001321	88.39	63	0	26.74	5	0	0	0	0	0	0	0
TCONS_00001324	11.1	4	6	0	0	0	29.76	12	11.73	3	0	0
TCONS_00001328	14.17	16	0	43.05	12	1	0	0	0	0	0	0
TCONS_00001337	7.51	14	0	3.8	1	1	1.52	1	0	0	3.5	2
TCONS_00001345	8.16	32	2	27.34	31	0	5.15	7	1.2	1	12.72	14
TCONS_00001353	27.89	36	78	17.76	20	8	13.58	25	4.42	5	103.94	170
TCONS_00001355	29.51	34	0	197.64	59	1	2.19	1	0	0	0	0
TCONS_00001367	9.21	9	1	10.34	3	0	0	0	0	0	19.42	5
TCONS_00001368	10.36	24	0	4.94	3	0	0	0	2.04	1	9.34	6
TCONS_00001369	8.74	36	3	8.12	9	1	0.69	1	3.33	3	4.84	6
TCONS_00001370	36.08	11	0	25	2	0	0	0	0	0	0	0
TCONS_00001378	26.4	28	3	30.58	5	6	4.25	2	16.84	5	100.86	32
TCONS_00001390	23.39	24	1	14.17	4	0	0	0	0	0	9.09	3
TCONS_00001396	29.31	10	1	41.42	4	0	0	0	38.79	5	0	0
TCONS_00001402	104.5	44	0	45.26	5	0	0	0	0	0	17.95	2
TCONS_00001413	23.45	19	0	14.11	3	0	0	0	0	0	12.06	3
TCONS_00001423	5.9	9	0	12.49	5	0	12.39	7	0	0	142.09	57
TCONS_00001433	13.89	13	2	9.49	2	1	0	0	3.57	1	0	0
TCONS_00001442	8.51	27	0	6.01	5	0	0	0	0	0	0	0
TCONS_00001446	23.31	7	0	126.91	10	0	0	0	0	0	0	0
TCONS_00001454	5.1	12	0	17.79	11	0	0	0	0	0	13.32	9
TCONS_00001460	29.15	10	4	16.57	2	0	0	0	0	0	51.52	6
TCONS_00001479	28.88	12	0	0	0	0	0	0	0	0	0	0
TCONS_00001480	26.83	47	1	4.23	2	0	0	0	0	0	0	0
TCONS_00001482	64.27	28	1	42.46	5	0	0	0	0	0	9.58	1
TCONS_00001483	79.15	22	0	27.43	2	0	0	0	0	0	0	0
TCONS_00001485	22.63	39	0	0	0	0	0	0	0	0	0	0
TCONS_00001489	63.27	19	0	25.39	2	0	0	0	0	0	0	0
TCONS_00001497	11.33	11	0	31.41	8	0	0	0	0	0	7.79	2
TCONS_00001503	15.75	10	3	13.63	3	0	2.76	1	0	0	12.89	3
TCONS_00001505	14.62	23	0	36.33	15	0	0	0	0	0	17.14	7
TCONS_00001513	2.8	7	0	1.38	0	1	0	0	0	0	54.23	36
TCONS_00001777	6.77	2	0	25.78	2	0	0	0	0	0	386.7	30
TCONS_00001784	2.98	10	2	4.28	5	0	15.73	22	0	0	8.72	10
TCONS_00001790	40.94	10	0	15.61	1	0	0	0	0	0	35.19	2

TCONS_00001797	8.43	16	0	20.08	10	0	0	0	0	0	0	0
TCONS_00001798	61.38	34	3	19.06	3	0	0	0	0	0	18.03	3
TCONS_00001799	17.65	44	12	5.19	2	3	0	0	0	0	29.06	28
TCONS_00001800	23.13	27	8	8.57	2	2	3.37	2	2.69	1	6	3
TCONS_00001801	10.56	12	11	10.44	1	7	0	0	16.1	9	29.13	21
TCONS_00001828	14.05	22	0	0	0	0	0	0	0	0	4.83	2
TCONS_00001831	43.66	14	1	0	0	0	0	0	0	0	0	0
TCONS_00001841	48.34	10	4	28.77	2	0	6.74	1	10.01	1	85.53	6
TCONS_00001847	37.17	27	0	15.74	3	0	0	0	0	0	5.92	1
