Correlation of mast cell in gastric cancer, insight study of HRH4 receptor by NGS technologies

A Major Project dissertation submitted

in partial fulfilment of the requirement for the degree of

Master of Technology

In

Bioinformatics

Submitted by

Aniket Shrotriya

(2k12/BIO/02)

Delhi Technological University, Delhi, India

Under the supervision of

Dr. Asmita Das

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

DECLARATION

I, **Aniket Shrotriya**, hereby declare that the work entitled **“Correlation of mast cell in gastric cancer, insight study of HRH4 receptor by NGS technologies.”** has been carried out by me under the guidance of Dr. Asmita Das, in Delhi Technological University, Delhi.

This dissertation is part of partial fulfillment of requirement for the degree of M.Tech in Bioinformatics. This is the original work and has not been submitted for any other degree in any other university.

Aniket Shrotriya

Roll No.: 2K12/BIO/02

ACKNOWLEDGEMENT

I would like to acknowledge my deep sense of gratitude to ***Prof. B. D Malhotra, (Head of Department) Department of Biotechnology, Delhi Technological University, Delhi-110042*** *for giving me an opportunity to study and work in this prestigious Institute.*

I am extremely thankful to my mentor, ***Dr. Asmita Das****,* ***Assistant Professor, Department of Biotechnology, Delhi Technological University, Delhi-110042*** *for her exemplary guidance, monitoring and constant encouragement throughout the M. Tech course. I would also like to thank her for sparing the efforts in compiling the work presented here.*

At last, I am extremely thankful to my parents, family members and friends specially Prateek, Prashant and Kunal whose blessings and support were always with me.

ANIKET SHROTRIYA

Roll No.: 2K12/BIO/02

LIST OF FIGURES

LIST of TABLE

CONTENTS

Study of Correlation of HRH4 receptor in gastric cancer by using NGS technologies

Aniket Shrotriya

Delhi Technological University, Delhi, India

1. ABSTRACT

Gastric cancer is one of the most prevalent type of cancer and is the second most cancer to cause deaths annually. Although there is no perfect treatment of any type of cancer in the world, yet if diagnosed at early stage, can be controlled by surgery, chemotherapy, radiation therapy, or combinations of all which is known as multimodality therapy.

We have used RNA seq (from dnanexus.com) of 10 gastric cancer patient and compare them to healthy individuals by keeping reference of already sequenced human genome from UCSC genome browser. We use softwares like BOWTIE, TopHat, Cufflink package and CummeRbund package in this pipeline to analyze the data. We have analyzed 10 RNA seq which are sequenced by Next Generation Sequencing by comparing them to normal data and find that the expression level of HRH4 is decreased by almost twice fold.

Role of HRH4 in colorectal cancer, breast cancer and gastric cancer was also previously believed and is yet to be confirmed but our study is step finding towards the association. As previously reported HRH4 is linked to tumor progression in colorectal cancer but our results also clearly shows there is down-regulation of HRH4 in gastric malignancies by almost double fold in both the isoforms of HRH4. Although this work is yet to be validated, so that it can also use as marker in gastric cancer diagnosis.

2. Introduction

Gastric cancer is one of the most prevalent cancer that causes numerous deaths around the world. Although there is no perfect treatment of any type of cancer in the world, yet if diagnosed at an early stage, can be controlled by surgery, chemotherapy, radiation therapy, or combinations of all which is known as multimodality therapy. But in Gastric cancer mostly symptoms occur at an advanced stage so it is not easy to diagnose it in early stages. Previous studies have reported that. Mast cell plays important role in various types of cancers via allergies or by mediators secreted by it.

Histamine is most important mediator among all of them which is largely secreted by mast cells and is very well known growth factor for gastrointestinal malignancies. Its effect is largely determined locally by the histamine receptor expression pattern. Histamine receptor H4 (HRH4), the newest member of the histamine receptor family, is positively expressed on the epithelium of the gastrointestinal tract, and its function remains to be elucidated.

Recently, some evidence indicates that HRH4 also plays a role in cell proliferation, both in normal and malignant cells, including hematopoietic progenitor cells [12], breast cancer cells [13], and pancreatic carcinoma cells [14] It is well noted that the decreased expression of HRH4 in colorectal cancers and its correlation with tumour proliferation. Next Generation Sequencing (NGS) is a high throughput technology to study NGS data may be present in many forms, usage depends on aim on the study. Here RNA seq is used for gene expression profiling. NGS data came from technologies like Roche 454, Illumina Genome Analyzer and Applied Biosystems SOLiD platforms. Sequencing technologies such as 454 or the classic capillary electrophoresis approach can be used for large-scale cDNA sequencing. Relating the expression pattern of HRH4 receptor in gastric cancer by comparing the data of cancerous and non-cancerous patients through several computational tools.

Here we aimed to investigate the abnormalities of HRH4 gene in gastric carcinomas (GCs) in humans. We take RNA seq data of 10 samples of cancerous patients & normal sample and compare them with the reference genome. We analyse NGS data by using several softwares like BOWTIE, TopHat, Cufflink package and CummeRbund for visualization in Linux. In this study, we came to know that there is down-regulation of HRH4 receptor as the gastric malignancy proceeds so HRH receptor can be serve as a diagnostic marker in gastric cancer

3. Review of Literature

3.1 Gastric cancer

Cancer which originates in any part of the stomach is known to be gastric cancer. After lung cancer, it is the second most cancers, leading to death and in occurrence, it is forth most which leads to around millions of deaths across the globe (Kamangar, Dores, & Anderson, 2006). The incidence of gastric cancer varies across the globe and doesn't depend on gender and ethnicity. However the mortality rate declined in western countries, but still it is one of the most common problems in the eastern countries of Asia (Jemal et al., 2006). There are many conventional treatments for gastric cancer like chemotherapy, radiation therapy, but it would be best if they are used in combinations. To cure gastric cancer one needs to diagnosed it earliest, only then complete removal by surgical procedure takes place (Ferlay et al., 2010). Although it is very difficult to diagnose it in early stages as the most symptoms occur only in advanced stages. The survival rate of gastric patient is few years and it doesn't improve significantly from past 4 decades. There are several cases where patients develops another tumor even after surgical removal (Yamashita et al., 2011). One of the most prominent characteristic of gastric cancer is heterogeneity and every patient have a distinct genetic and molecular profile (Zheng, Wang, Ajani, & Xie, 2004). Tissues of most of the gastric malignancies are adenocarcinomas which are differentiated into poorly(diffuse) or well differentiated(intestinal) and both have different epidemiological and genetic patterns (Resende, Thiel, Machado, & Ristimäki, 2011; Yasui et al., 1999). Etiologically, association of gastric cancer came from genetic variation and environmental factors and also the accumulation of alterations of genetic and epigenetic profile(Resende et al., 2011). Diet and lifestyle are always an important factor in several cancers, tobacco and obesity sometime associated with(Compare, Rocco, & Nardone, 2010). Some of the common gastric cancer is associated with *Helicobacter pylori* (Bouvard et al., 2009), and increases the risk of gastric cancerup too 80%. *H. pylori* induces generalized mutations and genomic instability in the host DNA(Machado, Figueiredo, Seruca, & Rasmussen, 2010), and this increases diversity in oncogenic mechanisms. So this leads to different routes in the occurrence of malignancy so it is very much difficult to treat it with a particular drug. So therapy based on genetic and molecular profile of the patient is one of the best way for treatment, it basically involves studying the molecular biology of tumor and then target the specific mechanisms with anti-tumor. These therapies specifically inactivate the mechanisms which are crucial for survival of tumor cells rather than normal gastric cells, which increases its benefit and decreasing the side effects. Trastuzumab a monoclonal antibody produced in human which target the extracellular domain of HER2/neu receptor. It recently used in combination of chemotherapy for treating gastric cancer, which is ERBB2 positive and advanced metastatic. (Bang et al., 2010). Up to 30% of gastric tumor responds to this treatment (Grabsch, Sivakumar, Gray, Gabbert, & Müller, 2010; Wainberg et al., 2010). (Arkenau, 2009; Ku & Ilson, 2010). All the therapies till date are based only little information we have regarding oncogenic genes, so their efficacies are under doubt. So better to develop new strategies for therapeutics which specifically targets gastric cancer by understanding all the molecular mechanisms which happens in the initiation and progression of the disease

3.2 Histamine

Histamine is one of the major mediators of acute anaphylactic reactions. This biogenic amine has several other physiological functions. In the gastrointestinal tract, histamine has 3 major functions:

(1) Increase in gastric acid production (Tanaka S et.,al 2002)
(2) Modulation of gastrointestinal motility (Bertaccini G, Coruzzi G 1995) and
(3) Change in secretion of mucosal ion (Kelly S J et.,al 1995, Wang Y Zet.,al 1990)

Most of these were only performed in animal models, whereas replicated these findings in the human had gained very limited success ((Hemedah, Loiacono, Coupar, & Mitchelson, 2001). Decarboxylation of amino acid L-histidine (E.C. 4.1.1.22 or EC. 4.1.1.26) which is stored mostly in mast cells, and also in basophils, results in Histamine production. Previously Histamine, was known as compound which have potent vasoactive properties (Ring et.,al 1979). Also, it was mostly known for regulation in several processes. Histamine is synthesized in mast cell and stored in cytoplasmic granules and release in response to inflammation, gastric acid secretion and neurotransmission. It is synthesized in central nervous system in only a few neurons of the posterior hypothalamus and specifically its tuberomammillary nucleus. These neurons can diffuse to the cerebra and regulate several functions of the brain in mammals like sleep and hormonal secretion (Waldman et al., 1977), control in cardiovascular activities (Stasiewicz and Gabryelewicz, 1979), etc.

The production of this compound is the result of decarboxylation *L*-histidine by using the pyridoxal-5' phosphate-dependent *L*-histidine decarboxylase enzyme (HDC) via a histidine-PLP Schiff base intermediate (Finch and Hicks, 1976). After its release from cytoplasmic granules, it maintains its level by two major metabolic pathways, i.e., Histamine-*N*-methyl transferase and diamine oxidase. For production of monoamine oxidase-B and diamine oxidase,methyl histamine act as substrate. Aldehyde dehydrogenase further oxidized it to methylimidazole acetic acid. Diamine oxidase convert histamine to imidazole in oxidative pathway, and then instantly converted into imidazole-4-acetic acid by aldehyde dehydrogenase (Yatsunami et al., 1994). It is still unclear out of 4 which histamine receptor is expressed in the human gastrointestinal tract. Particularly function and expression of H3R receptor is not confirmed. Its role in inhibition on releasing neurotransmitters from nerves in the intestine has been previously suggested (Lovenberg TW et.,al 2000) but not yet confirmed. Several studies had declared histamine as an effective modulator in many immune functions (Jutel M et.,al 200s2). As gastrointestinal tract represents one of the largest immune organ in the human body, which produced, regulate, possess several mucosal mast cells (Atkins FM 1987) histamine can be a key regulator for immune regulation which mainly depends on mast cells.

