

Studies on Information Content in Different Type of DNA Sequences

DISSERTATION

Submitted in the partial fulfilment for the award of the degree of

MASTER OF SCIENCE

IN

BIOTECHNOLOGY

Submitted by

Gurpreet Kaur

301401004

Under the Guidance of

Dr. Vikas Handa

Assistant Professor

Department of Biotechnology

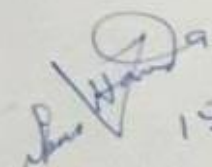


THAPAR UNIVERSITY, PATIALA

JULY, 2016

CERTIFICATE

This is to certify that thesis entitled "Studies on Information Content in Different Type of DNA Sequences" submitted by GurpreetKaur (301401004) in the partial fulfilment of the requirement for the award of Degree of Masters in Science in Department of Biotechnology, Thapar University, Patiala, is a record of student's own work carried out by her. The report has not been submitted for the award of any other degree or certificate in this or any other university or institute.



15.07.2016

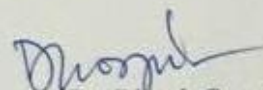
Dr. Vikas Handa

Assistant Professor

Department of Biotechnology

Thapar University

Patiala



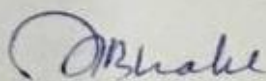
Dr. Dinesh Goyal

Professor and Head

Department of Biotechnology

Thapar University

Patiala



Dr. S.S. Bhatia

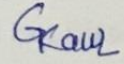
Dean, Academic Affairs

Thapar University

Patiala

DECLARATION

I hereby declare that the work which is being presented in the thesis "*Studies on Information Content in Different Type of DNA Sequences*" submitted by me for the award of the degree of Masters in Science in Department of Biotechnology, Thapar University, Patiala, is true and original record of my own independent and original research work carried out under the supervision of **Dr. Vikas Handa**. The matter embodied in this thesis has not been submitted in part or full to any other university or institute for the award of any degree in India or Abroad.



Gurpreet Kaur

301401004

Date: 15/07/16

ACKNOWLEDGEMENT

Though only my name appears on the cover of this dissertation, a great many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever. I would also like to acknowledge the Head of Department **Dr. Dinesh Goyal** and my course coordinator **Dr. Vikas Handa**. He has always been there to help me at every single step during my project work, motivating me, guiding me and encouraging me to complete my work.

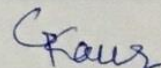
My acknowledgement will never be complete without the special mention of all my lab mates **Ms. Shivangi, Ms. Sonika and Ms. Nehakshi** who have taught me the lab culture and willingly devoted their valuable time whenever I needed their help.

Some friends are never too busy to give us a hand whenever they are needed. I want to make a special thanks to my friends who showed faith in me. I could not have asked for more than what I got from you all. I express my sincere love and affection towards all those benevolent souls and true friends, **Gurpreet, Swati and Rahul**.

I find it difficult to pen down my deepest sense of thanks towards my parents and all other family members who soulfully provided me their constant support and right impetus to undertake the challenge of this proportion like all other spheres of life, for what I cannot measure but treasure. It was my Father and Mother who gave me unconditional support during this time. They have elucidated the meaning of love, life and living. My sincere thanks to my parents who had always keep my morals high through tough time.

It has been great privilege to spend two years in the **Department of Biotechnology, Thapar University, Patiala** and its members will always remain dear to me.

Last but not the least, I wish to acknowledge all those, whose names have not figured here, but who helped me in any form during the period of my project work.


Gurpreet Kaur

Date: 15/07/16

Place: Patiala



DEDICATED TO

GURPREET

TABLE OF CONTENTS

ABBREVIATIONS (i)

LIST OF FIGURES (ii)

LIST OF TABLES(iii)

ABSTRACT(iv)

1. INTRODUCTION1

1.1 Compression

1.2 Data Compression

1.2.1 Lossy

1.2.2 Lossless Compression

1.2.3 Run Length Encoding

2. REVIEW OF LITERATURE6

2.1 Related Work

2.1.1 LZ Algorithm

2.1.2 Biocompress&Biocompress 2

2.1.3 Cfact

2.1.4 Gencompress

2.1.5 DNA Compress

2.1.6 LZ77

2.1.7 HuffBit Compress

2.1.8 GENBIT Compress

2.1.9 BIND

2.1.10 DELIMINATE

2.1.11 DNA Compact

2.1.12 BEETL-fastq

3. SCOPE OF STUDY 9

4. OBJECTIVES10

5. DATA SOURCE11

6. MATERIALS AND METHODS38

6.1 Sequence Analysis Tools

6.1.1 Microsoft Excel:

6.1.2 Notepad ++

6.1.3 SPSS

6.2 Methods

7. RESULT47

7.1 Compression of intergenic sequences, exons and non-coding sequences of genes of human Chromosome 21

7.2 Validation of lossless nature of the compression algorithm

7.3 Analyzing means of PCR of introns and exons:

7.4 Comparing the PCR of coding sequences and non-coding sequence using ANOVA

8. DISCUSSION62

9. CONCLUSION65

REFERENCES66

LIST OF ABBREVIATIONS

A	Adenine
C	Cytosine
DNA	Deoxyribonucleic acid
G	Guanine
NCBI	National Center for Biotechnology Information
PCR	Percentage Compression Ratio
RNA	Ribonucleic Acid
T	Thymine
UTR	Untranslated Region

LIST OF FIGURES

Fig. 1.1 Genome Complexity analysed by Reassociation Kinetics	2
Fig. 1.2 Example of lossless data compression	4
Fig. 1.3 Compression of DNA sequence by RLE Algorithm	4
Fig 5.1 Screenshot showing the genome of <i>Homo sapiens</i>	11
Fig 5.2 Screenshot showing few genes of Chromosome 21 of <i>Homo sapiens</i> in Map Viewer	11
Fig 5.3 Screenshot showing few genes of Chromosome 21 of <i>Homo sapiens</i> in Map Viewer	12
Fig 5.4 Screenshot showing a genes of Chromosome 21 of <i>Homo sapiens</i> in Evidence Viewer with position of Exons.	13
Fig 6.1 Screenshot showing the <i>Homo sapiens</i> chromosome 21 sequence.	38
Fig 7.1 Screenshot showing the Pairwise sequence Alignment results.	58

LIST OF TABLES

Table 1. Position of Exons of Split Genes of Chromosome 21 of *Homo sapiens*

Table2. Nucleotide Repeats to be considered for compression

Table3. Replacement of nucleotide repeats using RLE

Table4. Compression Scheme

Table5. PCR value of all the sequences

ABSTRACT

The organization of eukaryotic DNA is highly complex and it includes sequences such as of exons, introns, 3'UTR, 5'UTR, promoter and intergenic. From the above sequences the exons are coding sequences and rest all the non-coding sequences. Coding sequences are unique sequences and do not contain any repetitive sequences, while non-coding sequences contain mostly repetitive DNA sequences. Therefore the compressibility of exons can be less as compare to other non- coding sequences. We calculated percentage compression ratio for analysing the compressibility of different sequences. We analysed PCR means with two sample t-test, ANOVA and Tukey's test. So we have found a novel tool with the help of which we can measure the information content of DNA sequence. Low Compressibility of coding DNA when compared to non-coding DNA is just a beginning to show the application of information theory in genomics.

DNA is a biopolymer consisting of four different deoxyribonucleotides. The length of DNA is defined as number of nucleotides present in it. A nucleotide has three components – a nitrogenous base, a pentose sugar (deoxyribose) and a phosphate group. There are two types of nitrogenous bases- Purines (Adenine and Guanine) and Pyrimidines (Cytosine and Thymine). Four bases of DNA i.e. A, T, G and C can be considered as four alphabets of DNA molecule. A nitrogenous base is linked to the pentose sugar through a N-glycosidic linkage to form a nucleoside. When a phosphate group is linked to 5'-OH of a nucleoside through phosphoester linkage, a corresponding nucleotide is formed. Two nucleotides are linked through 3'-5' phosphodiester linkage to form dinucleotide. The backbone in a polynucleotide chain is formed due to sugar and phosphate group. DNA as genetic material in most of the living organisms and it carries genetic information in the form of sequence of four different bases from parent cell to daughter cells. Central dogma states that the genetic information flows from DNA → RNA → Protein. DNA sequence of genome can be transcribed into RNA and most of them translated into proteins. DNA as the genetic material has the following features: (i) It can store genetic information (ii) Able to generate its replica through the process of replication (iii) It should have its own mechanism to decode the genetic information into functional molecules such as RNAs and their translation into proteins.

DNA sequence which encodes a protein or RNA is called gene. In prokaryotes, most of the DNA sequence is coding. Prokaryotic DNA rarely contains any intron or non-coding sequence. Lower eukaryotic DNA contains very few genes and the number of introns per gene is less. They contain introns with small length. In higher eukaryotes e.g. mammals have ~2% coding sequence. Most of the genes contain introns and the number of intron per gene is more. Length of the introns in higher eukaryotes is ~10 fold large than lower eukaryotes. Genes that contain introns are called split genes. Repetitive sequence is mainly found in intergenic region i.e. the region between two adjacent genes which do not contain any gene. Introns are removed from the primary transcript through the process of Splicing. After the removal of introns, the exons are joined together to form messenger RNA.

Coding sequences primarily genes encode for proteins. The process begins with transcription. Promoters are the sequences which are recognized by RNA Polymerase and transcription machinery resulting in transcription of coding part of gene. Promoters contain conserved sequences upstream the coding region which are recognized by the transcription machinery. In eukaryotes, the primary transcript not only consists of introns (non-coding sequences) but also certain untranslated sequences in the 5' and 3' ends of the genes in addition to exons (coding sequences). These sequences found at extreme ends of the primary transcript are known as 5' UTR and 3' UTR. 5'UTR sequence is present upstream the gene or coding region. It contains a sequence where ribosome binds and begins

translation. 3'UTR present immediately after the stop codon; and helps in the termination of translation and in PTMs.