TCONS_00001853	0.63	1	0	0	0	0	0	0	27.66	14	0	0
TCONS_00001860	0	0	0	3.7	0	1	0	0	0	0	126.74	22
TCONS_00001863	9.72	13	0	17.09	6	0	0	0	0	0	43.51	15
TCONS_00001872	17.25	14	2	11.01	2	1	7.61	3	15.98	4	91.63	22
TCONS_00001874	53.37	348	12	6.72	12	0	2.42	5	3.94	5	29.68	55
TCONS_00001884	4.59	10	0	5.24	3	0	0	0	0	0	64.33	38
TCONS_00001886	12.02	15	1	5.63	2	0	0	0	0	0	3.18	1
TCONS_00001900	27.17	17	0	6.09	1	0	0	0	0	0	5.21	1
TCONS_00001901	20.64	22	2	31.14	2	10	2.15	1	6.8	2	253.56	79
TCONS_00001904	10.82	21	1	21.84	11	1	0	0	2.23	1	1.58	1
TCONS_00001905	7.25	18	0	26.09	17	0	0	0	0	0	1.74	1
TCONS_00001916	6.98	2	0	13.29	1	0	0	0	0	0	274.82	21
TCONS_00001917	0	0	0	37.78	1	7	23.13	6	77.3	13	0	0
TCONS_00001920	15.89	11	0	0	0	0	0	0	0	0	6.21	1
TCONS_00001926	16.83	14	0	0	0	0	0	0	0	0	0	0
TCONS_00001953	43.48	15	1	10.5	1	0	0	0	0	0	0	0
TCONS_00001978	19.29	54	3	8.78	7	0	1.01	1	1.63	1	4.29	4
TCONS_00001981	15.66	26	0	13.77	6	0	0	0	0	0	27.3	12
TCONS_00001986	26.37	79	1	21.95	12	6	1	1	9.69	6	16.25	13
TCONS_00002000	6.36	11	0	8.46	3	1	0	0	0	0	48.08	23
TCONS_00002005	6.99	11	0	58.12	24	0	0	0	0	0	2.74	1
TCONS_00002021	37.63	10	2	38.35	2	2	5.98	1	8.96	1	88.89	8
TCONS_00002025	5.65	15	8	9.81	4	9	41.81	65	3.14	3	67.6	92
TCONS_00002028	10.48	15	1	12.24	5	0	0	0	5.56	2	0	0
TCONS_00002066	21.88	15	0	27.8	5	0	0	0	0	0	9.51	2
TCONS_00002067	124.61	48	3	80.67	8	1	0	0	14.47	2	29.29	3
TCONS_00002068	286.7	113	0	183.73	19	0	9.7	2	44.25	6	104.12	11
TCONS_00002070	27.42	12	0	60.97	7	0	0	0	0	0	0	0
TCONS_00002077	15.83	18	0	15.95	4	1	0	0	7	2	58.89	18
TCONS_00002081	4.41	2	1	0	0	0	0	0	0	0	230.46	41

TCONS_00002088	46.77	13	0	41.14	3	0	0	0	0	0	0	0
TCONS_00002091	33.67	18	1	6.79	1	0	3.7	1	28.57	5	28.74	4
TCONS_00002098	12.48	14	0	16.99	5	0	0	0	0	0	60.5	17
TCONS_00002101	22.22	10	2	53.37	5	4	15.42	4	47.57	8	0	0
TCONS_00002107	16.35	21	1	21.82	7	1	0	0	0	0	0	0
TCONS_00002112	13.45	5	5	21.97	3	2	32.76	11	4.66	1	31.62	7
TCONS_00002113	5.41	30	1	3.25	4	1	1.12	2	0.91	1	1.29	2
TCONS_00002114	10.75	19	1	9.76	4	1	0	0	0	0	6.8	3
TCONS_00002117	23.86	34	1	48.47	18	1	0	0	2.89	1	164.85	64
TCONS_00002122	17.37	32	0	22.76	11	0	0	0	0	0	54.44	26
TCONS_00002125	14.43	15	4	7.12	1	2	45.54	24	9.06	3	7.77	3
TCONS_00002127	15.65	6	0	109.29	11	0	0	0	0	0	11.21	1
TCONS_00002133	11	18	0	9.32	4	0	0	0	0	0	2	1
TCONS_00002142	8.02	13	0	4.71	2	0	0	0	0	0	0	0