3.2.1 Histamine H4 receptor

Histamine H4 receptor has up to 45% homology to H3 receptor and it varies from species to species and 55% in the transmembrane domain

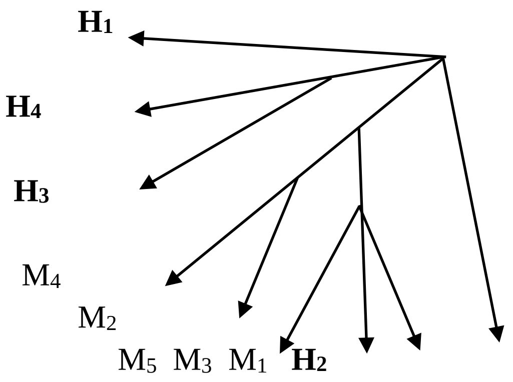

Fig1 -Phylogenetic tree represents the homology between the GPCRs family member (modified from Stark et al., 2003)

	hRH1	hRH2	hRH3	hRH4
Chromosomal Gene location	3p25	5q35.2	20q13.33	18q11.2
Amino acids	487	359	445	390
Isoforms			+	+
G-protein coupling	Gq/11	Gs	Gi/Go	Gi/Go
Principal signal transduction	PLC Ca2+	cAMP	cAMP Ca2+ MAPK	cAMP Ca2+ MAPK
Tissue	Lung, brain, vessels	Heart, stomach, brain	Neurones (CNS, PNS)	Mast cells, eosinophils
Physiological revealance	Contraction of smooth muscles,food intake, sleep-wake regulation	Gastric acid secretion	Sleep, food intake	Chemotaxis
Pathophysiological	Allergic reaction	Gastric ulcer	Cognitive impairment, seizure, metabolic syndrome.	Inflammation, immune reaction

Table 1. Comparison of properties of HRH receptors

3.3 Copy no Variation

Genetic variation in the human genome takes many forms, ranging from large microscopically-visible chromosome anomalies to single nucleotide changes. Several, multiple studies have discovered an abundance of sub-microscopic copy number variation of DNA segments ranging from kilobases (kb) to megabases (Mb) in size ((Iafrate et al., 2004), (Sharp et al., 2005)). Deletions, insertions, duplications, and complex multi-site variants ((Fredman et al., 2004)), collectively termed copy number variants (CNVs) or copy number polymorphisms (CNPs), are found in all humans and other mammals ((Feuk, Carson, & Scherer, 2006))

A CNV can be simple in structure, such as tandem duplication, or may involve complex gains or losses of homologous sequences at multiple sites in the genome. An early association of CNV with a phenotype was described 70 years ago, with the duplication of the Bar gene in Drosophila melanogaster being shown to cause the Bar eye phenotype ((Bridges, 1936)). CNV^{s} influence gene expression, phenotypic variation and adaptation by disrupting genes and altering gene dosage (Buckland, 2003; Nguyen, Webber, & Ponting, 2006; Repping et al., 2006)13-15, and can cause disease, as in micro-deletion or micro-duplication disorders (Inoue & Lupski, 2002; Lupski & Stankiewicz, 2005; Shaw-Smith et al., 2004)16-18, or confer risk to complex disease traits such as HIV-1 infection and glomerulonephritis (Aitman et al., 2006; Gonzalez et al., 2005)19,20. CNVs often represent an appreciable minority of causative alleles of genes at which other types of mutation are strongly associated with specific diseases: CHARGE syndrome, Parkinson and Alzheimer disease (Chartier-Harlin et al., 2004; Jongmans et al., 2006; Rovelet-Lecrux et al., 2005)21 22,23. Furthermore, CNVs can influence gene expression indirectly through position effects, predispose to deleterious genetic changes, or provide substrates for chromosomal change in evolution,(Feuk, Marshall, Wintle, & Scherer, 2006)

3.4 Next-generation sequencing technology

It is a one of the most advanced technology in diagnosing/therapeutics of human cancer and identifying the potential targets (Guan et al., 2012). It is also popularly known as 2nd -generation sequencing, and has relation to capillary-based Sanger sequencing, which was the former and 1st -generation sequencing technology (Mardis, 2011). Cyclic-array sequencing and its various implementations are included in the NGS (Shendure & Ji, 2008). Most used NGS around the globe commercially the 454 Genome Sequencer (Roche Applied Science), Solexa technology (Illumina) and the SOLiD platform (Life Sciences). Despite some changes in technical process, the concept of all is very similar: Firstly the sample DNA fragments produced randomly, followed by in vitro ligation of common adaptor sequences, then a library is produced; after then PCR colonies which are spatially fixed produced; and then at last, by using enzyme biochemical reactions takes place and data processing is performed which is based on imaging in parallel. So for a sample NGS produced millions of reads very efficiently. In the downstream analysis, several short reads mapped to the genome of source which results in reads distribution at nucleotide resolution level and from it we can know about several biological features of the molecules., The NGS applications are of mainly DNA sequencing (DNA-seq), RNA sequencing (RNAseq) and chromatin immunoprecipitation sequencing (ChIP-seq) and it solely depends on the material which is used as input. Previously NGS is used in sequencing genomes of species/individuals people or various tissues and their transcriptome but recently the trends shifts towards studying oncology (Meyerson, Gabriel, & Getz, 2010) and changed it in many ways. Mainly NGS has ability to produce unbiased, comprehensive and, very limited catalogs of various aberrations genomes of the cancer patients. Like, pre-NGS sequencing studies will focus on cancer somatic mutations for few know about genes which are already well known, such as TP53, EFGR and KRAS; and so this study is not easily able to discover novel genes (Meyerson et al., 2010). And also, there are ''analog'' signals in previous sequencing technology, while count of the short reads is the basis of this signaling. This digitalization of signal is can be used for quantification (Z. Wang, Gerstein, & Snyder, 2009). Thus, it's very beneficial to detect biological signals due to high sequence coverage and gather sufficient information in cancer tissues due to heterogeneity. And also, when NGS platforms becomes commercialized, the DNA sequencing cost decreased severely, and reach almost $1,000 per personal genome (Schloss, 2008). Due to this cheap cost clinical applications increased by using information about personal genome. However, in the last 2 years there are very few NGS studies in gastric cancer have been published.

3.4.1 RNA- Sequencing

NGS has several applications in studies which studies of whole genome involved, like single-nucleotide polymorphisms (Auer & Doerge, 2010), epigenetic events (Park 2009), copy number variants (Alkan et al. 2009), differential expression (Bloom et al. 2009), and alternative splicing (Sultan et al. 2008). By use of NGS RNA-Seq can be used to sequence, map and quantify the transcript population. (Mortazavi et al. 2008; Morozova et al. 2009). Although it is a relatively new, yet it can tell about all the transcriptional complexities of an organism (Nagalakshmi et al. 2008), mice (Mortazavi et al. 2008), Arabidopsis (Eveland et al. 2008), and humans (Sultan et al. 2008). There are 3 NGS devices which can be used commercially [Illumina’s Genome Analyzer, Applied Biosystems’ (Foster City, CA) SOLiD, and the 454 Genome Sequencer FLX] for RNA-Seq (Cloonan et al. 2008; Eveland et al. 2008;Marioni et al. 2008). Different platform has same RNA-Seq methodology. Several steps, including RNA isolations, then fragmented randomly, and convert in complementary DNA (cDNA). PCR amplification performed on fragments of specific size like 200–300 bases long. Then by using NGS cDNA is sequenced; then a mapping of the genome of reference, and the reads for particular gene are recorded. These gene counts in the form of digital gene expression (DGE) recorded and can be used to find out the differential gene expression (Morozova et al. 2009). Although there are some loopholes in this process which can lead to an error or bias, yet it is believed to be the future of transcriptome research (Shendure 2008) as it produces an almost unlimited dynamic range so have greater sensitivity than microarray so closely homologous region can be discriminated, and do not require our prior knowledge about the expressed regions (Cloonan et al. 2009; Morozova et al. 2009). As the use of microarrays come in mainstream (Schena et al. 1995), many researchers sought use of proper experimental designed experiments

(Kerr et al. 2000; Lee et al. 2000; Kerr and Churchill 2001a, b; Churchill 2002).

Although In-depth knowledge of molecular biology of gastric cancer and identification of novel therapeutic targets provided by NGS but they have some limitations too. As they are performed on small sample sizes, 30 samples at most, as cost and resources become the limitations. As we know heterogeneity of gastric cancer due to distinct genetic and molecular profile of every cancer patient, so statistics have very limited power to accurately determine the targets of the small sample cohort. As previously, only two candidate driver mutated genes (TP53 and ARID1A) were only simultaneously identified in both studies of sequencing of (Hahn & Weinberg, 2002; K. Wang et al., 2011). Only DNA-seq or RNA-seq based studies used to find out the alterations in the genome

The complexity of gastric cancer involves interactions of different layers of aberrations. By integrating several multi-dimensional profiles of genomes, one can easily understand the effect of the alteration in driver gene function in context of tumor. High throughput functional assays are able to find out the genetic aberrations and its functional consequences (Liang et al.).Size of datasets results from NGS is huge and it is not easy to access. These issues are largely solved by 2 consortium projects. One project is The Cancer Genome Atlas (TCGA), which is funded by the U.S. National Cancer Institute and the National Human Genome Research Institute. They systematically characterize the large sample size almost 500 patients by use of several profiling techniques like miRNA expression profiling, DNA methylation profiling, copy number variation profiling, exome sequencing, SNP arrays and mRNA-seq and RPPA based protein expression. And the access for TCGA generated datasets is user oriented. Glioblastoma is the first cancer to be evaluated by data released from TCGA (McLendon et al., 2008)and ovarian cancer (Network, 2011). In the second phase gastric cancer is done. Another consortium which is doing similar type of work is International Cancer Genome Consortium (ICGC), which have the aim to target at least 50 types of cancer and systematically studying almost 25k cancer genomes at transcriptomic, epigenomic and genomic level. Unlike TCGA, ICGC involves ethnicity as it includes samples from different country. In every selected cancer type the sample of gastric cancer comes from USA and China. It is expected these consortium-based cancer genomic projects can prove themselves as valuable resources in finding several novel targets for gastric cancer in therapeutics over a few years. It is interesting to find out some important points in the application of the NGS to gastric cancer samples. Firstly, it would be interesting to find out the effect of ethnicity on the gastric cancer; thus, it is interesting to determine up to what extent to which the molecular basis of gastric cancer depends on the ethnicity of the individual (e.g. East Asian vs. Caucasian). This topic can be better dealt with the use of data sets obtained from ICGC. Also, not only inter-tumoral heterogeneity between gastric cancers also the intra tumoral heterogeneity which is in single gastric tumor is an important consideration in effective treatment and it can impose drug resistance. NGS can answer it by following these approaches:

(i) In detection of rare clones ultra-deep sequencing of primary tumor;

(ii) In identification of dominate clones by low depth characterization (Network, 2011).
(iii) It is interesting to find out the role of aberrant splicing in gastric tumor as RNA-seq allows the detection of splicing variants
(iv) By using NGS we are able to discovering novel targets by identifying gene fusion
With the accelerated progress in sequencing technologies, the availability of microbial genome data has grown in an accelerated manner. This easy availability has made it easier to elucidate relevant biological processes. But only the nucleotide sequence information does not provide direct information about these processes. To obtain information about these pathways, a need to study the post-genomic process such as transcriptomics was felt. The information about the role and function of factors that are participating in the process can be identified by the transcriptome analysis.