In simpler organisms almost the entire DNA consists of unique sequences. In higher organisms there can be large amounts of repetitive DNA. There are two types of repetitive DNA: tandem repeats and dispersed repeats. The length of the non-repetitive DNA component tends to increase as the complexity of organisms increase. Large amount of DNA present in the plants and animals indicates the presence of repetitive DNA. Most genes are present in non-repetitive DNA. This indicates that genetic complexity is proportional to the amount of non-repetitive DNA. **(Primrose, 2013)**

Genome complexity can also be analysed using **reassociation kinetics**. An organism's DNA can be heated in solution until it melts, and then cooled to allow DNA strands to reassociate forming double-stranded DNA. This is typically done after shearing the DNA to form many fragments a few hundred bases in length. The larger and more complex an organism's genome is, the longer it will take for complementary strands to bump into one another and hybridize. **(Primrose, 2013)**

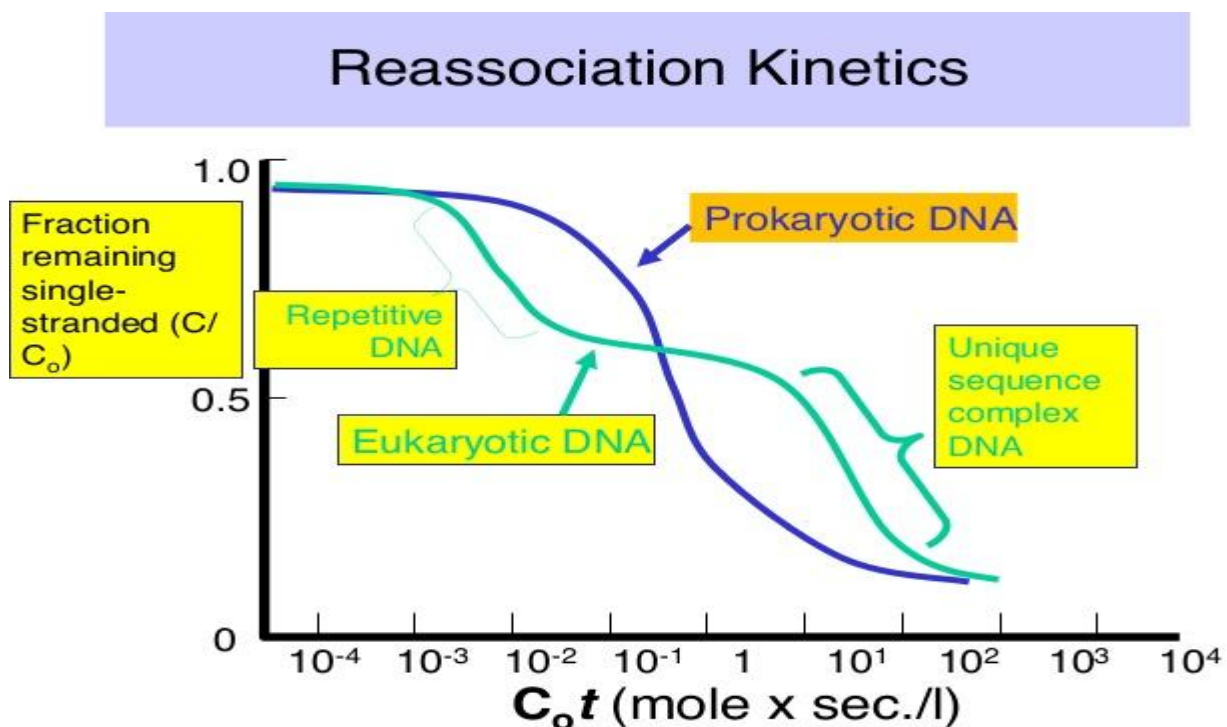


Fig. 1.1 Genome Complexity analysed by Reassociation Kinetics

(Source: <http://image.slidesharecdn.com/genomestructure-medics>)

Complexity of sequence can be measured in the form of compressibility of sequence known as Kolmogorov complexity. Kolmogorov complexity is based on algorithmic information theory considering objects as individual symbol strings, whereas the measures effective measure complexity, excess entropy, predictive information or thermodynamic depth relate objects to random variables and are ensemble based. **(Emmert-Streib F 2010)**

Nature of DNA sequence is not random because DNA sequence contains tandem repeats, ORF, Transposons, Pseudo-genes and duplicate genes. Coding Region contains ORF and Non- Coding Region contains tandem repeats, dispersed repeats, transposons and Pseudo-genes. Coding region of DNA sequence contains the information for synthesizing proteins whereas non coding region does not carry any information necessary to make proteins. There is not any fundamental difference in coding and non-coding DNA sequence except the ORF in coding region. Exons and coding regions are less complex as compare to non-coding regions i.e. introns and intergenic regions as they contain repetitive DNA sequence. Therefore the compressibility of coding is expected to be less than the non- coding region of DNA.

More than 98% portion of the DNA is non-coding, which contain introns and intergenic region. Introns and intergenic regions contain repetitive DNA sequences. Therefore these sequences can be compressed. Compression can be used to find out the relatedness of coding and non-coding sequences. The percent compression ratio of coding and non-coding regions of DNA must be different. Therefore, the compression technique can be used to test the non- randomness of different type of DNA sequence. **(Tungadri Bose et al., 2012)**

1.1 Compression

DNA is the genetic material in most of the organisms and consists genetic information in the form of sequence of bases. This sequence of bases is similar to data in the form of strings just like English alphabets and syntax. So DNA sequence can also be subjected to lossless compression just like other data used in computer files. Compression has been used as a tool for different genome comparison and for studying various properties of Genome. Compression is a process for reducing the size of data for efficient transfer and storage of data. Compressibility can also be used as a measure of information content, for the purpose of making inferences about sequences such as relatedness of the two sequences. For compressing the DNA sequence Lossless compression process can be used. **(Afify H. et. al, 2011)**

1.2 Data Compression

There are various types of compression algorithms which can be used to compress the different types of document. The compression algorithm used to compress the text document is called Data Compression. Data compression is of two types depending upon the type of data i.e. (i) Lossless compression and (ii) lossy compression. For DNA sequence compression Lossless compression technique can be used, because it is necessary that after decompression the decompressed data could be identical to the identical data.

1.2.1 Lossy Compression

It is a type of data compression where the data is compressed in such a way that after decompression the original data cannot be revived. Example of such type of data is compression of Image File (.jpg etc.). Image file once compressed cannot be regained its original size after decompression. The compression algorithm example is gzip compression algorithm.

1.2.2 Lossless Compression

Lossless compression is a type of data compression algorithms by which the original data can be perfectly reconstructed after decompression of the compressed data.

Lossless compression is used in cases where there is important that the original and the decompressed data be identical, or where deviations from the original data could be deleterious. Example of this type of document is text document.

Sequence → GGGGATATATGGCGGC
Compression → [G4[AT3[GGC2
Decompression → GGGGATATATGGCGGC

Fig. 1.2 Example of lossless data compression

Examples of lossless data compression are:

Run-Length Encoding, Huffman coding, Ziv and Lempel Algorithm and Burrow Wheeler transform. Most simple algorithm for DNA sequence compression is Run-Length Encoding Algorithm.

1.2.3 Run Length Encoding

Run-length encoding (RLE) is a form of lossless data compression in which runs of data are stored as a single data value and count. This is most useful on data that contains many such runs. For DNA sequence, RLE can be used to compress the homopolymeric nucleotide repeat, di-nucleotide repeat, tri-nucleotide repeat and tandem repeats. It is not useful with files that don't have many runs as it could greatly increase the file size.

Length

Sequence → GGGGATATATGGCGGC = 16
Compression → [G4[AT3[GGC2 = 12

Fig. 1.3 Compression of DNA sequence by RLE Algorithm

Compression percentage of sequences with different functional roles such as introns, exons, intergenic region, 3'UTR, 5'UTR and promoter region may vary according to sequence complexity or information content. For example the presence of repetitive DNA sequence has low complexity, lower information content and high compressibility. Different types of DNA sequences can be compressed and their level of compression may be expressed in terms of compression percentage.

Most commonly used DNA compressors are Biocompress & Biocompress 2, Cfact, DNA compress, LZ77, HuffBit Compress, BIND, DELIMINATE, and DNA- Compact etc. But the motive of all these compression techniques is just to reduce the size of DNA sequence for easy transfer and storage. None of these algorithms define the information content present in DNA on the basis of compression percentage.

On an average in *Homo sapiens*, the length of 3' UTR and 5' UTR is 800 and 200 nucleotides respectively (**Mignone et al., 2002**).

2.1 Related Work:

Most of the compression algorithms designed to the date is mainly focused to reduce the size of the DNA sequence. Neither of the algorithms are used to compare relatedness of different types of DNA sequences. Mostly designed algorithms for compression are:

2.1.1 LZ Algorithm:

For sequential data compression the first algorithm designed was LZ algorithm. It is based on dictionary coders and relies on exact repeats. It is the universal algorithm for text data compression. Therefore it can also be used to compress the DNA sequence. English alphabets from A-Z assigned Binary Codes such as A - 0001 (1), B – 00010 (2) and Z - 11010 (26). For DNA sequence binary codes are assigned as: T- 10100, G- 111, C- 11, A-1

The whole DNA sequence is converted to binary numbers and then compressed (**J. Ziv and A. Lempel, 1977**).

2.1.2 Biocompress & Biocompress 2:

Biocompress uses LZ compression scheme. It detects tandem repeats, multiple copies of genes and palindromic sequences in the target sequence and then encodes them by repeat length and the position of a previous repeat occurrence. After detection of repeat regions each region is assigned a position which is further used to compress the sequence (**Grumbach A. et al., 1993**).