TCONS_00002150	19.8	11	0	34.29	5	0	0	0	5.76	1	0	0
TCONS_00002161	15.82	15	0	0	0	0	0	0	0	0	9.07	2
TCONS_00002162	105.58	62	0	0	0	0	0	0	0	0	5.55	1
TCONS_00002167	0	0	0	13.93	1	0	281.71	43	0	0	0	0
TCONS_00002249	31.57	5	18	6.9	0	1	3.02	1	0	0	0	0
TCONS_00002382	5.86	12	0	1.87	1	0	0	0	2.25	1	0	0
TCONS_00002383	30.58	27	1	15.4	3	1	0	0	0	0	0	0
TCONS_00002387	9.42	11	0	22.08	6	1	0	0	0	0	0	0
TCONS_00002392	10.88	28	1	5.62	4	0	0	0	0	0	0	0
TCONS_00002395	68.46	17	0	15.35	1	0	0	0	0	0	0	0
TCONS_00002399	14.64	18	0	15.5	5	0	0	0	0	0	0	0
TCONS_00002402	24.8	17	0	27.8	5	0	0	0	0	0	106.97	19
TCONS_00002407	10.56	12	2	4.42	0	2	9.57	5	3.05	1	2.37	1
TCONS_00002413	47.49	17	0	48.47	4	1	0	0	7.93	1	0	0
TCONS_00002421	5.51	12	2	20.95	15	0	0	0	0	0	17.37	12
TCONS_00002423	26.69	43	14	2.88	0	2	1.24	1	0	0	5.48	4
TCONS_00002430	11.86	10	2	0	0	0	2.37	1	0	0	7.28	2
TCONS_00002437	17.86	18	0	15.13	4	0	0	0	0	0	11.76	3
TCONS_00002445	14.49	42	24	9.39	9	5	1.15	2	0	0	0	0
TCONS_00002446	18.34	4	11	6.17	0	2	5.38	2	0	0	0	0
TCONS_00002449	25.16	12	2	0	0	0	18.79	5	5.81	1	41.25	6
TCONS_00002454	66.84	18	0	28.31	2	0	0	0	0	0	0	0
TCONS_00002484	14.3	56	1	19.87	21	0	0	0	0	0	33.5	36
TCONS_00002485	21.62	14	4	21.16	4	1	0	0	0	0	43.25	9
TCONS_00002491	33.99	11	0	11.78	1	0	0	0	0	0	10.07	1

TCONS_00002512	13.56	15	0	23.26	6	1	0	0	0	0	30.26	9
TCONS_00002517	4.89	14	0	6.56	4	1	1.06	1	5.12	3	4.14	3
TCONS_00002533	7.25	13	0	12.74	6	0	0	0	0	0	35.99	16
TCONS_00002538	2.95	2	0	0	0	0	0	0	5.01	1	109.56	19
TCONS_00002540	14.32	14	0	50.65	13	0	0	0	0	0	19.84	5
TCONS_00002553	45.12	21	3	58.07	8	0	0	0	6	1	65.75	9
TCONS_00002555	28.25	15	0	0	0	0	7.71	2	0	0	26.5	4
TCONS_00002558	55.76	17	0	37.5	3	0	0	0	0	0	14.1	1
TCONS_00002559	6.28	4	2	0	0	0	2.43	1	0	0	61.31	15
TCONS_00002589	9.84	3	0	0	0	0	11.96	2	0	0	378.97	31
TCONS_00002592	31.93	31	0	0	0	0	4.96	2	3.91	1	12.21	3
TCONS_00002596	28.17	18	0	0	0	0	0	0	0	0	0	0
TCONS_00002617	9.49	11	0	8.3	1	2	6.53	3	3.45	1	0	0
TCONS_00002645	6.37	2	0	24.26	2	0	0	0	0	0	800.15	66
TCONS_00002648	19.29	13	1	27.98	4	2	3.06	1	4.78	1	50.57	10
TCONS_00002651	50.35	18	1	10.22	1	0	0	0	0	0	20.25	2
TCONS_00002659	12.37	11	0	29.99	7	0	0	0	0	0	8.5	2
TCONS_00002676	22.47	32	0	42.82	16	0	0	0	2.97	1	9.88	4
TCONS_00002698	21.44	45	4	27.14	16	1	0	0	0	0	3.63	2