3.5 Transcriptomics and its significance

The genome of an organism contains of all the information that is needed for the metabolic and regulatory processes taking place within a cell. The genome is only a source of information. In order to function, the genome must be expressed. The first step of gene expression is transcription. Transcription is a process by which the genome is transcribed into a complete set of RNA transcripts.

Full set of transcripts produced in a cell and their expression at specific condition. For specific and complete interpretation of the genome and its function and also for understanding disease and its progression understanding transcriptome is very important ((Z. Wang et al., 2009).

The aims to study transcriptomics are

- Find out all transcripts, which includes all types of RNA^s
- Identification of gene structure, of their start sites, at both 5′ and 3′ ends, post-transcriptional modifications and other splicing patterns.

- To quantitate the change in expression levels of all transcripts during the development stage and under different conditions.

Unlike the genome, the transcriptome is extremely dynamic. Most of our cells contain the same genome irrespective of the type of cell, its stage of development or environmental conditions. Conversely, the transcriptome also varies considerably in differing circumstances due to different patterns of gene expression. Transcriptomics is the study of the transcriptome and is therefore a global way of looking at gene expression patterns.

Out of all the transcripts produced in the cell some code for proteins involved in the cellular functions and some are non-coding which are involved in regulatory processes. Non-coding RNAs are predicted to be involved in differential response in stress environment.

3.6 Technique adopted for transcriptome analysis.

To define a precise map of all genes as well as their alternative isoforms and expression in a species is critical to understand the genomics of that species. But to carry out such mapping and analyses is very expensive and experimentally arduous.

Until now, the major methods for annotation of transcriptome included cloning of cDNAs or expressed sequence tag (EST) libraries. Cloning was followed by sequencing, which lead to high cost and limited data yield. It contributes to the high complexity of this method. Sophisticated computational tools are required for the analysis of these data, which can provide the basis for the programs used today for high-throughput RNA sequencing (RNA-seq) data. (Adams et al., 1991)Adams, M.D. et al.,1991, Wu et al., 2005)

Some alternative strategies are present such as genome-wide tiling arrays which allows the identification of transcribed regions at a larger scale. It is cost-efficient scale, but has limited resolution. Splicing arrays with probes across exon- exon junctions enabled researchers to analyze predefined splicing events, but could not be used to identify previously uncharacterized events. Expression quantification required hybridization of RNA for gene-expression microarrays, a process that is limited to studying the expression of known genes for defining isoforms.

High-throughput RNA sequencing (RNA-seq) is used to get a comprehensive picture of the transcriptome. It allows for the complete annotated and quantification of all genes and their isoforms across samples. It requires increasingly complex computational methods. These computational challenges fall into three main categories:

- Read mapping.

- Transcriptome reconstruction.

- Expression quantification.

RNA-seq is a technique which is used to represent the transcriptome revealed by sequencing cDNA through Next Generation Sequencing technologies. With the introduction of such revolutionary method, researchers are now equipped with techniques which are highly sensitive and are able to identify and characterize the organisms' transcriptome. This technique is employed in finding out novel transcripts, mutation identification, INDELS characterization. Excellent coverage provided so a single run is able to generate >600 million reads. (Sorek & Cossart, 2009)

3.7 Advantages and Biases of RNA-seq

The pros of RNA- seq are:-
1. No noise at all.
2. It allows unequivocal mapping of the sequences in a single region of the genome.
3. High coverage of the genome.
4. Large range transcripts can be detected as it can find out many copies of RNA in a single cell.
5. It is relatively cheaper rather than other previously employed methods.
6. With the help of NGS technology, RNA seq not only reduced the errors but also simplified the method of sample preparation and the cloning step is eliminated also this method is very precise in quantifying the results so it replace the use of quantitative PCR(qpcr).

Transcriptome analysis should be free from bias like abundance, their size, these things are basically followed in new generation sequencers. Also, it is the first technology which revolutionizes the research by allowing precise examination of whole the transcriptome in qualitative manner not only rapid but also at very low cost than the previous methods of Sanger sequencing of EST^{s}

Research shows that RNA sequencing produce highly reproducible data used for Illumina's same library, thus, it is necessary to sequence only once. (Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011; Z. Wang et al., 2009)

In detecting the differential expression without bias in the samples one has to be careful in choosing the correct statistical test before analysis so we can get the exact amount of the gene. The results of the analysis of some bacterial transcriptome studies are beginning to re-shape the previous understanding of the complexity of the bacterial transcriptome. This also describes ways in which whole- transcriptome studies are providing insights into functional genomic elements and their regulatory roles in bacteria.

3.8 Steps of RNA-seq analysis

3.8.1 Mapping short RNA-seq reads

One of the most basic and preliminary tasks in RNA-seq analysis is the alignment of reads. The reads are either aligned to either a reference transcriptome or genome. Alignment of reads is a classic problem in Bioinformatics with several solutions specifically for EST mapping(Kent, 2002) (Kent et al.,2002, Wu et al.,2005). RNA-seq reads, however, pose particular challenges because they are short (~36–125 bases), error rates are considerable and many reads span exon-exon junctions.

Additionally, the number of reads per experiment is increasingly large, currently as many as hundreds of millions. There are two major algorithmic approaches to map RNA-seq reads to a reference transcriptome. The first, to which we collectively refer as ‘unspliced read aligners’, align reads to a reference without allowing any large gaps. The unspliced read aligners fall into two main categories, ‘seed methods’ and ‘Burrows-Wheeler transform methods’(Homer, Merriman, & Nelson, 2009). Seed methods find matches for short sub-sequences, termed ‘seeds’, assuming that at least one seed in a read will perfectly match the reference. Each seed is used to narrow candidate regions where more sensitive methods (such as Smith-Waterman) can be applied to extend seeds to full alignments.(Lunter & Goodson, 2011)

In contrast, the second approach includes Burrows-Wheeler transform methods such as Burrows -Wheeler alignment (BWA) and Bowtie, which compact the genome into a data structure that is very efficient when searching for perfect matches. When allowing mismatches, the performance of Burrows-Wheeler transform methods decreases exponentially with the number of mismatches as they iteratively perform perfect searches

3.8.1.1 TopHat for Mapping Reads

We used TopHat for mapping RNA-seq reads onto reference genome. This is a Genome-guided approach. It works by finding splice junctions. Firstly, it map RNA seq reads to the genome, then it find out the regions which are potentially exons in total all the regions which align to the genome. After this initial step, a probable splice site junctions’ database formed and then these reads are mapped
.(Kim & Salzberg, 2011; Trapnell et al., 2012)

3.8.1.2 Basis of Tophat Algorithm

More than one source of evidence is used in generating the probable splice sites by TopHat.
One of the most important evidence for splice junction is if 2 segments of the same read are mapped at a certain distance on the same genomic se**quence**. With this approach, "GT-AG", "GC-AG" and "AT-AC" introns will be found ab initio. The second source is pairings of "coverage islands", which are distinct regions of piled up reads in the initial mapping. Neighboring islands are often spliced together in the transcriptome, so TopHat looks for ways to join these with an intron. We only suggest users use this second option (--coverage-search) for short reads (< 45bp) and with a small number of reads (<= 10 million). This latter option will only report alignments across "GT-AG" introns. (http://tophat.cbcb.umd.edu/manual.shtml)

3.8.1.3 Drawback of Tophat

Along with the advantages of Tophat aligner , there is one major drawback that we encountered during our project. The drawback is that only those reads can be used for mapping which are produced by the Illumina Genome Analyzer.

3.8.2 Transcriptome reconstruction

The process of defining a precise map of all transcripts and isoforms that are expressed in a particular sample which requires the assembly of these reads or read alignments into transcription units is collectively referred as transcriptome reconstruction. Transcriptome reconstruction is a difficult computational task for three main reasons.

First, gene expression spans several orders of magnitude, with some genes represented by only a few reads.

Second, reads originate from the mature mRNA (exons only) as well as from the incompletely spliced precursor mRNA (containing intron sequences), making it difficult to identify the mature transcripts.

Third, reads are short and genes can have many isoforms, making it difficult to identify which isoforms produced each read.

Several methods are present for reconstruction of the transcriptome, and they fall into two main categories:

1. Genome-guided
2. Genome Independent

The genome guided methods rely on a reference genome to first map all the reads to the genome and assemble overlapping reads into transcripts genome-independent methods assemble the reads directly into transcripts without using a reference genome.

3.8.3 Genome-guided reconstruction

Existing genome-guided methods can be classified into two main categories: ‘exon identification’ and ‘genome-guided assembly' approaches. Exon identification methods were developed early when reads were short (~36 bases) and few aligned to exon-exon junctions. They first define putative exons as coverage islands, and then use spliced reads that span across these coverage islands to define exon boundaries and to establish connections between exons. Exon identification methods provided a first approach to solve the transcript reconstruction problem best suitable for short reads, but they are underpowered to identify full-length structures of lowly expressed, long and alternatively spliced genes.

To take advantage of longer read lengths, genome-guided assembly methods such as Cufflinks and Scripture were developed. These methods use spliced reads directly to reconstruct the transcriptome. Scripture initially transforms the genome into a graph topology, which represents all possible connections of bases in the transcriptome either when they occur consecutively or when they are connected by a spliced read. Scripture uses this graph topology to reduce the transcript reconstruction problem to a statistical segmentation problem of identifying significant transcript paths across the graph. Scripture provides increased sensitivity to identify transcripts expressed at low levels by working with significant paths, rather than significant exons. Cufflinks utilizes an approach originally developed for EST assembly, by connecting fragments into a graph if the overlapping fragments agree on their spliced alignment locations. (Gottesman, 2005; Trapnell et al., 2012)

Cufflinks reports the minimal number of compatible isoforms (maximum precision). Cufflinks finds the set of all incompatible assemblies of reads, splice sites does not overlap. These sets of incompatible assemblies represent the minimum possible number of reconstructed isoforms. Since this minimal transcript set is not guaranteed to be unique, coverage level compatibility across the graph is used to decide between minimal sets of transcripts.