2.1.3 Cfact:

Cfact is used to compress the biological sequence. It searches the longest exact matching repeat in the entire sequence. Detect repeats and select some of the repeats according to the maximum compressibility (**Rivals E. et al, 1995**).

2.1.4 Gencompress:

It is based on the fact that DNA sequence repetitive sequence and Palindromic sequences. This algorithm is based on approximate matching. It involves modified version of LZ algorithm i.e. LZW. Gencompress searches for reverse complements and approximate repeats. Approximate repeats are subsequences which can be transformed into original sequence using deletion, insertion and substitution. Gencompress has higher compression percentage than Biocompress and Cfact algorithms. It can be used to measure the relatedness of two DNA sequences or two genomes (**Chen X. et al., 2000**).

2.1.5 DNA Compress:

DNA Compress uses LZW compression scheme. The basic idea of LZW scheme is to replace repeat regions by references to a Dictionary. This algorithm is much faster than Gencompress. Compression is done in two phases:

- (i) Find all approximate repeats including complementary palindromes.
- (ii) Encode approximate repeat regions and non-repeat regions.

Each repeat region is checked to see whether it reduces the size, otherwise that repeat is discarded (**Chen X. et al., 2002**).

2.1.6 LZ77:

It is based on LZ algorithm. The compression is achieved by replacing repeated occurrence of data with reference to a dictionary. It is the modified form of LZ algorithm, which is used to compress the biological sequences such as DNA and RNA (**Sheng Bao et al., 2005**).

2.1.7 HuffBit Compress:

It is an algorithm to compress the DNA sequences using extended binary trees. Binary bit codes are assigned for each base i.e. A- 0, T- 111, G-110 and C- 10. It can compress both repetitive and non-repetitive DNA sequence (**Rajarajeshwari P. et al. 2010**).

2.1.8 GENBIT Compress:

It compresses repetitive and non-repetitive DNA sequence. It has the best compression ratio compared with other compression tools. Each fragment i.e. ACGT is replaced with 9 bits binary code (0 or 1). The bits assigned for any fragment followed by different fragment are “000000000”. When there is a repetitive segment followed by any segment then the value of the 9th bit raises to “1” otherwise “0”. (**Rajarajeshwari P. et al. 2010**).

2.1.9 BIND:

All sequence headers and non-ATGC characters are separated and stored in separate files. Headers are compressed using 7-Zip compression tool. Binary codes are assigned to the four nucleotide bases (A- 00, T- 01, G- 10, C- 11). BIND is based on a block-length encoding methodology. It is a lossless compression algorithm. Compression and decompression can be achieved in a very reasonable time (**Tungadri Bose et al., 2012**).

2.1.10 DELIMINATE:

There are two variants of DELIMINATE referred to as DELIM-1 and DELIM-2. These two perform the compression of sequence in two phases. Information of all non-ATGC characters and low complexity regions are recorded. All characters are converted to uppercase. The positions of the two bases having the highest frequencies of occurrence are delta encoded and these bases are subsequently eliminated from the sequence. The remaining two bases are then represented with a binary code. Sequence is compressed using 7-zip tool (**Mohammed M.H. et al., 2012**).

2.1.11 DNA Compact:

In first step, repetitions, reverse complement and palindromes removed. It is designed to compress the sequence with and without reference. Binary codes are assigned to the nucleotide bases: A- 01, T- 00, G- 10, C- 11. After that the sequence is compressed. Compression performance can be improved by the synthesis of complementary contextual models (**Li P. et al., 2013**).

2.1.12 BEETL-fastq:

It is a tool that compresses FASTQ- formatted DNA reads more compactly than gzip. It also permits k-mer queries within the archived sequences. It is a lossless compression algorithm (**Cox J. A. et al., 2014**).

Genome Complexity:

Genome complexity can also be analysed using **Reassociation kinetics**. An organism's DNA can be heated in solution until it melts, and then cooled to allow DNA strands to reassociate forming double-stranded DNA. This is typically done after shearing the DNA to form many fragments a few hundred bases in length. The larger and more complex an organism's genome is, the longer it will take for complimentary strands to bump into one another and hybridize. (**Primrose, 2013**)

Sequence complexity

Kolmogorov Complexity:

Complexity of sequence can be measured in the form of compressibility of sequence known as Kolmogorov complexity. Kolmogorov complexity is based on algorithmic information theory considering objects as individual symbol strings, whereas the measures effective measure complexity, excess entropy, predictive information or thermodynamic depth relate objects to random variables and are ensemble based. (**Emmert-Streib F 2010**).

Since exons are coding they have higher complexity. They consist largely of unique sequences and lack repeat sequences. So they cannot be compressed very effectively. However converse is this for tandem repeat sequences. Lower the Percent compression ratio (PCR) value lower the complexity. Exons have lower compressibility and tandem repeats have higher compressibility. Rest of the various types of sequences should have PCR value between the two extremes.

Expected PCR from lowest to highest are:

- ✓ Intergenic
- ✓ Introns
- ✓ 3' UTR
- ✓ 5' UTR
- ✓ Promoter
- ✓ Exons

- ✓ Development of lossless compression algorithm for efficiently compressing the DNA sequence.
- ✓ Comparison of compressibility of coding and various types of non-coding sequences.

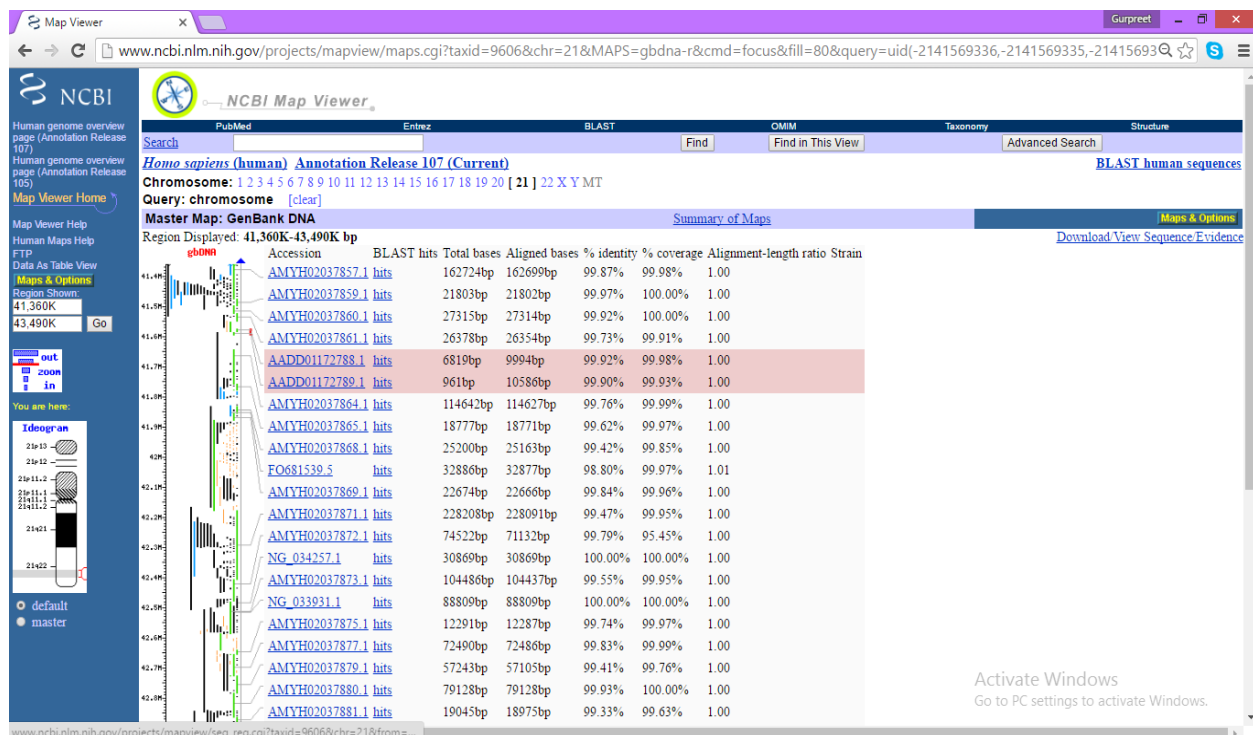
Genome of *Homo sapiens* was searched in Genome database of NCBI Website (<http://www.ncbi.nlm.nih.gov/>).