TCONS_00002699	32.28	22	1	42.78	8	0	0	0	0	0	10.61	2
TCONS_00002706	4.79	26	1	11.19	14	3	7.88	14	21.96	24	6.17	10
TCONS_00002711	19.55	85	1	11.18	13	0	1.44	2	1.17	1	24.84	29
TCONS_00002724	13.08	58	2	17.08	21	0	0	0	2.22	2	10.34	12
TCONS_00002740	23.97	15	0	48.72	8	0	0	0	0	0	0	0
TCONS_00002741	25.55	15	0	10.56	1	1	0	0	0	0	5.55	1
TCONS_00002779	1.97	1	3	4.04	3	0	19.22	18	0	0	6.49	3
TCONS_00002943	4.55	52	2	26.11	50	12	3.57	9	3.88	5	4.09	14
TCONS_00002944	6.97	52	2	2.12	50	12	0	9	0	5	7.92	14
TCONS_00002945	60.16	25	0	45.86	5	0	0	0	0	0	44.21	5
TCONS_00002946	14.99	15	0	22.86	6	0	0	0	7.66	2	0	0
TCONS_00002947	2.97	5	0	6.79	3	0	0	0	2.62	1	1277.09	576
TCONS_00002952	16.06	12	0	5.11	1	0	0	0	0	0	0	0
TCONS_00002962	40.85	11	0	14.16	1	0	0	0	0	0	0	0
TCONS_00002978	70.31	9	13	7.7	0	1	0	0	0	0	33.5	2
TCONS_00002985	31.65	43	2	10.59	4	0	1.85	1	0	0	5.25	2
TCONS_00002990	7	6	10	1.25	1	0	13.97	14	9.67	6	15.83	12
TCONS_00002991	146.88	158	1	112.54	32	0	2.29	1	0	0	12.99	4
TCONS_00002993	22.56	4	10	8.07	0	2	14.19	4	5.5	1	5.49	1
TCONS_00003008	42.14	11	1	82.95	5	2	19.39	3	28.9	3	66.09	5

TCONS_00003015	36.81	24	1	33.69	6	0	0	0	5.01	1	0	0
TCONS_00003018	11.76	14	0	0	0	0	0	0	0	0	6.35	2
TCONS_00003020	5.32	4	1	6.82	1	1	2.5	1	0	0	160.24	41
TCONS_00003036	21.7	11	0	30.07	4	0	0	0	0	0	6.43	1
TCONS_00003047	12.3	28	1	20.67	13	0	1.23	1	0	0	5.44	4
TCONS_00003051	11.83	12	6	6.37	2	1	0	0	0	0	2.53	1
TCONS_00003058	9.53	4	1	7.43	1	0	0	0	0	0	294.31	39
TCONS_00003065	71.61	12	30	36.27	4	2	18.79	5	29.03	5	143.41	21
TCONS_00003068	3.85	8	10	6.47	6	5	11.11	22	0	0	33.07	58
TCONS_00003069	12.49	3	0	31.73	2	0	0	0	0	0	669.12	42
TCONS_00003075	5.49	22	1	1.78	2	0	11.8	16	0	0	5.78	7
TCONS_00003087	6.82	4	0	38.95	6	0	0	0	0	0	1181.24	181
TCONS_00003095	11.34	10	0	0	0	0	0	0	4.18	1	4.88	1
TCONS_00003096	3.38	6	0	4.29	2	0	0	0	0	0	0	0
TCONS_00003114	17.17	12	0	3.6	0	1	0	0	0	0	44.72	8
TCONS_00003115	31.86	10	1	21	2	0	5.18	1	15.68	2	11.84	1
TCONS_00003124	2.96	1	0	11.28	1	0	0	0	0	0	613.36	56
TCONS_00003132	2.02	2	0	15.35	4	0	0	0	0	0	226.13	60
TCONS_00003151	13.46	11	0	7.88	1	1	0	0	17.62	4	0	0
TCONS_00003159	30.81	14	0	8.39	1	0	0	0	0	0	0	0
TCONS_00003162	14.46	70	16	3.98	5	2	0.49	1	0	0	0	0
TCONS_00003164	9.78	10	1	6.66	2	0	0	0	0	0	0	0
TCONS_00003168	18.4	58	16	4.91	1	5	13.26	19	0	0	15.84	20