Cufflinks constructs a parsimonious set of transcripts that "explain" then reads observed in an RNA-Seq experiment. It does so by reducing the comparative assembly problem to a problem in maximum matching in bipartite graphs. In essence, Cufflinks implements a constructive proof of Dilworth’s Theorem, which states that “the number of mutually incompatible reads is the same as the minimum number of transcripts needed to ‘explain’ all the fragments” by constructing a covering relation on the read alignments, and finding a minimum path cover on the directed acyclic graph of the relation.

The algorithm of cufflinks takes input of cDNA sequences which have been aligned to the genome by TopHat which can produce spliced alignment. Cufflink treats each pair of fragment reads as a single alignment. Overlapped fragments are aligned separately by this algorithm as every bundle contains few genes. It then finds out the frequency of assembled transcripts

The first task is to identify the pairs of fragment which are incompatible and are from different spliced mRNA pairs. These are overlapped and connected to each other when they are compatible and in the whole genome their alignments overlap each other

Between every pair of compatible fragments every fragment placed which is usually had one node in the graph, and an edge, directed from left to right in the genome. (Kim & Salzberg, 2011; Trapnell et al., 2012)

The cufflinks produces following output files:-

Transcripts. gtf

Cufflinks' isoforms are assembled in the form of GTF file as output. The first 7 columns are standard GTF, and the last column contains attributes, some of which are also standardized ("gene_id", & "transcript_id"). There one GTF record per row, and each record represents either a transcript or an exon within a transcript.

isoforms.fpkm_tracking

This file contains the estimated isoform-level expression values in the generic FPKM Tracking Format.

genes.fpkm_tracking

This file contains the estimated gene-level expression values in the generic FPKM Tracking Format. We got transcripts. gtf containing the final transcripts which are needed in the next step. (http://cufflinks.cbcb.umd.edu/manual.html)

Analysis of the transcripts

The transcripts are combined from all the files and unique transcripts are taken into a single file. The transcripts are combined in parallel with the help of Cuffmerge and Cuffcompare. Both of these methods use different basis for merging of transcripts and removes transcripts which are present in more than one file. From the output of these methods, we obtained unique transcripts. Cuffcompare takes Cufflinks' GTF output as input, and optionally can take a "reference" annotation.

3.9 Differential analysis with Cuffdiff

Cuffdiff, which is also included in Cufflinks is used in calculating expression in samples and find out statistical significance of every change in expression between them. Changes are evaluated by the statistical model which assumes that abundance of every transcript is the basis of the number of reads production, but it can be varied due to some bias which are introduced during Despite its exceptional overall accuracy, RNA-seq, like all other assays for gene expression, has sources of bias. These biases have been shown to depend greatly on library preparation protocol. Cufflinks and Cuffdiff can automatically model and subtract a large fraction of the bias in RNA-seq read distribution across each transcript, thereby improving abundance estimates. Although RNA-seq is often noted to have substantially less technical variability than other gene expression assays (e.g., micro-arrays), biological variability will persist. Cuffdiff allows you to supply multiple technical or biological replicate sequencing libraries per condition. With multiple replicates, Cuffdiff learns how read counts vary for each gene across the replicates and uses these variance estimates to calculate the significance of observed changes in expression. We strongly recommend that RNA-seq experiments be designed in replicate to control for batch effects such as variation in culture conditions.

Advances in multiplexing techniques during sequencing now make it possible to divide sequencing output among replicates without increasing total sequencing depth (and thus cost of sequencing).

Cuffdiff reports numerous output files containing the results of its differential analysis of the samples. Gene and transcript expression level changes are reported in simple tabular output files that can be viewed with any spreadsheet application (such as Microsoft Excel). These files contain familiar statistics such as fold change (in log2 scale), P values (both raw and corrected for multiple testing) and gene- and transcript-related attributes such as a common name and location in the genome.

Cuffdiff also reports additional differential analysis results beyond simple changes in gene expression. The program can identify genes that are differentially spliced or differentially regulated via promoter switching. The software groups together isoforms of a gene that have the same TSS. These TSS groups represent isoforms that are all derived from the same premRNA; accordingly, changes in abundance relative to one another reflect differential splicing of their common pre-mRNA. Cuffdiff also calculates the total expression level of a TSS group by adding up the expression levels of the isoforms within it. When a gene has multiple TSSs, Cuffdiff looks for changes in relative abundance between them, which reflect changes in TSS (and thus promoter) preference between conditions. The statistics used to evaluate significance of changes within and between TSS groupings are somewhat different from those used to assess simple expression level changes of a given transcript or gene.

3.10 Visualization with CummeRbund

Regulation and differential expression of a gene or transcript are analyzed by the Cuffdiff. These results can be opened and viewed in a spreadsheet. These files-formats made in a way so that its use is simplified in downstream. Still, these files are not viewable with eye and working with the multiple files across the platform is very difficult, like finding out the genes which are differentially expressed across the sample, but finding out the relative expression level of the different isoforms and plot them is not as easy task

CummeRbund is user-friendly tool, which can manage, visualize and can integrate the output data produced from Cuffdiff analysis. It can drastically

4. Methodology

We needed RNA-seq reads of our organism and reference genetic sequence of the same organism to proceed our analysis.

1) The Sequence Read Archive (SRA) stores raw sequencing data from next-generation sequencing platforms, including Applied Biosystems SOLiD® System, Complete Genomics®, Helicos Heliscope®, llumina Genome Analyzer®, Pacific Biosciences SMRT®, and Roche 454 GS System®. The SRA is the single best resource for useful data from initiatives such as the 1,000 Genomes Project and institutions like the Broad Institute,Washington University, and the Wellcome Trust Sanger Institute. We obtained sequence reads from http://sra.dnanexus.com for our species.

2) Reference genome data can be downloaded from any genome browser like NCBI, UCSC etc. The file should be in .fasta format.

3) Index files were created for the reference genome by using bowtie- build binaries of Bowtie aligner packages which can be obtained from
http://sourceforge.net/projects /bowtie-bio/files/bowtie2/

4) All sra files were converted into a readable fastq file format using fastqdump obtained from Sra toolkit downloaded from
http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi\?cmd=show&f=software&m=software&s=software
These files are to be used as input in Tophat aligner in the next step.

5) The alignment was done with the help of Tophat aligner. Tophat was run for each RNA-seq file using.fastq files and reference genome index files as input. From this we obtained output in bam file format with filename acceptedhits.bam which contains map reads against reference genome.

6) Converted output .bam file format or binary format into .sam file format which is Sequence Alignment Map for using it as input in cufflinks which is the next step.
This was done using samtools package downloaded from
http://sourceforge.net/projects/samtools/files/samtools/
As well as installing this package from the Ubuntu software repository from the local machine terminal.

7). Now acceptedhits.sam is taken as input in cufflinks which is a program that assembles aligned RNA-Seq reads into transcripts, estimates their abundances, and tests for differential expression and regulation transcriptome-wide. Cufflinks runs on Linux and OS X.

8). Then we proceed to Cuffmerge for assembling all the output files of cufflinks. Cuffmerge many GTF (Gene transfer format) files which are output of Cufflinks'. Input GTF files are specified in a "manifest" file listing full paths to the files. Cuffmerge produces a GTF file that contains an assembly that merges together the input assemblies.

9). Transcriptome assembly merged and Cuffdiff run along with BAM files for the output of TopHat of every replicate

10). Open the R and download the CummeRbund

11). A CummeRbund database formed by using the output from Cuffdiff

12). The distribution of expression levels for each sample is plotted

13). Each gene expression is compared in different conditions in a scattered plot

14). A volcano plot can be created to inspect the genes which are differentially expressed.

15). Bar plots of expression of interested genes are plotted

16). Expression levels of every isoform of the gene of interest with bar created

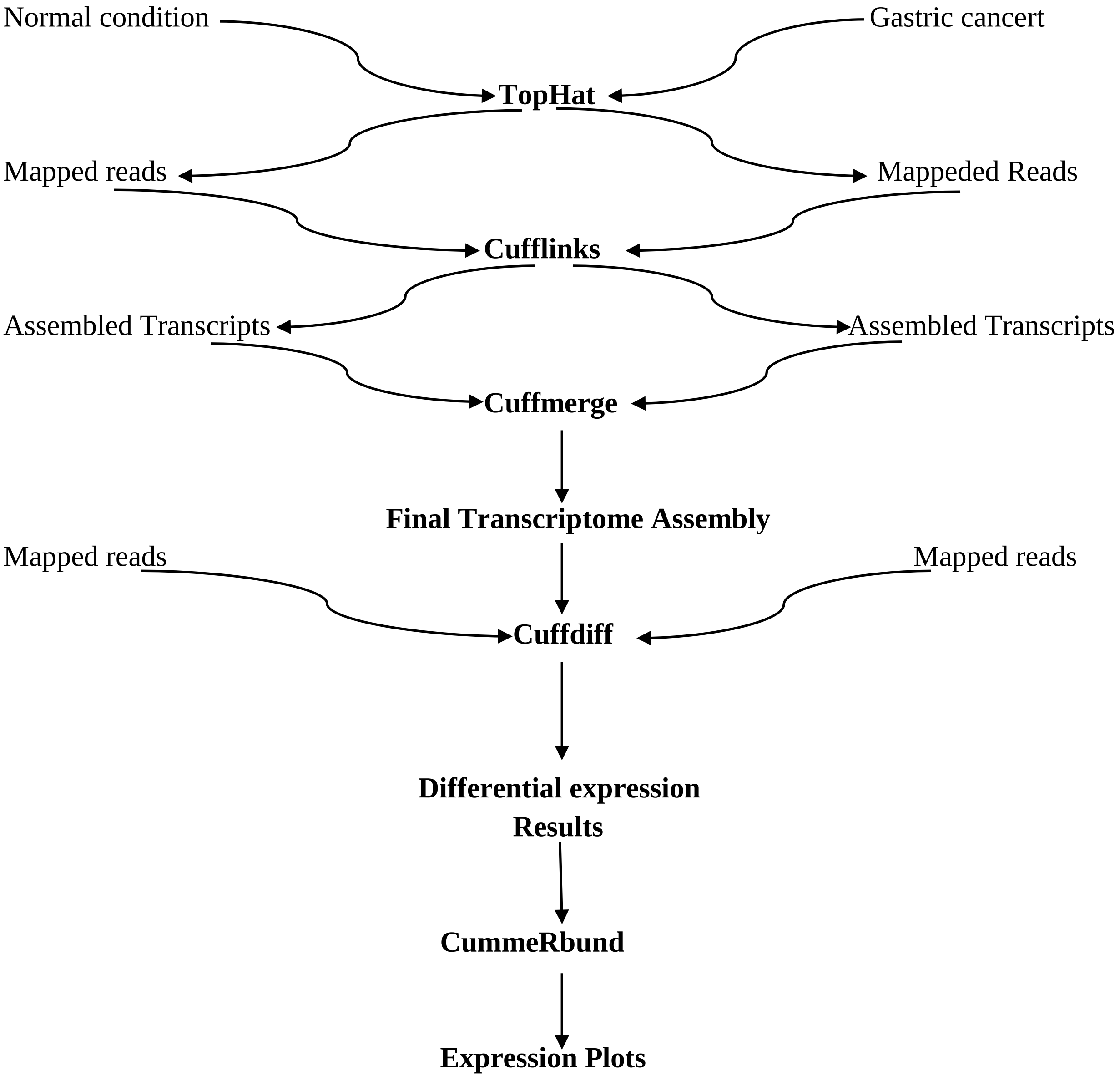

Fig2.Overview of the Methodology

4. RESULTS AND DISCUSSIONS

We have chosen the pipeline which can able to handle the huge RNA seq data efficiently, these include the TopHat pipeline mapped the RNA-Seq reads against the whole reference genome, and those reads that do not map are set aside. Sequences flanking potential donor/acceptor splice sites within neighbouring regions are joined to form potential splice junctions. The Initial Unmapped (IUM) reads are indexed and aligned to these splice junction sequences. TopHat captured around 80% of splice junctions in more actively transcribed genes.