Fig 5.1 Screenshot showing the genome of *Homo sapiens*

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Nuc	Chr	1	NC_000001.11	CM000663.2	248.96	42.3	11,046	17	90	4,350	5,078	1,372
Nuc	Chr	2	NC_000002.12	CM000664.2	242.19	40.3	8,054	-	8	3,638	3,862	1,166
Nuc	Chr	3	NC_000003.12	CM000665.2	198.3	39.7	6,790	-	4	2,723	2,971	887
Nuc	Chr	4	NC_000004.12	CM000666.2	190.22	38.3	4,374	-	1	2,209	2,441	799
Nuc	Chr	5	NC_000005.10	CM000667.2	181.54	39.5	4,599	-	17	2,225	2,578	766
Nuc	Chr	6	NC_000006.12	CM000668.2	170.81	39.6	5,338	-	144	2,408	3,000	876
Nuc	Chr	7	NC_000007.14	CM000669.2	159.35	40.7	4,843	-	21	2,232	2,774	896
Nuc	Chr	8	NC_000008.11	CM000670.2	145.14	40.2	3,962	-	5	1,975	2,152	661
Nuc	Chr	9	NC_000009.12	CM000671.2	138.4	42.3	4,572	-	3	2,135	2,262	702
Nuc	Chr	10	NC_000010.11	CM000672.2	133.8	41.6	5,237	-	3	2,053	2,174	631
Nuc	Chr	11	NC_000011.10	CM000673.2	135.09	41.6	6,187	-	13	2,267	2,920	835
Nuc	Chr	12	NC_000012.12	CM000674.2	133.28	40.8	5,648	-	9	2,313	2,521	680
Nuc	Chr	13	NC_000013.11	CM000675.2	114.36	40.2	1,946	-	4	1,235	1,381	477
Nuc	Chr	14	NC_000014.9	CM000676.2	107.04	42.2	3,252	-	18	1,704	2,055	583
Nuc	Chr	15	NC_000015.10	CM000677.2	101.99	43.4	3,436	-	9	1,778	1,814	555
Nuc	Chr	16	NC_000016.10	CM000678.2	90.34	45.1	4,453	-	27	1,700	1,920	451
Nuc	Chr	17	NC_000017.11	CM000679.2	83.26	45.3	6,004	-	33	2,131	2,432	541
Nuc	Chr	18	NC_000018.10	CM000680.2	80.37	39.8	1,812	-	1	1,013	988	295
Nuc	Chr	19	NC_000019.10	CM000681.2	58.62	47.9	6,452	-	6	1,750	2,481	514
Nuc	Chr	20	NC_000020.11	CM000682.2	64.44	43.9	2,794	-	-	1,276	1,349	329
Nuc	Chr	21	NC_000021.9	CM000683.2	46.71	42.2	1,240	-	1	687	756	202
Nuc	Chr	22	NC_000022.11	CM000684.2	50.82	47.7	2,493	-	-	965	1,172	348
Nuc	Chr	X	NC_000023.11	CM000685.2	156.04	39.6	3,739	-	4	1,196	2,158	859
Nuc	Chr	Y	NC_000024.10	CM000686.2	57.23	45.4	325	-	-	298	577	395
MT	Chr	MT	NC_012920.1	J01415.2	0.016569	44.4	13	2	22	-	37	-
Un	-	-	-	-	144.26	44.5	5,172	-	186	2,840	5,824	1,757

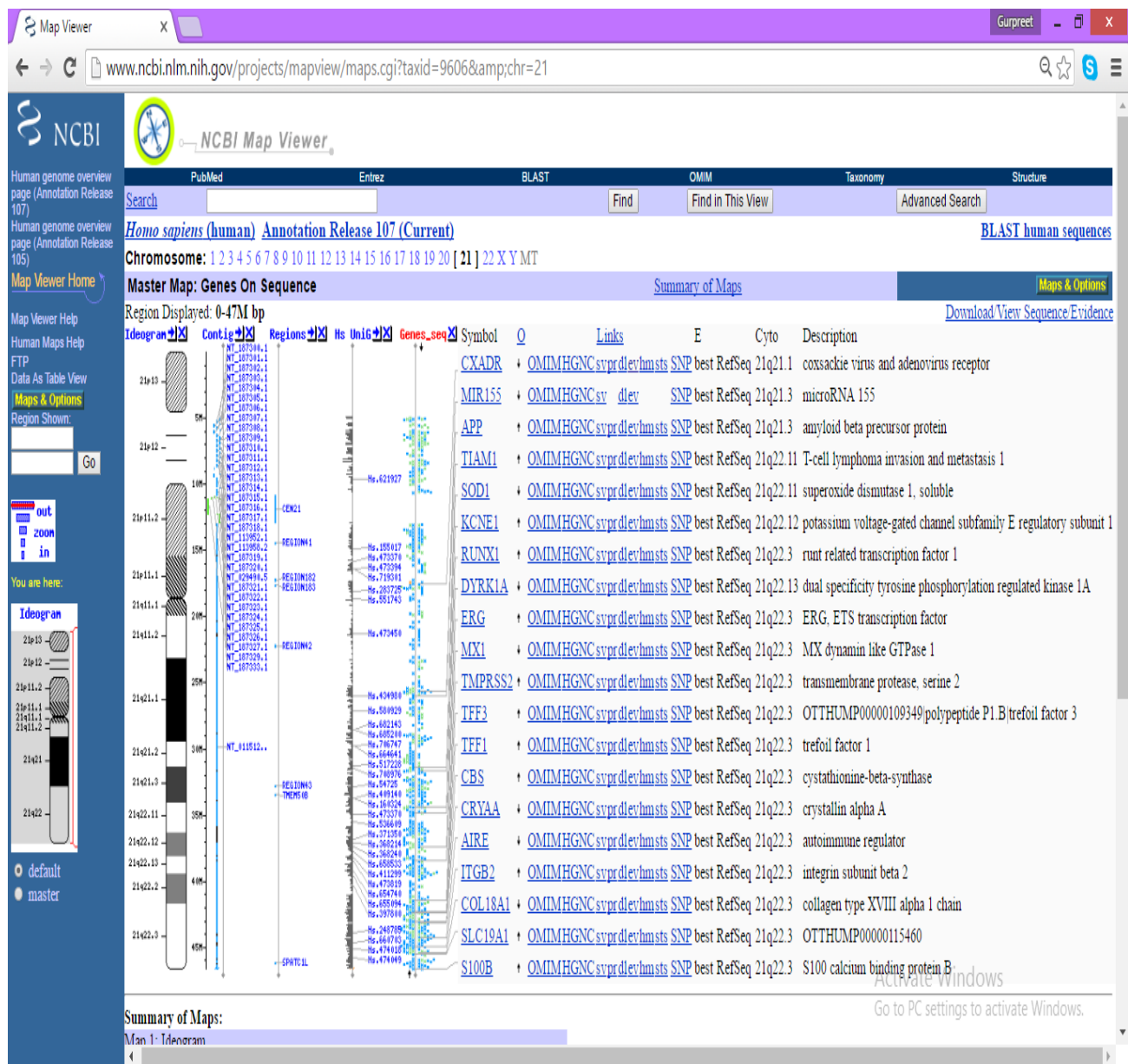
Then the chromosome 21 was selected and Map Viewer was opened.

Fig 5.2 Screenshot showing few genes of Chromosome 21 of *Homo sapiens* in Map Viewer



In Map Viewer, the Evidence viewer was opened for each gene. Genes of chromosome 21 of *Homo sapiens* that contain exons and introns were selected for compression. Chromosome 21 has 301 split genes.

Fig 5.3 Screenshot showing few genes of Chromosome 21 of *Homo sapiens* in Map Viewer



Positions and number of exons were noted down from the Map Viewer using Evidence Viewer.

Fig 5.4 Screenshot showing a genes of Chromosome 21 of *Homo sapiens* in Evidence Viewer with position of Exons.



Table 1. Position of Exons of Split Genes of Chromosome 21 of *Homo sapiens*

nuclear receptor interacting protein 1			chromosome 21 open reading frame 91			superoxide dismutase 1, soluble			riply transcriptional repressor 3		
EXON	NT_011512.12		EXON	NT_011512.12		EXON	NT_011512.12		EXON	NT_011512.12	
1	2099917	2100102	1	4853495	4853578	1	18693814	18694033	1	24040342	24040425
2	2098937	2098997	2	4852384	4852519	2	18697982	18698078	2	24040755	24041068
3	2077687	2077766	3	4830774	4831310	3	18700641	18700710	3	24042349	24042415
4	2048536	2048658	4	4829400	4829462	4	18701450	18701567	4	24047743	24047810
5	1995427	2002718	5	4823159	4827773	5	18702663	18703123	5	24051981	24053851
uncharacterized LOC100133286			small integral membrane protein 11A			long intergenic non-protein coding RNA 310			mitochondrial ribosomal protein S6		
EXON NT_011512.12			EXON NT_011512.12			EXON NT_011512.12			EXON NT_011512.12		
1	23160679	23160832	1	21409642	21409759	1	21214869	21214985	1	21107715	21107937
2	23132316	23132451	2	21409760	21410371	2	21216900	21217437	2	21126477	21140454
3	23109723	23109794	3	21413620	21413708	3	21220058	21220138	3	21141716	21143998

4	23109073	23109159	4	21419670	21420812	4	21222409	21222854	4	21156366	21159672
5	23103834	23104677	5	21422821	21423346	5	21222960	21224114	5	21176600	21177226
long intergenic non-protein coding RNA 945			C21orf62 antisense RNA 1			chromosome 21 open reading frame 59			t-complex 10-like		
EXON			EXON			EXON			EXON		
1	20092021	20092090	1	19806292	19806512	1	19646297	19646800	1	19619613	19619727
2	20092293	20092396	2	19818989	19819111	2	19644030	19644197	2	19618353	19618497
3	20096984	20097008	3	19822804	19822893	3	19641841	19641908	3	19616392	19616607
4	20098468	20098640	4	19825015	19825350	4	19638317	19638705	4	19612886	19613023
5	20098795	20099271	5	19831131	19831899	5	19635859	19637484	5	19609033	19611115
D21S2088E			uncharacterized LOC105372804			Down syndrome critical region 4			ybeY metalloproteinase (putative)		
EXON			EXON			EXON			EXON		
1	10418880	10419053	1	26383095	26383323	1	25155320	25155552	1	33320522	33320929
2	10407140	10407276	2	26382737	26382769	2	25154499	25154600	2	33321062	33321315
3	10406519	10406640	3	26382614	26382643	3	25088203	25088972	3	33325526	33326713
4	10398230	10398385	4	26379338	26382531	4	25044670	25044810	4	33330354	33330422
5	10395296	10396431	5	26376501	26377351	5	25040736	25041018	5	33331731	33331943
ITGB2 antisense RNA 1			NADH:ubiquinone oxidoreductase subunit V3			zinc finger and BTB domain containing 21			uncharacterized LOC105372809		
EXON			EXON			EXON			EXON		
1	31955227	31955543	1	29927460	29927573	1	29044444	29044579	1	28615457	28615511
2	31957105	31957175	2	29931119	29931239	2	29043659	29043789	2	28610327	28610617
3	31960999	31961196	3	29937374	29938468	3	29037089	29037153	3	28597099	28597150
4	31961817	31961859	4	29941032	29941094	4	29036542	29036755	4	28594524	28594644
5	31962041	31963872	5	29943056	29943855	5	29021023	29028300	5	28593317	28593565
B melanoma antigen family member 2			interferon (alpha, beta and omega) receptor 2			potassium channel, inwardly rectifying subfamily J, member 15			GA binding protein transcription factor alpha subunit		
EXON	NT_029490.5		EXON			EXON			EXON		
1	89205	89427	1	20264087	20264408	1	25264107	25264215	1	12768762	12769770
2	90484	90588	2	20264662	20264749	2	25290890	25290979	2	12775765	12775867
3	129806	131484	3	20276032	20276169	3	25290996	25291377	3	12779402	12779546
4	138509	138680	4	20277865	20277906	4	25304943	25307212	4	12783228	12783312
5	140535	140649	5	20279143	20279266	5	25307668	25309726	5	12786181	12786426
6	148701	149402	6	20280910	20281082	6	25325798	25325939	6	12792202	12792396
7	158413	158532	7	20282901	20283046	7	25330922	25331016	7	12796504	12796557
8	161291	161407	8	20286854	20287022	8	25331118	25331215	8	12798402	12798542
9	166919	167020	9	20294789	20294919	9	25332336	25332474	9	12798787	12798979
10	167213	167288	#	20296753	20298717	10	25333436	25336018	10	12803196	12806652
transmembrane phosphatase with tensin homology			neural cell adhesion molecule 2			ATP binding cassette subfamily G member 1			trafficking protein particle complex 10		
EXON	NT_029490.5		EXON			EXON			EXON		
1	197187	197367	1	8032507	8032810	1	29233881	29234041	1	31046517	31046752
2	200253	200361	2	8039150	8039581	2	29235769	29235915	2	31066283	31066364