TCONS_00003171	39.07	15	1	9.42	1	0	0	0	14.47	2	18.67	2
TCONS_00003180	10.74	22	0	9.31	5	0	0	0	0	0	9.99	5
TCONS_00003183	4.19	1	3	0	0	0	23.01	10	0	0	86.58	27
TCONS_00003186	5.67	5	0	17.78	2	3	15.95	6	25.07	6	3864.05	904
TCONS_00003192	10.45	11	0	18.09	5	0	0	0	0	0	0	0
TCONS_00003203	6.34	1	1	0	0	0	0	0	0	0	244.79	18
TCONS_00003206	16.15	38	1	17.15	11	0	0	0	0	0	3.1	2
TCONS_00003219	18.02	29	1	9.06	4	0	0	0	5.24	2	10.29	5
TCONS_00003222	24.97	19	0	50.08	10	0	0	0	4.63	1	0	0
TCONS_00003223	33.49	11	0	0	0	0	0	0	25.37	3	0	0
TCONS_00003234	24.35	24	0	3.87	1	0	0	0	0	0	25.25	7
TCONS_00003237	8.95	17	2	1.72	1	0	1.31	1	0	0	5.87	4
TCONS_00003238	3.29	3	3	6.84	4	0	11.76	9	0	0	92.45	55
TCONS_00003239	23.18	55	0	38.38	23	1	1.25	1	2	1	75.04	47
TCONS_00003240	8.24	8	0	3.93	1	0	2.48	1	0	0	329.16	85
TCONS_00003247	50.72	15	0	25.78	2	0	0	0	0	0	51.1	4

TCONS_00003248	31.61	61	0	57.01	28	1	0	0	0	0	2.23	1
TCONS_00003249	10.32	11	1	11.39	2	2	0	0	0	0	72.12	23
TCONS_00003250	45.1	10	2	15.1	1	0	0	0	0	0	12.91	1
TCONS_00003255	24.3	62	4	21.67	15	1	26.95	25	5.22	3	62.99	50
TCONS_00003258	8.8	21	2	11.17	8	0	0	0	0	0	0	0
TCONS_00003268	17.66	11	1	11.23	2	0	0	0	0	0	33.39	6
TCONS_00003269	1.67	1	0	0	0	0	0	0	5.46	1	388.18	60
TCONS_00003285	41.77	12	1	12.5	1	0	5.98	1	0	0	0	0
TCONS_00003286	30.85	25	0	9.41	2	0	0	0	0	0	5.31	1
TCONS_00003287	29.65	12	0	28.25	3	0	0	0	0	0	26.72	3
TCONS_00003288	19.8	11	0	0	0	0	0	0	0	0	37.05	6
TCONS_00003293	13.15	24	2	24.28	13	0	0	0	6.77	3	18.52	10
TCONS_00003300	35.01	11	0	0	0	0	0	0	0	0	10.37	1
TCONS_00003315	14.03	14	11	14.65	2	7	5.45	4	15.31	7	20.34	11
TCONS_00003318	13.77	10	0	31.48	6	0	0	0	0	0	35.69	7
TCONS_00003319	2.61	2	0	34.74	7	0	0	0	0	0	528.82	110
TCONS_00003320	7.89	4	0	7.52	1	0	0	0	0	0	148.76	19
TCONS_00003331	7.8	14	1	11.57	6	0	0	0	0	0	6.6	4
TCONS_00003334	0	0	0	0	0	0	0	0	7.42	2	247.63	68
TCONS_00003577	23.92	28	1	21.84	7	0	0	0	3.33	1	0	0
TCONS_00003578	16.07	16	1	6.23	1	1	2.32	1	0	0	6.1	2
TCONS_00003579	11.76	6	3	13.18	2	1	14.6	5	45.7	10	59.83	12
TCONS_00003581	25.95	17	3	14.75	3	0	0	0	0	0	0	0
TCONS_00003585	8.95	14	2	7.85	3	1	12.05	8	24.1	10	151.6	77
TCONS_00003600	10.63	7	0	57.85	10	0	0	0	0	0	13.05	2
TCONS_00003609	38.09	11	2	17.9	1	1	0	0	8.46	1	0	0
TCONS_00003615	32.33	43	1	19.48	7	0	0	0	0	0	7.14	3
TCONS_00003637	5.67	2	0	64.79	6	0	10.6	2	0	0	685.41	65