CummeRbund package produces bar plot and sequence coverage in 2 different conditions. We have taken FPKM value in consideration in comparing the differential expression. FPKM is fragments per kilobase of transcript per million mapped fragments, using it Bar graph are plotted. A linear statistical model which find out the abundance of every transcript is used by both Cufflinks and Cuffdiff and it also defines the reads with most probable likelihood.

Cufflinks count the reads to find out the exact expression for every transcript, the reads were counted by Cufflinks which are map to every transcript and then there was normalization of this number to the length. In the same manner same library can produce different amount of sequencing reads in two sequencing runs. FPKM is the method of normalizing the total yield developed for comparing the expression of the transcripts between different runs.

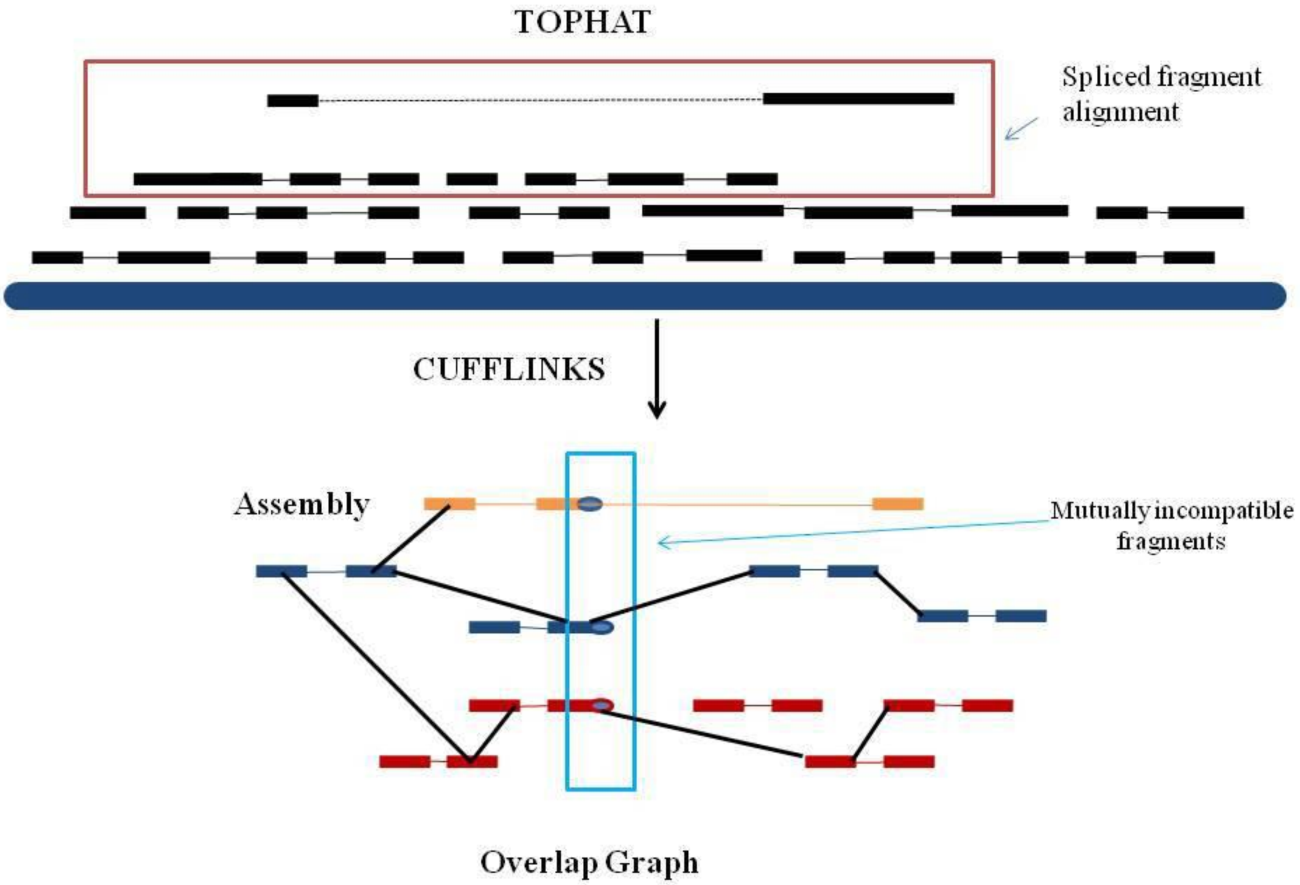

Fig 3. Fragments overlapped and connected in the graph

The algorithm takes as input cDNA fragment sequences that have been aligned to the genome by software capable of producing spliced alignments, such as TopHat. With paired-end RNA-Seq, Cufflinks treats each pair of fragment reads as a single alignment. The algorithm assembles overlapping 'bundles' of fragment alignments separately, which reduces running time and memory use, because each bundle typically contains the fragments from no more than a few genes. Cufflinks then estimates the abundances of the assembled transcripts .The first step in fragment assembly is to identify pairs of 'incompatible' fragments that must have originated from distinct spliced mRNA isoforms. Fragments are connected in an 'overlap graph' when they are compatible and their alignments overlap in the genome.

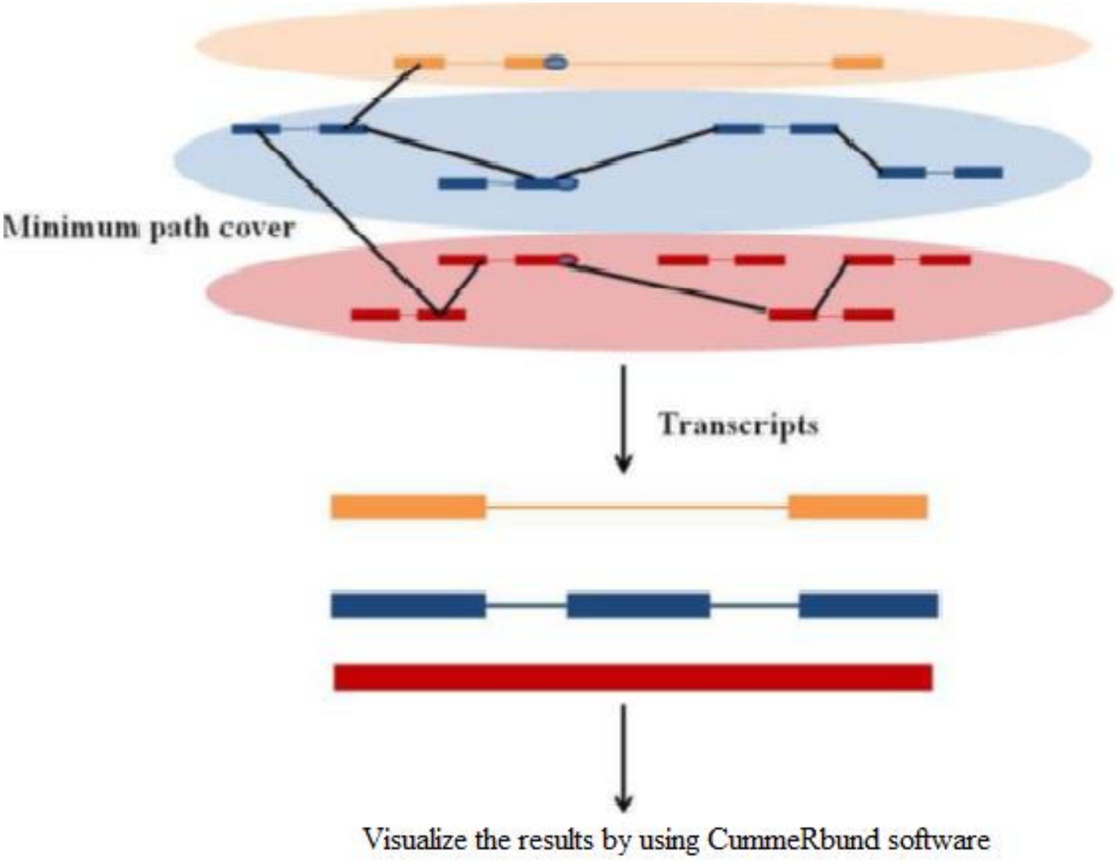

Fig 4. Cufflinks methodology

Fig 4 shows the method by which cufflinks produces transcripts. Each fragment has one node in the graph, and an edge, directed from left to right along the genome, is placed between each pair of compatible fragments. In this figure, the orange, blue and red fragments must have originated from separate isoforms, but any other fragment could have come from the same transcript as one of these three. Isoforms are then assembled from the overlap graph. Paths through the graph correspond to sets of mutually compatible fragments that could be merged into complete isoforms. The overlap graph here can be minimally 'covered' by three paths, each representing a different isoform. Dilworth's Theorem states that the number of mutually incompatible reads is the same as the minimum number of transcripts needed to 'explain' all the fragments. Cufflinks implements a proof of Dilworth's Theorem that produces a minimal set of paths that cover all the fragments in the overlap graph by finding the largest set of reads with the property that no two fragments could have originated from the same isoform

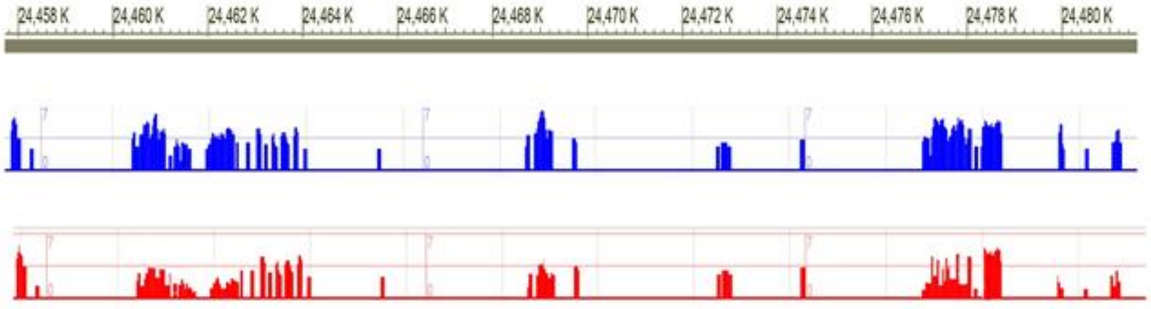

Fig 5. Comparison of expression of HRH4 between cancerous and normal conditions. Here blue colour pattern is of normal sample and cancerous sample represented by red colour which clearly shows significant decrease in the expression pattern than the previous one.