3	203028	203085	3	8114660	8116473	3	29241138	29241178	3	31071984	31072119
4	214354	214407	4	8181312	8181443	4	29250269	29250380	4	31086472	31086668
5	216785	216838	5	8214124	8216220	5	29253339	29253496	5	31089890	31090085
6	217212	217239	6	8244733	8244883	6	29254013	29254238	6	31093295	31093406
7	218068	218121	7	8314770	8314844	7	29259863	29260106	7	31093657	31094426
8	219002	219055	8	8318386	8318592	8	29290276	29290808	8	31097245	31097360
9	228330	228389	9	8320461	8320604	9	29293507	29293694	9	31097730	31097977
10	235167	235217	10	8326296	8326433	10	29294181	29294331	10	31108516	31108662
11	236703	236864	11	8358575	8358692	11	29302744	29304473	11	31109231	31109345
12	243343	243462	12	8369697	8369857	12	29305262	29305546	12	31110744	31110820
13	245110	245209	13	8372581	8372726	13	29307495	29307627	13	31111885	31111976
14	245356	245419	14	8408055	8408205	14	29310536	29311137	14	31113756	31113896
15	246158	246222	15	8419796	8419866	15	29316466	29316611	15	31114207	31114319
16	253133	253193	16	8444466	8444653	16	29318752	29318875	16	31116980	31117494
17	254010	254091	17	8452665	8452761	17	29320072	29320186	17	31118314	31118455
18	254190	254278	18	8466300	8466473	18	29322081	29322504	18	31120994	31121152
19	266135	266196	19	8500798	8500917	19	29324242	29324410	19	31121891	31122120
20	267966	268046	20	8502854	8502975	20	29325284	29325384	20	31122231	31124125
21	271655	271760	21	8511483	8511663	21	29325690	29325848	21	31126115	31126241
22	273688	273767	22	8521281	8521571	22	29328734	29329544	22	31128255	31128425
23	277731	277823	23	8543043	8543247	23	29329762	29331436	23	31130772	31133251
24	279235	279305	24	8568729	8568848	24	29333186	29335001	24	31136970	31137164
25	281090	281943	25	8572038	8574388	25	29338255	29338579	25	31137485	31140744
SAMS1 antisense RNA 1			uncharacterized LOC105379499			ICOS ligand			uncharacterized LOC101928269		
EXON			EXON	NT_187311.1		EXON	NT_187300.1		EXON		
1	1616394	1616483	1	13373	13546	1	12531	12693	1	23038756	23038859
2	1619554	1619730	2	10155	11105	2	15009	15049	2	23022322	23022363
3	1625514	1625660	3	8938	9377	3	16280	16630	3	23018764	23019022
4	1628886	1628932	4	7891	8103	4	17934	18225	4	23015943	23016108
5	1632117	1632495	5	2231	2429	5	22053	22217	5	23004267	23004359
						6	23408	26775	6	22988852	22989017
uncharacterized LOC100506403			long intergenic non-protein coding RNA 160			chloride intracellular channel 6			ES1 protein homolog, mitochondrial		
EXON			EXON			EXON			EXON	NT_187300.1	
1	22614852	22614956	1	21771302	21771374	1	21703355	21704954	1	118214	118438
2	22499376	22499521	2	21768878	21768948	2	21740219	21740272	2	117898	117950
3	22496153	22496217	3	21768153	21768233	3	21741472	21741581	3	115879	115992
4	22469853	22473263	4	21767101	21767271	4	21742136	21742261	4	114756	114875
5	22456250	22456277	5	21758238	21758388	5	21742892	21742998	5	111711	111803
6	22406699	22406985	6	21757999	21758145	6	21743549	21743730	6	108691	108847
						7	21750513	21752419	7	106329	107231

uncharacterized LOC102724219			putative uncharacterized protein C21orf125			GRIK1 antisense RNA 2			BACH1 intronic transcript 2		
EXON	NT_187303.1		EXON	NT_187306.1		EXON			EXON		
1	83216	86975	1	120752	121649	1	16630232	16630970	1	16406693	16406755
2	89125	92992	2	119168	120500	2	16631762	16631886	2	16407434	16408172
			3	117771	118552	3	16664731	16664939			
ES1 protein homolog, mitochondrial			CBR3 antisense RNA 1			ATP synthase, H ⁺ transporting, mitochondrial F1 complex, O subunit			interleukin 10 receptor subunit beta		
EXON	NT_187300.1		EXON			EXON			EXON		
1	118214	118438	1	23190409	23190500	1	20948642	20950046	1	20300552	20300706
2	117898	117950	2	23182162	23184902	2	20946481	20946591	2	20302120	20302709
3	115879	115992	3	23180448	23181817	3	20943274	20943403	3	20310788	20310945
4	114756	114875	4	23175588	23176138	4	20940641	20941645	4	20313944	20314110
5	111711	111803	5	23172017	23172124	5	20938874	20939773	5	20317286	20317433
6	108691	108847	6	23167879	23167953	6	20938127	20938766	6	20322296	20322453
7	106329	107231	7	23165959	23167266	7	20937645	20937831	7	20330376	20331426
uncharacterized LOC105372831			long intergenic non-protein coding RNA 189			long intergenic non-protein coding RNA 161			long intergenic non-protein coding RNA 314		
EXON	NT_187300.1		EXON			EXON			EXON		
1	63458	63685	1	16227672	16228452	1	15573510	15574076	1	15047555	15047624
2	77617	77867	2	16263153	16263225	2	15574245	15574547	2	15056889	15047555
			3	16322176	16322397						
CAF-1 subunit B			SET domain containing 4			long intergenic non-protein coding RNA 649			hCG_401147		
EXON			EXON			EXON			EXON		
1	23419583	23419643	1	23094474	23095010	1	20965314	20965934	1	18907507	18908127
2	23420252	23420454	2	23093008	23094453	2	20972417	20972564	2	18958660	18958952
3	23421790	23421922	3	23091576	23091669	3	20977748	20977788	3	18974357	18974412
4	23425743	23425860	4	23091301	23091396	4	20977817	20978068	4	18980228	18980363
5	23428739	23428842	5	23087775	23087812	5	20978780	20979127	5	18993035	18993162
6	23431607	23431703	6	23082500	23082588	6	20980002	20980087	6	19002442	19002577
7	23433713	23433797	7	23079774	23080203	7	20982918	20982972	7	19008747	19008909
8	23436950	23437043	8	23077876	23078148	8	20982980	20983484	8	19017718	19017801
9	23442953	23443022	9	23075999	23077565	9	20985302	20985546	9	19024321	19024368
10	23443566	23443657	10	23074767	23074847	10	20996257	20998152	10	19026897	19029350
11	23445655	23445796	11	23072342	23072465	11	21001216	21001710	11	19029960	19030140
12	23447076	23447507	12	23068732	23071013	12	21003429	21005378	12	19032718	19038256
13	23448739	23451019									
regulator of calcineurin 1			potassium channel, voltage gated subfamily E regulatory beta subunit 1			eva-1 homolog C (C. elegans)			Beta-1,3-galactosyltransferase 5		
EXON			EXON			EXON			EXON		
1	21648952	21649334	1	21546219	21546467	1	19446629	19447205	1	26647132	26647259
2	21648470	21648638	2	21543636	21545630	2	19475969	19476872	2	26649415	26649522