TCONS_00003646	19.52	1	16	5.67	0	2	2.47	1	3.89	1	47.55	13
TCONS_00003656	9.73	16	0	9.27	4	0	3.34	2	0	0	45.86	19
TCONS_00003659	57.48	17	0	25.78	2	0	0	0	0	0	22.03	2
TCONS_00003667	46.93	18	0	19.87	2	0	0	0	0	0	0	0
TCONS_00003692	28.37	21	0	0	0	0	9.06	3	0	0	5.81	1
TCONS_00003693	27.42	12	0	0	0	0	0	0	0	0	0	0
TCONS_00003694	18.35	27	0	10.37	4	0	0	0	2.9	1	13.2	5
TCONS_00003705	33.81	10	0	25.78	2	0	0	0	9.18	1	120.7	10
TCONS_00003706	49.88	21	0	36.21	4	0	0	0	0	0	379.65	43
TCONS_00003715	29.54	30	0	157.6	42	0	0	0	0	0	3.21	1
TCONS_00003719	16.37	20	0	24.96	8	0	2.1	1	6.65	2	5.34	2

TCONS_00003720	19.06	27	0	29.6	11	0	1.88	1	0	0	26.67	10
TCONS_00003721	14.05	30	1	32.35	19	0	0	0	2.1	1	12.11	8
TCONS_00003728	1.82	1	0	0	0	0	0	0	0	0	196.52	29
TCONS_00003745	14.31	12	0	57.67	12	1	0	0	0	0	0	0
TCONS_00003748	18.71	19	0	11.26	3	0	0	0	3.79	1	0	0
TCONS_00003751	10.11	14	0	22.02	8	0	0	0	0	0	0	0
TCONS_00003752	9.4	28	3	27.31	24	1	0	0	0	0	0	0
TCONS_00003765	17.41	10	0	6.64	1	0	3.64	1	0	0	5.68	1
TCONS_00003767	10.6	30	0	1.35	1	0	0	0	0	0	0	0
TCONS_00003773	15.16	40	2	22.4	13	4	1.07	1	8.59	5	4.18	3
TCONS_00003784	31.94	39	2	19.91	6	1	0	0	19.06	6	80.61	28
TCONS_00003790	94.67	28	0	12.89	1	0	0	0	0	0	11.02	1
TCONS_00003798	10.55	9	0	13.4	3	0	0	0	0	0	0	0
TCONS_00003804	3.17	2	0	0	0	0	0	0	0	0	191.16	32
TCONS_00003806	9.73	22	0	11.8	7	0	0	0	0	0	0	0
TCONS_00003811	18.81	24	0	23.25	7	1	0	0	3.22	1	2.56	1
TCONS_00003813	17.47	38	0	38.34	21	1	0	0	0	0	1.5	1
TCONS_00003814	31.01	23	1	39.34	8	0	0	0	0	0	0	0
TCONS_00003818	38.45	13	0	67.63	6	0	0	0	0	0	0	0
TCONS_00003822	12.53	21	0	6.83	3	0	0	0	0	0	18.03	8
TCONS_00003825	33.99	11	0	0	0	0	0	0	0	0	0	0
TCONS_00003829	14.92	14	1	15.01	4	0	0	0	0	0	0	0
TCONS_00003830	26.16	19	0	12.24	1	2	0	0	0	0	0	0
TCONS_00003836	6.37	43	2	19.32	37	0	0.46	1	1.48	2	4.43	8
TCONS_00003837	7.2	21	0	14.36	11	0	0	0	3.36	2	47.25	35
TCONS_00003847	17	6	0	0	0	0	10.6	2	8.01	1	202.29	20
TCONS_00003851	16.05	48	0	7.65	6	0	0	0	0	0	57.06	46
TCONS_00003852	5.34	26	1	19.85	27	0	0	0	0	0	0.63	1
TCONS_00003855	26.17	13	1	14.36	2	0	0	0	11.9	2	0	0
TCONS_00003858	6.9	15	1	6.39	4	0	0	0	5.96	3	3.6	2
TCONS_00003859	8.25	15	0	16.77	8	0	0	0	0	0	32.67	16
TCONS_00003863	1.35	3	0	5.11	3	0	3.91	3	81.58	39	7.22	4
TCONS_00003864	23.39	13	0	0	0	0	0	0	0	0	5.87	1