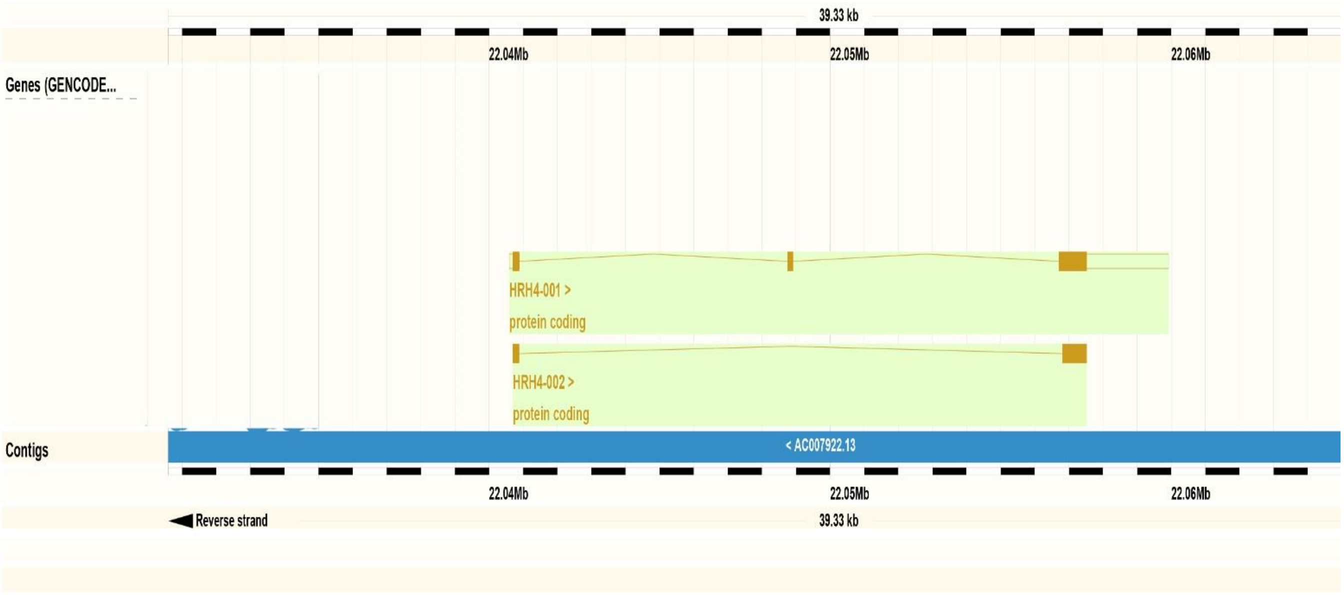

Fig 6. Comparison of both the isoforms of HRH-001 and HRH-002 and their exons and their alignment with the contigs produced in the genome

Chemotactic property of histamine binding to H4R, enhancing leucocyte migration and recruitment from bone marrow. This observation could tempt speculation that H4R are down regulated after migration of leucocytes from the blood stream into the tissue.

Here healthy normal count are represented as condition C1 and Gastric cancerous patients are C2. We have compared the two isoforms of the gene in two different conditions. Expression of a transcript is proportional to the number of reads sequenced from that transcript after normalizing for that transcript's length. Each gene and transcript expression value is annotated with error bars that capture both cross-replicate variability and measurement uncertainty as estimated by Cuffdiff's statistical model of RNA-seq. Changes in HRH4 expression are attributable to a large decrease in the expression of one of two alternative isoforms. The read coverage, can also be viewed through the genome browsing application IGV.

Isoforms of a gene are compared between the healthy individual and the gastric cancer patient. Comparison of isoforms are done as changes in structure may lead to behavioural changes towards the gastric cancer. Although we didn’t find remarkable changes in both the isoforms of HRH4 but it is proposed that a study relating these isoforms should be done with large sample size to come on definite conclusion.

HRH4 possess 2 isoforms which are quite similar but many gene may have several isoforms which can show quite distinct behaviour

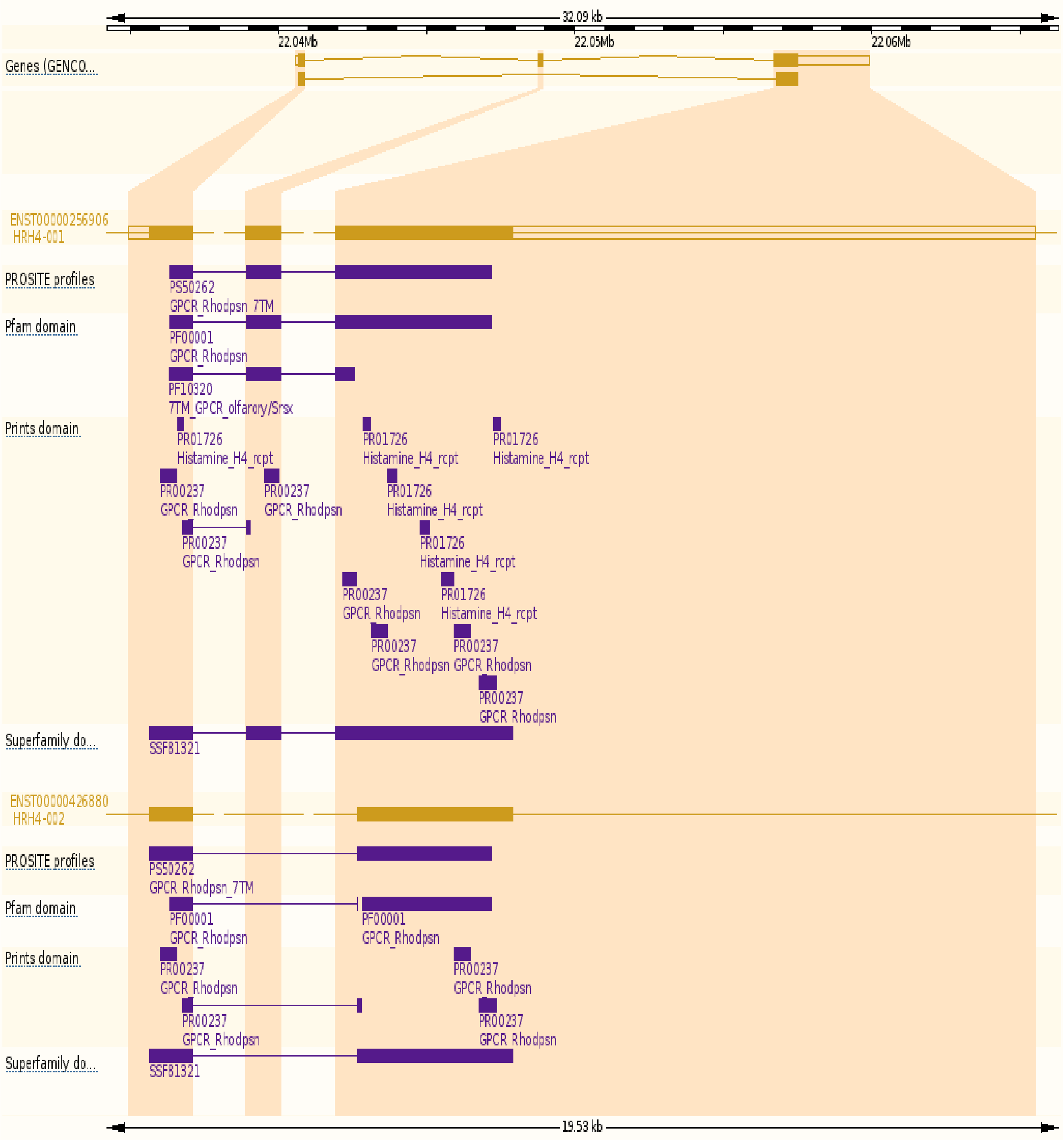

Fig 7.Difference in splice junctions of both the isoforms of HRH4; splice junctions can be easily detected by TopHat