3	21647979	21648035	3	21494091	21494130	3	19486377	19487700	3	26664483	26664682
4	21611709	21612440	4	21493721	21493796	4	19487886	19492652	4	26676577	26676668
5	21560855	21561202	5	21492846	21492956	5	19501888	19502040	5	26680584	26680814
6	21559280	21559659	6	21489578	21490001	6	19529219	19529362	6	26691512	26691791
7	21558719	21558829	7	21480880	21483876	7	19535607	19535687	7	26691977	26692085
8	21557729	21558481	8	21458025	21458158	8	19538118	19538207	8	26693945	26694104
9	21553610	21557309	9	21452907	21453822	9	19549006	19549592	9	26694661	26697081
10	21550634	21552448									
PAXBP1 antisense RNA 1			melanocortin 2 receptor accessory protein			GRIK1 antisense RNA 1			N-6 adenine-specific DNA methyltransferase 1 (putative)		
EXON			EXON			EXON			EXON		
1	19762306	19764853	1	19326005	19326041	1	16782367	16782492	1	15919404	15919565
2	19766711	19767108	2	19327185	19327324	2	16783026	16783337	2	15917177	15917263
3	19767781	19767849	3	19333042	19333269	3	16793023	16793206	3	15916352	15916442
4	19771565	19771665	4	19340832	19340932	4	16794212	16794257	4	15914062	15914145
5	19772464	19772533	5	19345874	19346482	5	16794586	16794659	5	15912384	15912525
6	19776454	19777314	6	19348743	19348976	6	16795449	16795611	6	15910256	15910683
						7	16797770	16798198	7	15906383	15909266
RWD domain containing 2B			uncharacterized LOC101927973			uncharacterized LOC284825			ADAM metalloproteinase with thrombospondin type 1 motif 5		
EXON			EXON			EXON			EXON		
1	16053403	16053570	1	15262554	15262859	1	15082606	15082729	1	13999480	14001312
2	16046395	16046896	2	15260076	15260183	2	15124595	15124713	2	13988931	13989063
3	16045800	16046010	3	15251370	15251539	3	15135458	15135511	3	13977572	13977740
4	16045326	16045349	4	15246432	15246543	4	15136742	15137772	4	13968658	13968941
5	16042587	16042813	5	15204270	15204335	5	15164722	15164814	5	13967053	13967236
6	16042432	16042499	6	15201418	15201511	6	15169506	15169900	6	13966196	13966371
7	16041953	16042315	7	15197794	15197909	7	15170882	15171801	7	13964078	13964253
8	16039951	16040843	8	15150286	15154066				8	13952104	13958812
uncharacterized LOC105372763			uncharacterized LOC105372760			MIR155 host gene			DSCAM antisense RNA 1		
EXON			EXON			EXON			EXON		
1	14672885	14673115	1	13923568	13923636	1	12596101	12596206	1	27417275	27417776
2	14682857	14682957	2	13944058	13944222	2	12596317	12596449	2	27417844	27418330
3	14687548	14687683	3	13973808	13973934	3	12604029	12604113	3	27418771	27419550
						4	12608085	12609360			
uncharacterized LOC105372753			Purkinje cell protein 4			carbonyl reductase 1			OTTHUMP00000109349 polypeptide P1.B trefoil factor 3		
EXON			EXON			EXON			EXON		
1	12552468	12552530	1	26901613	26901702	1	23104116	23104596	1	29349485	29349843
2	12549665	12550647	2	26932666	26932719	2	23105142	23105775	2	29347677	29347823
			3	26963176	26963589	3	23106007	23107369	3	29345859	29346461
B3GALT5 antisense RNA 1			tryptophan rich basic protein			Down syndrome critical region 9 (non-protein coding)			BTG family member 3		
EXON			EXON			EXON			EXON		

1	26646489	26647016	1	26414479	26414678	1	24242695	24242736	1	4646891	4647142
2	26643775	26643857	2	26421957	26422047	2	24242846	24242948	2	4643164	4643344
3	26639791	26640539	3	26424890	26425668	3	24244197	24244306	3	4639052	4639189
4	26634013	26634041	4	26425961	26426712	4	24248288	24248345	4	4638337	4638468
5	26633533	26633848	5	26427358	26427472	5	24250597	24250685	5	4632809	4633016
6	26631339	26632181	6	26431058	26432087	6	24254430	24255932	6	4627842	4628524
ATP synthase, H+ transporting, mitochondrial Fo complex subunit F6			CPI-B OTTHUMP00000109464			S100 calcium binding protein B			BRWD1 antisense RNA 1		
EXON			EXON			EXON			EXON NT_011512.12		
1	12769045	12769865	1	30810396	30810567	1	33637920	33639314	1	26349899	26350007
2	12763823	12763993	2	30808850	30808951	2	33636470	33636608	2	26351237	26351401
3	12759418	12759542	3	30807857	30808522	3	33634474	33634567	3	26356994	26357410
4	12758672	12758869				4	33632810	33633695			
SH3 domain binding glutamate-rich protein			Down syndrome critical region 3			coxsackie virus and adenovirus receptor			TSPEAR antisense RNA 1		
EXON			EXON			EXON			EXON		
1	26480057	26480273	1	24301430	24302142	1	4547098	4547364	1	31540236	31540520
2	26481202	26481261	2	24300984	24301039	2	4581219	4581385	2	31540839	31541291
3	26486022	26486333	3	24274688	24274831	3	4585941	4586145	3	31542229	31542501
4	26496567	26496752	4	24272652	24272801	4	4593168	4593323	4	31542919	31543121
5	26509327	26509407	5	24267554	24267634	5	4594894	4595016	5	31543444	31543598
6	26534015	26534107	6	24266569	24266837	6	4595530	4595668	6	31544015	31544119
7	26543190	26543219	7	24265685	24266167	7	4599620	4604303	7	31544878	31544991
8	26545872	26546001	8	24262415	24263675	8	4625453	4629208	8	31545131	31545263
9	26549280	26549698	9	24257612	24261998	9	4645965	4648277	9	31549689	31550767
mir-99a-let-7c cluster host gene			MORC family CW-type zinc finger 3			crystallin zeta like 1			chaperonin containing TCP1, subunit 8 (theta)		
EXON			EXON			EXON			EXON		
1	3104714	3104759	1	23354381	23354495	1	20675873	20676046	1	16106462	16107989
2	3105306	3105590	2	23364100	23364185	2	20675371	20675497	2	16106144	16106180
3	3215783	3215879	3	23367838	23367910	3	20673937	20674380	3	16104439	16104529
4	3228571	3228825	4	23371066	23371198	4	20665678	20665749	4	16103615	16103694
5	3265248	3265307	5	23371924	23372138	5	20658875	20658952	5	16101748	16101897
6	3364796	3364903	6	23372966	23373691	6	20656188	20656260	6	16101083	16101263
7	3401555	3403425	7	23375591	23375738	7	20650898	20650942	7	16100908	16100969
8	3425806	3425901	8	23378771	23378899	8	20650479	20650795	8	16099160	16099297
9	3453520	3453933	9	23379104	23379223	9	20647730	20647798	9	16097544	16097722
10	3454647	3454764	10	23383503	23383600	10	20637596	20637729	10	16096682	16096748
11	3521682	3522087	11	23390812	23390916	11	20636426	20636537	11	16096520	16096607
12	3555911	3555996	12	23393264	23395732	12	20633342	20633477	12	16096320	16096435
13	3571473	3571631	13	23395866	23396232	13	20631472	20631593	13	16095688	16095759
14	3573281	3573361	14	23396375	23396420	14	20629923	20630028	14	16095445	16095609

15	3574020	3574103	15	23398285	23398451	15	20628831	20629868	15	16094733	16094852
16	3624429	3624517	16	23403180	23404068	16	20625354	20625674	16	16093614	16094287
17	3628532	3628646	17	23406566	23406723	17	20623533	20624113	17	16092265	16092384
18	3641186	3643967	18	23409335	23410838				18	16090514	16090744
junctional adhesion molecule 4			coxsackie virus and adenovirus *receptor			single-minded family bHLH transcription factor 2			POTE ankyrin domain family, member D		
EXON			EXON			EXON			EXON NT_011512.12		
1	26761947	26762184	1	4951532	4951592	1	23733325	23734113	1	644369	644941
2	26763325	26764872	2	5037776	5038479	2	23743360	23744805	2	649315	649429
3	26779599	26779718	3	5038548	5040538	3	23746725	23746814	3	649589	649762
4	26780408	26780490	4	5062064	5062163	4	23754013	23754121	4	652541	652647
5	26799727	26800044	5	5279025	5279494	5	23757237	23757322	5	654350	654487
6	26805108	26805407	6	5280067	5280161	6	23760311	23760510	6	656997	657067
7	26813282	26813497	7	5290701	5291010	7	23765237	23765343	7	662574	662644
8	26822359	26822380	8	5291162	5291319	8	23775909	23776056	8	664048	664092
9	26826200	26826291	9	5294392	5294478	9	23777579	23777747	9	665133	665299
10	26827726	26827805	10	5296983	5297085	10	23778920	23780516	10	673707	673830
11	26835454	26836288	11	5300146	5301562	11	23781857	23784109	11	675537	675777
transmembrane protease, serine 2			SON DNA binding protein			MAP3K7 C-terminal like			integrin subunit beta 2		
EXON			EXON			EXON			EXON		
1	28542273	28542350	1	20577230	20577361	1	16111663	16111905	1	31962846	31963065
2	28542142	28542257	2	20580405	20580571	2	16114744	16114900	2	31955013	31955242
3	28531050	28533258	3	20583668	20589583	3	16119264	16119335	3	31944917	31945018
4	28528548	28528770	4	20591348	20591508	4	16120005	16120109	4	31944476	31944607
5	28523699	28523785	5	20591738	20593248	5	16125693	16125768	5	31941107	31941287
6	28522586	28522705	6	20593422	20593568	6	16125860	16125953	6	31937557	31937727
7	28520659	28522216	7	20593779	20593967	7	16126446	16126773	7	31935684	31935925
8	28514668	28514794	8	20594069	20594859	8	16143232	16143983	8	31934512	31934667
9	28513364	28513474	9	20601349	20601659	9	16156246	16156405	9	31933259	31933354
10	28510769	28510812	10	20603163	20603279	10	16164677	16165115	10	31927109	31929252
11	28507517	28507688	11	20605943	20606197	11	16165153	16165242	11	31926001	31926188
12	28505998	28506173	12	20606561	20606822	12	16167498	16167606	12	31924170	31924414
13	28504840	28504935	13	20607500	20607647	13	16179751	16179899	13	31923468	31923687
14	28502588	28502730	14	20609788	20609859	14	16183381	16183442	14	31922885	31923087
15	28501926	28502078	15	20609970	20610085	15	16194133	16194248	15	31920145	31922808
16	28498743	28501280	16	20610557	20611706	16	16208904	16210081			
autoimmune regulator			interferon (alpha, beta and omega) receptor 1			BTB and CNC homology 1, basic leucine zipper transcription factor 1			serine/threonine-protein kinase SIK1		
EXON			EXON			EXON			EXON		
1	31320030	31320330	1	20358621	20358830	1	16332987	16333145	1	144541	144659
2	31320749	31320923	2	20359101	20359323	2	16333608	16333790	2	145477	145646
3	31321170	31321783	3	20369716	20369839	3	16339396	16339514	3	146146	146262