TCONS_00003871	20.62	229	10	47.46	149	0	0	0	13.87	30	40.24	128
TCONS_00003872	19.56	18	1	19.48	5	0	2.47	1	0	0	18.78	5
TCONS_00003875	10.07	13	0	20.43	3	5	0	0	0	0	5.85	2
TCONS_00003878	30.73	11	0	37.82	3	1	0	0	15.85	2	0	0
TCONS_00003886	12.54	37	0	18.08	14	0	0	0	6.65	4	26.26	19
TCONS_00003889	7.99	4	0	38.04	5	0	0	0	0	0	379.37	51

TCONS_00003892	15.82	22	2	22.02	9	0	0	0	0	0	0	0
TCONS_00003898	39.58	11	0	0	0	0	0	0	0	0	27.19	2
TCONS_00003902	9.34	22	1	4.55	3	0	0	0	0	0	13.31	9
TCONS_00003904	28.61	13	0	33.55	4	0	0	0	0	0	26.08	3
TCONS_00003907	17.08	13	0	3.39	0	1	0	0	4.63	1	0	0
TCONS_00003911	9.89	4	0	28.25	3	0	0	0	7.24	1	181.18	19
TCONS_00003912	14.1	103	3	17.4	34	1	0	0	0	0	16.94	37
TCONS_00003920	62.59	25	0	9.55	1	0	0	0	0	0	0	0
TCONS_00003947	31.23	13	1	0	0	0	0	0	0	0	9.7	1
TCONS_00003950	46.64	16	0	33.33	3	0	0	0	0	0	9.5	1
TCONS_00003952	27.17	37	0	27.99	10	0	0	0	0	0	11.1	4
TCONS_00003953	22.38	26	3	18.26	4	3	1.97	1	21.86	7	47.99	18
TCONS_00003960	10.72	25	0	14.7	9	0	0	0	0	0	1.85	1
TCONS_00003974	214.09	134	0	176.59	29	0	0	0	0	0	15.62	3
TCONS_00004251	47.49	17	0	42.59	4	0	0	0	0	0	4337.16	414
TCONS_00004252	23.22	28	0	8.76	2	1	0	0	3.36	1	5.41	2
TCONS_00004254	10.02	18	1	57.05	29	0	0	0	0	0	7.27	4
TCONS_00004259	21.59	20	0	98.74	24	0	0	0	12.12	3	12.8	3
TCONS_00004263	5.74	6	1	6.05	2	0	10.23	5	19.49	6	80.51	27
TCONS_00004264	14.59	10	0	25.88	4	1	0	0	9.95	2	11.03	2
TCONS_00004269	14.61	29	8	6.77	5	0	0	0	0	0	1.53	1
TCONS_00004272	29.47	21	0	21.39	4	0	0	0	0	0	0	0
TCONS_00004274	14.98	16	0	0	0	0	0	0	0	0	4.03	1
TCONS_00004281	12.89	11	0	4.47	1	0	0	0	0	0	365.4	82
TCONS_00004294	17.52	14	1	0	0	0	0	0	0	0	0	0
TCONS_00004303	14.8	35	0	3.23	2	0	0	0	0	0	0	0
TCONS_00004310	13.39	13	0	18.55	4	1	2.48	1	46.91	12	0	0
TCONS_00004315	23.13	12	0	44.06	6	0	0	0	0	0	0	0
TCONS_00004320	21.25	22	4	17.75	2	5	4.08	2	16.17	5	31.54	11
TCONS_00004328	23.78	86	0	27.4	26	0	0	0	0	0	14.38	15
TCONS_00004332	11.9	32	0	50.77	34	2	2.24	2	5.4	3	135.03	100
TCONS_00004350	26.43	10	0	45.92	4	1	0	0	7.61	1	8.61	1
TCONS_00004354	47.68	28	0	38.95	6	0	0	0	5.55	1	5.55	1
TCONS_00004363	15.74	11	0	10.91	2	0	0	0	0	0	0	0
TCONS_00004367	21.7	25	0	33.08	10	0	0	0	3.47	1	0	0
TCONS_00004379	35.08	9	0	14.86	1	0	0	0	0	0	341.72	25
TCONS_00004390	23.67	56	1	9.42	6	0	2.44	2	3.92	2	8.42	5