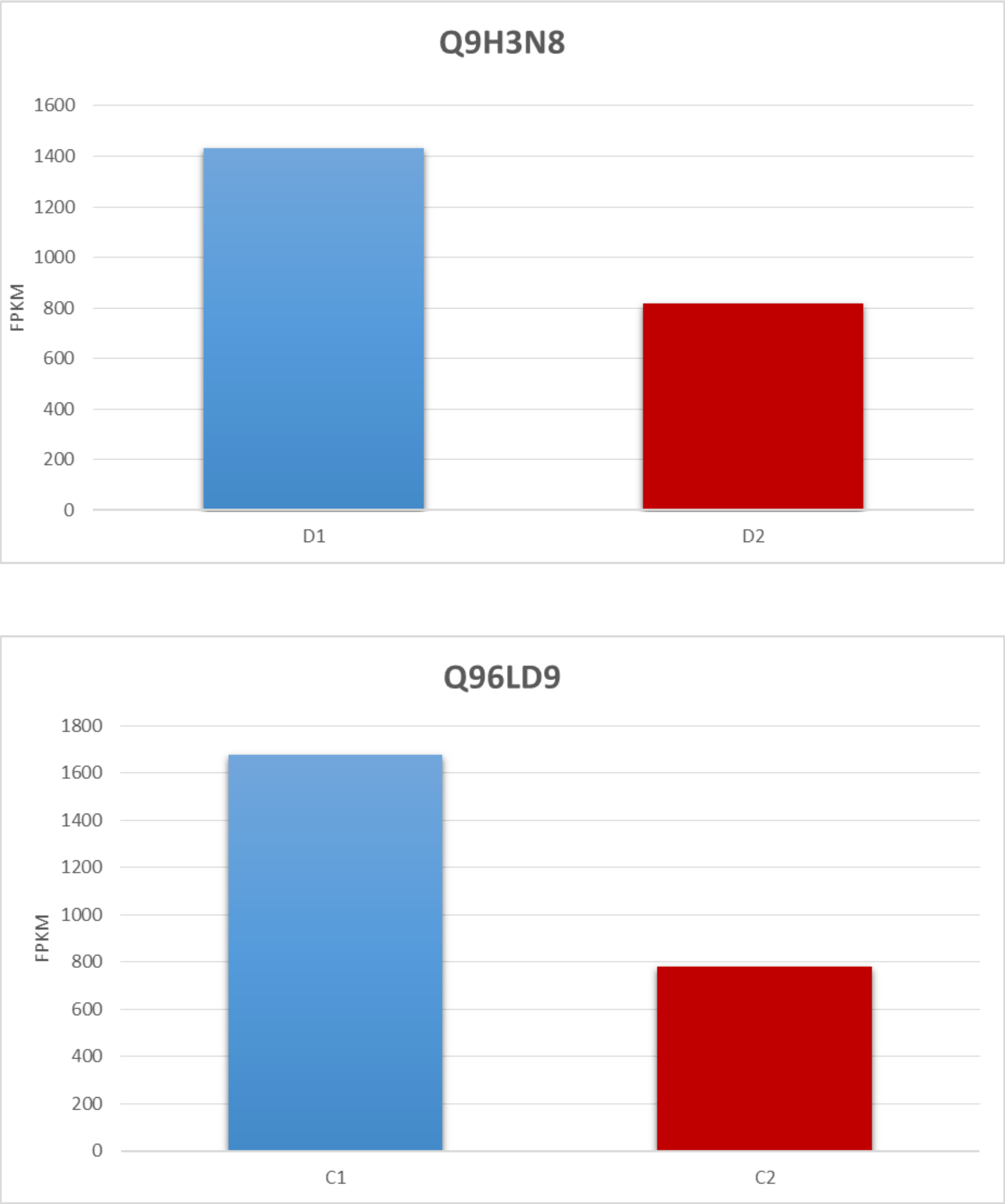

Fig.8 Bar plot showing comparison of both the isoforms (Q9H3N8 & Q96LD9) in normal and cancerous condition

Also there is a copy number variation between the healthy individuals and the cancerous patients. Till now little is known about the how the expression of HRH4 is regulated. Role of DNA methylation is not important in the progress as we do not detect CpG island in the proximal region HRH4 promoter. DNA deletion at chromosome position 18q11, also the chromosome locus that HRH4 gene resides, is frequent in gastrointestinal cancers. We wondered if copy number variations (CNVs) of HRH4 gene might play a role in the regulation of gene expression.

6. Conclusion

Out of all the receptors present on mast cell HRH4 is the newest and the most significant one of them. Mast cell can regulate growth of cell cycle through different mediators in cell cycle and also through inflammation which either may serve to cancer cells or immune system. Attenuated expression levels of the HRH4 protein were mainly observed in advanced Gastric cancer samples compared to adjacent normal tissues (ANTs). Gastric cancers suggested that deletion and down-regulation of HRH4 might mainly take place in the progression.

mRNA level in the group with deleted copies of HRH4 was significantly lower than those with unaltered copies. Thus, the copy number loss of HRH4 at least plays a down-regulation of HRH4 expression in gastric cancer. On the other hand, the reduced HRH4 expression was also observed in the Gastric cancer samples with unaltered copies, which indicated that there are other mechanisms involved as well.

So it can be concluded that expression of HRH4 influence the proliferation ability of the Gastric cancer cells upon exposure to histamine. In future more studies needed to be done in finding out how mast cells are related to gastric cancer

References

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., . . . Moreno, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science, 252*(5013), 1651-1656.

Aitman, T. J., Dong, R., Vyse, T. J., Norsworthy, P. J., Johnson, M. D., Smith, J., . . . Petretto, E. (2006). Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature, 439*(7078), 851-855.

Arkenau, H.-T. (2009). Gastric cancer in the era of molecularly targeted agents: current drug development strategies. *Journal of cancer research and clinical oncology, 135*(7), 855-866.

Auer, P. L., & Doerge, R. (2010). Statistical design and analysis of RNA sequencing data. *Genetics, 185*(2), 405-416.

Bang, Y.-J., Van Cutsem, E., Feyereislova, A., Chung, H. C., Shen, L., Sawaki, A., . . . Satoh, T. (2010). Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *The Lancet, 376*(9742), 687-697.

Bouvard, V., Baan, R., Straif, K., Grosse, Y., Secretan, B., El Ghissassi, F., . . . Galichet, L. (2009). A review of human carcinogens—part B: biological agents. *The lancet oncology, 10*(4), 321-322.

Bridges, C. B. (1936). The bar" gene" a duplication. *Science, 83*(2148), 210-211.

Buckland, P. R. (2003). Polymorphically duplicated genes: their relevance to phenotypic variation in humans. *Annals of medicine, 35*(5), 308-315.

Chartier-Harlin, M.-C., Kachergus, J., Roumier, C., Mouroux, V., Douay, X., Lincoln, S., . . . Hulihan, M. (2004). α-Synuclein locus duplication as a cause of familial Parkinson's disease. *The Lancet, 364*(9440), 1167-1169.

Compare, D., Rocco, A., & Nardone, G. (2010). Risk factors in gastric cancer. *Eur Rev Med Pharmacol Sci, 14*(4), 302-308.

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International journal of cancer, 127*(12), 2893-2917.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics, 7*(2), 85-97.

Feuk, L., Marshall, C. R., Wintle, R. F., & Scherer, S. W. (2006). Structural variants: changing the landscape of chromosomes and design of disease studies. *Human molecular genetics, 15*(suppl 1), R57-R66.

Fredman, D., White, S. J., Potter, S., Eichler, E. E., Den Dunnen, J. T., & Brookes, A. J. (2004). Complex SNP-related sequence variation in segmental genome duplications. *Nature genetics, 36*(8), 861-866.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., . . . Bamshad, M. J. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science, 307*(5714), 1434-1440.

Gottesman, S. (2005). Micros for microbes: non-coding regulatory RNAs in bacteria. *TRENDS in Genetics, 21*(7), 399-404.

Grabsch, H., Sivakumar, S., Gray, S., Gabbert, H. E., & Müller, W. (2010). HER2 expression in gastric cancer: rare, heterogeneous and of no prognostic value–conclusions from 924 cases of two independent series. *Analytical Cellular Pathology, 32*(1), 57-65.

Guan, Y.-F., Li, G.-R., Wang, R.-J., Yi, Y.-T., Yang, L., Jiang, D., . . . Peng, Y. (2012). Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer. *Chinese journal of cancer, 31*(10), 463.

Hahn, W. C., & Weinberg, R. A. (2002). Rules for making human tumor cells. *New England Journal of Medicine, 347*(20), 1593-1603.

Hemedah, M., Loiacono, R., Coupar, I., & Mitchelson, F. (2001). Lack of evidence for histamine H3 receptor function in rat ileum and human colon. *Naunyn-Schmiedeberg's archives of pharmacology, 363*(2), 133-138.

Homer, N., Merriman, B., & Nelson, S. F. (2009). BFAST: an alignment tool for large scale genome resequencing. *PloS one, 4*(11), e7767.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., . . . Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature genetics, 36*(9), 949-951.

Inoue, K., & Lupski, J. R. (2002). Molecular mechanisms for genomic disorders. *Annual review of genomics and human genetics, 3*(1), 199-242.

Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Smigal, C., & Thun, M. J. (2006). Cancer statistics, 2006. *CA: a cancer journal for clinicians, 56*(2), 106-130.

Jongmans, M., Admiraal, R., Van Der Donk, K., Vissers, L., Baas, A., Kapusta, L., . . . Veltman, J. (2006). CHARGE syndrome: the phenotypic spectrum of mutations in the CHD7 gene. *Journal of medical genetics, 43*(4), 306-314.

Kamangar, F., Dores, G. M., & Anderson, W. F. (2006). Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. *J Clin Oncol, 24*(14), 2137-2150. doi: 10.1200/JCO.2005.05.2308

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome research, 12*(4), 656-664.

Kim, D., & Salzberg, S. L. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol, 12*(8), R72.

Ku, G. Y., & Ilson, D. H. (2010). Esophagogastric cancer: targeted agents. *Cancer treatment reviews, 36*(3), 235-248.

Lunter, G., & Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research, 21*(6), 936-939.

Lupski, J. R., & Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS genetics, 1*(6), e49.

Machado, A. M. D., Figueiredo, C., Seruca, R., & Rasmussen, L. J. (2010). < i> Helicobacter pylori</i> infection generates genetic instability in gastric cells. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 1806*(1), 58-65.

Mardis, E. R. (2011). A decade/'s perspective on DNA sequencing technology. *Nature, 470*(7333), 198-203.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., . . . Aldape, K. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature, 455*(7216), 1061-1068.

Meyerson, M., Gabriel, S., & Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics, 11*(10), 685-696.

Network, C. G. A. R. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature, 474*(7353), 609-615.

Nguyen, D.-Q., Webber, C., & Ponting, C. P. (2006). Bias of selection on human copy-number variants. *PLoS genetics, 2*(2), e20.

Pinto, C., Di Fabio, F., Siena, S., Cascinu, S., Llimpe, F. R., Ceccarelli, C., . . . Funaioli, C. (2007). Phase II study of cetuximab in combination with FOLFIRI in patients with untreated advanced gastric or gastroesophageal junction adenocarcinoma (FOLCETUX study). *Annals of Oncology, 18*(3), 510-517.

Repping, S., van Daalen, S. K., Brown, L. G., Korver, C. M., Lange, J., Marszalek, J. D., . . . Page, D. C. (2006). High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nature genetics, 38*(4), 463-467.

Resende, C., Thiel, A., Machado, J. C., & Ristimäki, A. (2011). Gastric cancer: basic aspects. *Helicobacter, 16*(s1), 38-44.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology, 12*(3), R22.

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerrière, A., Vital, A., . . . Vercelletto, M. (2005). APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature genetics, 38*(1), 24-26.

Schloss, J. A. (2008). How to get genomes at one ten-thousandth the cost. *Nature biotechnology, 26*(10), 1113.

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., . . . Segraves, R. (2005). Segmental duplications and copy-number variation in the human genome. *The American Journal of Human Genetics, 77*(1), 78-88.

Shaw-Smith, C., Redon, R., Rickman, L., Rio, M., Willatt, L., Fiegler, H., . . . Colleaux, L. (2004). Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *Journal of medical genetics, 41*(4), 241-248.

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

Sorek, R., & Cossart, P. (2009). Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics, 11*(1), 9-16.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols, 7*(3), 562-578.