4	31322537	31322650	4	20375191	20375366	4	16355413	16355706	4	149631	149694
5	31323420	31323608	5	20377460	20377614	5	16360251	16361585	5	149869	150030
6	31323849	31323994	6	20377727	20377868	6	16363679	16363885	6	150301	150425
7	31324180	31324260	7	20379438	20379585	7	16376591	16380340	7	150535	150658
8	31324457	31325402	8	20382742	20383267	8	16385826	16386005	8	151211	151434
9	31326494	31326593	9	20383283	20383482	9	16386771	16386855	9	151663	151809
10	31327185	31327367	10	20383581	20383812	10	16388907	16392933	10	152190	152315
11	31327981	31328102	11	20386950	20387100	11	16393194	16393645	11	152431	152647
12	31328564	31328695	12	20387830	20387975	12	16395223	16395435	12	153127	153408
13	31330575	31330637	13	20389508	20394054	13	16395744	16396088	13	153894	154125
14	31331848	31332411							14	154551	157147
lipase, member I			SAM domain, SH3 domain and nuclear localization signals 1			protein C-ets-2			potassium channel, inwardly rectifying subfamily J, member 6		
EXON	NT_011512.12		EXON			EXON	NT_011512.12		EXON	NT_011512.12	
1	1244992	1245083	1	1617512	1617594	1	25839499	25839808	1	24950076	24950630
2	1241007	1241141	2	1616328	1616680	2	25840023	25840312	2	24944141	24945034
3	1223226	1223611	3	1579354	1580552	3	25842302	25845392	3	24874850	24874904
4	1220153	1220261	4	1559318	1559372	4	25847195	25847306	4	24849638	24850178
5	1215950	1216051	5	1557463	1557586	5	25848465	25848584	5	24816562	24817192
6	1200554	1200643	6	1555342	1555413	6	25848973	25849173	6	24753541	24753761
7	1199415	1199582	7	1551000	1551233	7	25851200	25851283	7	24751294	24751429
8	1197611	1197715	8	1546636	1546765	8	25852617	25852838	8	24748403	24749323
9	1187052	1187162	9	1544502	1544653	9	25853695	25853958	9	24706714	24709072
10	1186765	1186876	10	1534721	1534927	10	25855778	25855896	10	24692751	24695027
11	1178815	1178991	11	1532634	1532784	11	25856866	25859147	11	24684336	24684802
12	1143005	1143272	12	1519420	1520306				12	24658415	24659676
interferon gamma receptor 2 (interferon gamma transducer 1)			claudin 14			uncharacterized LOC105372795			uncharacterized LOC102724701		
EXON			EXON			EXON			EXON	NT_187309.1	
1	20437088	20437808	1	23610603	23610761	1	24107183	24107350	1	40341	40505
2	20449080	20449212	2	23576757	23577011	2	24107040	24107092	2	37971	38073
3	20455672	20455877	3	23544555	23544692	3	24106626	24106995	3	36248	36345
4	20461076	20461224	4	23513687	23514282	4	24106337	24106502	4	29338	29559
5	20466369	20466528	5	23511566	23511755	5	24103744	24103817	5	12489	14802
6	20466906	20467063	6	23500559	23500633	6	24101454	24101572	6	4177	8156
7	20471020	20471713	7	23494813	23495968	7	24099556	24099904	7	1	2256
crystallin alpha A			uncharacterized LOC105369308			uncharacterized LOC105369306			long intergenic non-protein coding RNA 1436		
EXON			EXON			EXON			EXON		
1	30203200	30205076	1	23670983	23671225	1	23139073	23139315	1	23039512	23040057
2	30206263	30207002	2	23666873	23667762	2	23141790	23143882	2	23041511	23042487
dual specificity tyrosine-(Y)-phosphorylation regulated kinase 1A			runt-related transcription factor 1			SR-related CTD-associated factor 4			phosphofructokinase, liver		

EXON	NT_011512.12		EXON			EXON			EXON		
1	24401126	24401820	1	22083031	22083536	1	18765855	18766310	1	31334226	31334382
2	24413017	24413321	2	22079301	22080101	2	18740466	18740549	2	31338467	31338647
3	24414664	24415466	3	22072117	22073618	3	18739615	18739659	3	31339380	31339609
4	24444762	24444862	4	22068334	22068349	4	18737957	18738118	4	31339992	31340127
5	24449682	24449757	5	22064419	22064520	5	18736436	18736571	5	31340873	31340946
6	24453097	24454576	6	21976023	21979232	6	18735968	18736110	6	31345198	31345275
7	24455966	24456058	7	21927117	21927162	7	18734951	18735363	7	31346297	31346486
8	24464461	24464592	8	21921035	21922882	8	18730761	18730942	8	31347170	31347335
9	24505883	24506253	9	21914749	21914905	9	18730305	18730413	9	31347830	31347874
10	24506876	24507072	10	21897765	21900500	10	18729005	18729172	10	31348104	31350531
11	24508761	24509649	11	21893526	21893770	11	18728396	18728481	11	31350624	31350716
12	24512373	24512492	12	21890481	21890639	12	18727486	18727676	12	31352662	31352787
13	24514830	24515018	13	21867153	21868793	13	18726541	18726641	13	31353543	31356067
14	24520659	24520806	14	21864044	21864163	14	18726009	18726122	14	31356325	31356395
15	24522977	24523055	15	21857656	21860598	15	18724989	18725145	15	31357154	31357241
16	24524367	24524653	16	21855469	21855888	16	18722499	18722656	16	31357958	31358110
17	24527209	24527355	17	21845564	21845576	17	18719760	18719925	17	31358683	31358847
18	24530310	24530450	18	21839616	21839620	18	18719590	18719676	18	31359048	31360252
19	24539475	24539781	19	21833493	21833656	19	18718187	18719432	19	31360351	31360456
20	24540291	24540546	20	21821993	21826802	20	18705192	18706546	20	31360907	31361573
21	24546103	24549568									
uncharacterized LOC105369301			carbonyl reductase 3			long intergenic non-protein coding RNA 1426			long intergenic non-protein coding RNA 1549		
EXON			EXON			EXON			EXON		
1	23464552	23464604	1	23169104	23169673	1	21779949	21780208	1	4473082	4473189
2	23480014	23480199	2	23172010	23172291	2	21781973	21782559	2	4475866	4476035
3	23494482	23494944	3	23180268	23180754	3	21816317	21816382	3	4478275	4478392
4	23514423	23515262				4	21816773	21819078	4	4483057	4483377
phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase, phosphoribosylaminoimidazole synthetase			transmembrane protease, serine 15			solute carrier family 37 (glucose-6-phosphate transporter), member 1			phosphodiesterase 9A		
EXON			EXON			EXON			EXON		
1	20576779	20577084	1	5437670	5437848	1	29533824	29533935	1	29687944	29688075
2	20576257	20576350	2	5432391	5432521	2	29536488	29536609	2	29720384	29720454
3	20573363	20573548	3	5432071	5432138	3	29542816	29542861	3	29722109	29722186
4	20569417	20569512	4	5417819	5417970	4	29543437	29543853	4	29730578	29733034
5	20568771	20568945	5	5413475	5413510	5	29548084	29548909	5	29733160	29733203
6	20566537	20566648	6	5406385	5406516	6	29549062	29549127	6	29758748	29758826
7	20565681	20565749	7	5399332	5399440	7	29552469	29552702	7	29765962	29766141
8	20564951	20565076	8	5393949	5394055	8	29559968	29560230	8	29766262	29766316