TCONS_00004397	6.87	10	0	10.48	4	0	0	0	2.93	1	13.34	5
TCONS_00004411	53.28	16	0	164.98	13	0	0	0	27.19	3	64.63	5

TCONS_00004418	2.68	1	0	20.43	2	0	0	0	0	0	205.22	20
TCONS_00004422	17.6	25	1	5.1	2	0	0	0	11.46	4	7.24	3
TCONS_00004435	16.05	11	0	9.2	1	1	0	0	0	0	0	0
TCONS_00004437	21.02	18	2	39.56	10	0	2.5	1	0	0	38.14	10
TCONS_00004438	43.28	26	2	23.62	4	0	0	0	0	0	61.96	11
TCONS_00004443	8.76	33	2	18.42	20	0	0	0	3.72	3	16.43	18
TCONS_00004447	8.5	24	0	5.4	4	0	0	0	5.18	3	14.53	11
TCONS_00004449	19.55	42	3	21.65	9	5	1.25	1	5.99	3	9.56	6
TCONS_00004453	7.63	12	0	16.96	7	0	0	0	0	0	44.62	19
TCONS_00004462	13.82	11	0	4.79	1	0	8.59	3	4.49	1	25.68	5
TCONS_00004477	0	0	0	0	0	0	32.62	10	0	0	0	0
TCONS_00004484	19.15	11	0	26.54	4	0	0	0	11.26	2	41.51	7
TCONS_00004489	31.39	22	1	140.32	27	0	0	0	4.75	1	84.03	17
TCONS_00004509	6.29	9	0	26.62	10	0	1.86	1	0	0	41.39	15
TCONS_00004511	26.39	11	1	8.5	1	0	4.39	1	0	0	170.66	20
TCONS_00004525	8.88	3	0	123.99	11	0	0	0	0	0	0	0
TCONS_00004531	22.01	10	0	8.39	1	0	0	0	0	0	9.46	1
TCONS_00004532	44.32	17	0	49.68	5	0	0	0	0	0	11.21	1
TCONS_00004543	44.9	32	0	35.63	6	1	0	0	0	0	0	0
TCONS_00004546	58.64	22	1	29.41	3	0	0	0	0	0	8.38	1
TCONS_00004547	12.45	23	1	15.55	8	0	1.46	1	16.28	7	7.18	4
TCONS_00004550	15.89	12	5	10.13	3	0	0	0	14.06	4	25.84	8
TCONS_00004560	38.97	10	0	89.12	6	0	0	0	0	0	0	0
TCONS_00004561	3.43	3	0	47.91	11	0	0	0	0	0	16.08	4
TCONS_00004575	24.22	15	0	22.38	3	1	0	0	0	0	10.52	2
TCONS_00004580	17.14	24	0	8.17	3	0	0	0	0	0	0	0
TCONS_00004587	14.26	13	0	53.14	12	1	0	0	0	0	0	0
TCONS_00004591	50.06	142	5	56.2	39	5	0	0	1.66	1	8.81	8
TCONS_00004593	35.98	10	0	7.2	0	1	0	0	0	0	0	0
TCONS_00004604	40.97	9	1	32.27	2	0	0	0	21.98	2	82.16	5
TCONS_00004606	28.21	31	0	6.94	2	0	0	0	0	0	3.92	1
TCONS_00004618	3.45	2	0	0	0	0	0	0	11.17	2	326.73	49
TCONS_00004627	20.21	25	0	9.25	3	0	0	0	3.3	1	81.3	28
TCONS_00004630	36.84	26	1	5.2	1	0	0	0	0	0	0	0
TCONS_00004641	40.87	24	0	123.32	19	0	0	0	0	0	12.87	2
TCONS_00004662	37.45	13	1	10.36	1	0	0	0	7.76	1	11.68	1
TCONS_00004664	12.3	27	2	10.8	6	1	0	0	0	0	1.77	1
TCONS_00004677	10.3	18	0	15.27	7	0	0	0	0	0	4.92	2
TCONS_00004679	10.68	11	0	3.7	1	0	0	0	0	0	28.32	8

TCONS_00004680	16.91	27	0	4.78	2	0	0	0	0	0	6.12	3
TCONS_00004681	55.55	24	0	8.83	1	0	0	0	0	0	0	0
TCONS_00004684	73.23	21	0	292.36	22	0	0	0	28.19	3	391.56	30