Wainberg, Z. A., Anghel, A., Desai, A. J., Ayala, R., Luo, T., Safran, B., . . . Finn, R. S. (2010). Lapatinib, a dual EGFR and HER2 kinase inhibitor, selectively inhibits HER2-amplified human gastric cancer cells and is synergistic with trastuzumab in vitro and in vivo. *Clinical Cancer Research, 16*(5), 1509-1519.

Wang, K., Kan, J., Yuen, S. T., Shi, S. T., Chu, K. M., Law, S., . . . Tsui, W. Y. (2011). Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nature genetics, 43*(12), 1219-1223.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*(1), 57-63.

Yamashita, K., Sakuramoto, S., Nemoto, M., Shibata, T., Mieno, H., Katada, N., . . . Watanabe, M. (2011). Trend in gastric cancer: 35 years of surgical experience in Japan. *World journal of gastroenterology: WJG, 17*(29), 3390.

Yasui, W., Yokozaki, H., Fujimoto, J., Naka, K., Kuniyasu, H., & Tahara, E. (1999). Genetic and epigenetic alterations in multistep carcinogenesis of the stomach. *Journal of gastroenterology, 35*, 111-115.

Zheng, L., Wang, L., Ajani, J., & Xie, K. (2004). Molecular basis of gastric cancer development and progression. *Gastric cancer, 7*(2), 61-77.

Appendix

Downloading and installing software—Create a directory to store all of the executable programs used in this protocol (if none already exists):

```
$ mkdir $HOME/bin
```

Add the above directory to your PATH environment variable:

```
$ export PATH = $HOME/bin:$PATH
```

To install the SAM tools, download the SAM tools (http://samtools.sourceforge.net/) and unpack the SAM tools tarball and cd to the SAM tools source directory:

```
$ tar jxvf samtools-0.1.17.tar.bz2 $
cd samtools-0.1.17
```

Copy the samtools binary to some directory in your PATH:

```
$ cp samtools $HOME/bin
```

To install Bowtie, download the latest binary package for Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) and unpack the Bowtie zip archive and cd to the unpacked directory:

```
$ unzip bowtie-0.12.7-macos-10.5-x86_64.zip
$ cd bowtie-0.12.7
```

Copy the Bowtie executables to a directory in your PATH:

```
$ cp bowtie $HOME/bin
$ cp bowtie-build $HOME/bin
$ cp bowtie-inspect $HOME/bin
```

To install TopHat, download the binary package for version 1.3.2 of TopHat (http://tophat.cbcb.umd.edu/) and unpack the TopHat tarball and cd to the unpacked directory:

```
$ tar zxvf tophat-1.3.2.OSX_x86_64.tar.gz
$ cd tophat-1.3.2.OSX_x86_64
```

Copy the TopHat package executable files to some directory in your PATH:

```
cp * $HOME/bin
```

To install Cufflinks, download the binary package of version 1.2.1 for Cufflinks (http://cufflinks.cbcb.umd.edu/) and unpack the Cufflinks tarball and cd to the unpacked directory:

```
$ tar zxvf cufflinks-1.2.1.OSX_x86_64.tar.gz
$ cd cufflinks-1.2.1.OSX_x86_64
```

Copy the Cufflinks package executuble files to some directory in your PATH:

```
$ cp * $HOME/bin
```

To Install CummeRbund, start an R session:

```
$ R
```

```
R version 2.13.0 (2011-04-13)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin10.6.0/x86_64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type ‘license()’ or ‘licence()’ for distribution details.
R is a collaborative project with many contributors.
Type ‘contributors()’ for more information and ‘citation()’ on how to cite R
or R packages in publications.
Type ‘demo()’ for some demos, ‘help()’ for online help, or
‘help.start()’ for an HTML browser interface to help.
Type ‘q()’ to quit R.
```

Install the CummeRbund package:

```
> source(‘http://www.bioconductor.org/biocLite.R’)
> biocLite(‘cummeRbund’)
```

1| Map the reads for each sample to the reference genome:

```
$ tophat -p 8 -G genes.gtf -o C1_R1_thout genome C1_R1_1.fq C1_R1_2.fq
$ tophat -p 8 -G genes.gtf -o C1_R2_thout genome C1_R2_1.fq C1_R2_2.fq
$ tophat -p 8 -G genes.gtf -o C1_R3_thout genome C1_R3_1.fq C1_R3_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R1_thout genome C2_R1_1.fq C1_R1_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R2_thout genome C2_R2_1.fq C1_R2_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R3_thout genome C2_R3_1.fq C1_R3_2.fq
```

2| Assemble transcripts for each sample:

```
$ cufflinks -p 8 -o C1_R1_clout C1_R1_thout/accepted_hits.bam
$ cufflinks -p 8 -o C1_R2_clout C1_R2_thout/accepted_hits.bam
$ cufflinks -p 8 -o C1_R3_clout C1_R3_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R1_clout C2_R1_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R2_clout C2_R2_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R3_clout C2_R3_thout/accepted_hits.bam
```

3| Create a file called assemblies.txt that lists the assembly file for each sample. The file should contain the following lines:

```
./C1_R1_clout/transcripts.gtf
./C2_R2_clout/transcripts.gtf
./C1_R2_clout/transcripts.gtf
./C2_R1_clout/transcripts.gtf
```

```
./C1_R3_clout/transcripts.gtf
./C2_R3_clout/transcripts.gtf
```

4| Run Cuffmerge on all your assemblies to create a single merged transcriptome annotation:

```
cuffmerge -g genes.gtf -s genome.fa -p 8 assemblies.txt
```

5| Run Cuffdiff by using the merged transcriptome assembly along with the BAM files from TopHat for each replicate:

```
$ cuffdiff -o diff_out -b genome.fa -p 8 –L C1,C2 -u merged_asm/merged.gtf \
./C1_R1_thout/accepted_hits.bam,./C1_R2_thout/accepted_hits.bam,./
C1_R3_thout/ accepted_hits.bam \
./C2_R1_thout/accepted_hits.bam,./C2_R3_thout/accepted_hits.bam,./
C2_R2_thout/ accepted_hits.bam
```

6| Open a new plotting script file in the editor of your choice, or use the R interactive shell:

```
$ R
R version 2.13.0 (2011-04-13)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin10.6.0/x86_64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type ‘license()’ or ‘licence()’ for distribution details.
R is a collaborative project with many contributors.
Type ‘contributors()’ for more information and
‘citation()’ on how to cite R or R packages in publications.
Type ‘demo()’ for some demos, ‘help()’ for on-line help, or
‘help.start()’ for an HTML browser interface to help.
Type ‘q()’ to quit R.
```

7| Load the CummeRbund package into the R environment:

```
> library(cummeRbund)
```

8| Create a CummeRbund database from the Cuffdiff output:

```
> cuff_data < - readCufflinks(‘diff_out’)
```

9| Plot the distribution of expression levels for each sample

```
> csDensity(genes(cuff_data))
```

10| Compare the expression of each gene in two conditions with a scatter plot

```
> csScatter(genes(cuff_data), ‘C1’, ‘C2’)
```

11| Create a volcano plot to inspect differentially expressed genes

```
> csVolcano(genes(cuff_data), ‘C1’, ‘C2’)
```

12| Plot expression levels for genes of interest with bar plots

```
> mygene < - getGene(cuff_data, ‘regucalcin’)
> expressionBarplot (mygene)
```

13| Plot individual isoform expression levels of selected genes of interest with bar plots

```
> expressionBarplot(isoforms (mygene))
```

14| Inspect the map files to count the number of reads that map to each chromosome (optional). From your working directory, enter the following at the command line:

```
$ for i in *thout/accepted_hits.bam; do
echo $i; samtools index $i ; done;
$ for i in *thout/accepted_hits.bam; do
echo $i; samtools idxstats $i ; done;
```

The first command creates a searchable index for each map file so that you can quickly extract the alignments for a particular region of the genome or collect statistics on the entire alignment file. The second command reports the number of fragments that map to each chromosome.

Compare transcriptome assembly to the reference transcriptome (optional)

15| You can use a utility program included in the Cufflinks suite called Cuffcompare to compare assemblies against a reference transcriptome. Cuffcompare makes it possible to separate new genes from known ones, and new isoforms of known genes from known splice variants. Run Cuffcompare on each of the replicate assemblies as well as the merged transcriptome file:

```
$ find . -name transcripts.gtf > gtf_out_list.txt
$ cuffcompare -i gtf_out_list.txt -r genes.gtf
$ for i in ‘find . -name *.tmap’; do echo
$i; awk ‘NR > 1 { s[$3] + + } END { \ for (j in s) { print j, s[j] }} ’ $i;
done;
```

The first command creates a file called gtf_out_list.txt that lists all of the GTF files in the working directory (or its subdirectories). The second command runs Cuffcompare, which, compares each assembly GTF in the list to the reference annotation file genes.gtf.

Cuffcompare produces a number of output files and statistics, and a full description of its behaviour and functionality is out of the scope of this protocol. Please see the Cufflinks manual (http://cufflinks.cbcb.umd.edu/manual.html) for more details on Cuffcompare's output files and their formats. The third command prints a simple table for each assembly that lists how many transcripts in each assembly are complete matches to known transcripts, how many are partial matches and so on.

Record differentially expressed genes and transcripts to files for use in downstream analysis (optional)

16| You can use CummeRbund to quickly inspect the number of genes and transcripts that are differentially expressed between two samples. The R code below loads the results of Cuffdiff's analysis and reports the number of differentially expressed genes:

```
> library(cummeRbund)
> cuff_data < - readCufflinks(‘diff_out’)
>
> cuff_data
CuffSet instance with:
2 samples
14353 genes
26464 isoforms
17442 TSS
13727 CDS
14353 promoters
17442 splicing
11372 relCDS
> gene_diff_data < - diffData(genes(cuff_data))
> sig_gene_data < - subset(gene_diff_data, (significant = = ‘yes’))
> nrow(sig_gene_data)
```

17| Similar snippets can be used to extract differentially expressed transcripts or differentially spliced and regulated genes:

```
> isoform_diff_data < - diffData(isoforms(cuff_data), ‘C1’, ‘C2’)
> sig_isoform_data < - subset(isoform_diff_data, (significant = = ‘yes’))
> nrow(sig_isoform_data)
> tss_diff_data < - diffData(TSS(cuff_data), ‘C1’, ‘C2’)
> sig_tss_data < - subset(tss_diff_data, (significant = = ‘yes’))
> nrow(sig_tss_data)
> cds_diff_data < - diffData(CDS(cuff_data), ‘C1’, ‘C2’)
> sig_cds_data < - subset(cds_diff_data, (significant = = ‘yes’))
> nrow(sig_cds_data)
> promoter_diff_data < - distValues(promoters(cuff_data))
> sig_promoter_data < - subset(promoter_diff_data, (significant = = ‘yes’))
> nrow(sig_promoter_data)
> splicing_diff_data < - distValues(splicing(cuff_data))
```