9	20563042	20563129	9	5387915	5388055	9	29568890	29569022	9	29767548	29767618
10	20562707	20562796	10	5387095	5387244	10	29569664	29569742	10	29777756	29777861
11	20562359	20562527	11	5378147	5378252	11	29573704	29573839	11	29777968	29778052
12	20558201	20559192	12	5377698	5377848	12	29576596	29576672	12	29785308	29785389
13	20557244	20557296	13	5375605	5375740	13	29577628	29577794	13	29788182	29788256
14	20556380	20556474	14	5366276	5366365	14	29581295	29581332	14	29789504	29789531
15	20555098	20555207	15	5363361	5363486	15	29588254	29588334	15	29790967	29793277
16	20554556	20554837	16	5360624	5360764	16	29593150	29593281	16	29794520	29794624
17	20551549	20551800	17	5349338	5349448	17	29596270	29596360	17	29795017	29795099
18	20551181	20551333	18	5347137	5347269	18	29598007	29598069	18	29796275	29796431
19	20545444	20545650	19	5331926	5332021	19	29598900	29598985	19	29799573	29800170
20	20543975	20544112	20	5328795	5328844	20	29600019	29600067	20	29802380	29802484
21	20540166	20540296	21	5328462	5328636	21	29601177	29601250	21	29803219	29803347
22	20539753	20539894	22	5315232	5315413	22	29602552	29602630	22	29803679	29806712
23	20538604	20538719	23	5313156	5313251	23	29609010	29609107	23	29809472	29809701
24	20538123	20538507	24	5309389	5309528	24	29609237	29611839			
			25	5303308	5304316	25	29613254	29615632			
uncharacterized LOC105372791			solute carrier family 5 (sodium/myo-inositol cotransporter), member 3			uncharacterized LOC105372788			IL10RB antisense RNA 1 (head to head)		
EXON			EXON			EXON			EXON		
1	21409429	21409540	1	21107715	21107937	1	21002425	21002765	1	20300255	20300452
2	21404994	21405352	2	21126477	21140454	2	21001293	21001542	2	20299824	20300094
downstream neighbor of SON			solute carrier family 19 (folate transporter), member 1			junctional adhesion molecule 2			Leber congenital amaurosis 5-like		
EXON			EXON			EXON			EXON		
1	20622513	20622900	1	32576560	32576663	1	12673474	12674080	1	26479642	26479936
2	20622150	20622158	2	32571959	32572200	2	12718075	12718140	2	26478327	26478394
3	20621714	20621794	3	32567625	32569974	3	12724058	12724165	3	26472876	26473128
4	20620170	20620373	4	32565581	32566340	4	12727948	12728100	4	26469612	26469765
5	20618782	20618960	5	32564962	32565163	5	12732869	12733071	5	26463323	26463403
6	20617680	20617858	6	32560009	32560150	6	12736362	12736461	6	26461610	26462695
7	20616357	20616438	7	32548907	32550332	7	12740171	12740278	7	26455751	26458101
8	20616143	20616247	8	32546759	32548219	8	12743626	12743641	8	26454898	26455035
9	20615494	20615692	9	32541699	32543360	9	12746532	12746574	9	26445910	26445994
10	20613542	20613754	10	32541395	32541656	10	12748423	12748765	10	26444456	26444559
11	20611743	20612636	11	32540994	32541329	11	12748832	12751754	11	26444171	26444288
			12	32540302	32540949				12	26440036	26440804
DnaJ heat shock protein family (Hsp40) member C28			long intergenic non-protein coding RNA 1548			uncharacterized LOC101928107			chromosome 21 open reading frame 62		
EXON			EXON			EXON			EXON		
1	20525794	20525915	1	20204333	20204428	1	20157804	20157895	1	19847784	19847935
2	20525258	20525675	2	20203888	20204025	2	20148578	20148815	2	19844766	19844846

3	20522247	20523616	3	20202635	20202773	3	20145891	20146921	3	19841834	19841908
			4	20199662	20199883				4	19824865	19828677
uncharacterized LOC102724502			uncharacterized LOC105372779			long intergenic non-protein coding RNA 159			homeobox protein cut-like 1		
EXON			EXON			EXON			EXON		
1	19993080	19993281	1	19342711	19342929	1	19120396	19120487	1	18593437	18593787
2	19994385	19994758	2	19340656	19341018	2	19117666	19117809	2	18594209	18594679
						3	19114508	19114709			
uncharacterized LOC105372776			long intergenic non-protein coding RNA 307			uncharacterized LOC339622			uncharacterized LOC101927869		
EXON			EXON			EXON			EXON		
1	17914836	17915747	1	17245926	17245975	1	11874742	11874920	1	11355442	11355569
2	17863231	17863924	2	17243930	17244117	2	11878971	11879121	2	11342408	11342525
3	17846889	17847444	3	17243343	17243528	3	12063976	12064051	3	11341811	11341951
						4	12089626	12091938	4	11338742	11340361
DSCAM intronic transcript 1			high mobility group nucleosome binding domain 1			uncharacterized LOC101928435			uncharacterized LOC101928398		
EXON			EXON			EXON			EXON		
1	27664850	27664959	1	26349899	26350007	1	26011903	26011966	1	25891731	25891817
2	27663732	27663833	2	26351237	26351401	2	26011645	26011759	2	25891968	25892147
3	27655447	27655592	3	26356994	26357410	3	26008621	26010039	3	25892269	25893054
4	27652699	27653060									
uncharacterized LOC105372801			Down syndrome critical region 10 (non-protein coding)			TTC3 antisense RNA 1			uncharacterized LOC105372842		
EXON			EXON			EXON			EXON		
1	25272305	25272393	1	25240348	25240552	1	24228000	24228118	1	33071244	33072090
2	25271409	25271606	2	25240662	25240808	2	24221804	24222835	2	33072649	33073999
			3	25242327	25242836						
Down syndrome critical region 8			uncharacterized LOC102724188			DIP2A intronic transcript 1			COL18A1 antisense RNA 2		
EXON			EXON			EXON			EXON		
1	25155643	25155793	1	580225	580315	1	33496663	33496714	1	32444140	32444257
2	25156504	25156860	2	581172	581320	2	33496904	33496955	2	32442649	32442755
3	25188607	25188721	3	590585	590703	3	33497258	33497420	3	32442200	32442291
4	25190496	25190709	4	592509	592653	4	33503338	33503498	4	32441578	32441812
uncharacterized LOC101928796			uncharacterized LOC100129027			uncharacterized LOC105372840			COL18A1 antisense RNA 1		
EXON			EXON			EXON			EXON		
1	33007162	33007368	1	32870478	32870611	1	32627870	32627900	1	32457962	32459262
2	33008286	33009145	2	32866800	32866850	2	32629765	32629853	2	32456604	32456853
			3	32862033	32865433	3	32630583	32631283	3	32453908	32455811
long intergenic non-protein coding RNA 316			uncharacterized LOC642852			long intergenic non-protein coding RNA 162			long intergenic non-protein coding RNA 163		
EXON			EXON			EXON			EXON		
1	32376017	32376182	1	32322244	32322436	1	32038636	32038919	1	32028096	32028278
2	32372782	32373070	2	32323543	32323630	2	32033400	32033848	2	32024056	32026027

			3	32324663	32331546						
long intergenic non-protein coding RNA 1547			small ubiquitin-like modifier 3			TSPEAR antisense RNA 2			uncharacterized LOC100505727		
EXON			EXON			EXOM			EXON		
1	31970147	31974105	1	31852140	31852321	1	31551407	31552027	1	31519763	31520422
2	31969819	31970082	2	31848054	31848296	2	31552451	31552524	2	31524223	31524480
3	31968917	31969288	3	31843239	31843310	3	31552820	31553169	3	31524586	31525331
4	31967476	31968560	4	31839809	31841232						
long intergenic non-protein coding RNA 1424			leucine rich repeat containing 3			uncharacterized LOC105372833			uncharacterized LOC105372832		
EXON			EXON			EXON			EXON		
1	31836433	31836828	1	31489678	31489847	1	31285005	31285133	1	31241395	31241591
2	31837939	31838909	2	31490690	31493048	2	31285319	31285712	2	31240717	31240982
LRRC3 antisense RNA 1 (head to head)			TRPM2 antisense RNA			uncharacterized LOC101930094			trefoil factor 1		
EXON			EXON			EXON			EXON		
1	31489436	31489476	1	31459342	31459755	1	29530731	29530857	1	29400603	29400727
2	31489169	31489244	2	31457568	31457673	2	29531015	29531125	2	29397447	29397599
3	31485599	31485961	3	31449227	31450932	3	29531425	29531635	3	29396474	29396687
4	31485178	31485481	4	31448780	31449136						
uncharacterized LOC105377138			apoptosis associated transcript in bladder cancer			uncharacterized LOC102723380			long intergenic non-protein coding RNA 319		
EXON			EXON			EXON			EXON		
1	31194017	31194268	1	30842993	30846759	1	30499501	30499539	1	30484216	30485113
2	31192932	31193546	2	30841156	30841431	2	30501250	30501490	2	30485365	30486697
			3	30839950	30840780				3	30487313	30488094
uncharacterized LOC101928399			long intergenic non-protein coding RNA 322			uncharacterized LOC101928284			uncharacterized LOC101928255		
EXON			EXON			EXON			EXON		
1	30395084	30396541	1	30362334	30366231	1	29775286	29775950	1	29684756	29685436
2	30392247	30393678	2	30360351	30360432	2	29767786	29768874	2	29682463	29684635
			3	30356401	30357041						
uncharacterized LOC101928233			UMODL1 antisense RNA 1			ZNF295 antisense RNA 1			long intergenic non-protein coding RNA 112		
EXON			EXON			EXON			EXON		
1	29649045	29649287	1	29142488	29142726	1	29043386	29043776	1	28750628	28750736
2	29633472	29635040	2	29136326	29138227	2	29056196	29059143	2	28751193	28751366
									3	28751534	28751774
trefoil factor 2			long intergenic non-protein coding RNA 111			long intergenic non-protein coding RNA 323			potassium channel, voltage gated subfamily E regulatory beta subunit 2		
EXON			EXON			EXON			EXON		
1	29385071	29385291	1	28713373	28713702	1	28182154	28182325	1	21398216	21398343
2	29384073	29384222	2	28722515	28722660	2	28181541	28181625	2	21399548	21399776
3	29381678	29381824	3	28731430	28731528	3	28179269	28179527	3	21400920	21401361
4	29380549	29380738				4	28175685	28177607	4	21404598	21405333

uncharacterized LOC105372784			long intergenic non-protein coding RNA 317			uncharacterized LOC105372746			CHODL antisense RNA 1		
EXON			EXON			EXON			EXON		
1	19993080	19993281	1	8771444	8771511	1	6933893	6933947	1	4919654	4919800
2	19994385	19994758	2	8761224	8761285	2	6932215	6932276	2	4874687	4874845
			3	8757485	8757966	3	6927471	6927765	3	4869864	4870018
C21orf91 overlapping transcript 1			uncharacterized LOC105379493			U2 small nuclear RNA auxiliary factor 1-like 5			alpha-crystallin A chain		
EXON			EXON NT_187307.1			EXON NT_187308.1			EXON NT_187308.1		
1	4826669	4826701	1	107124	107233	1	64362	67906	1	133475	133713
2	4821220	4821402	2	142987	143151	2	59064	60801	2	134943	135065
3	4819740	4819804	3	147184	147813	3	57365	58972	3	136497	136999
4	4811596	4811811									
uncharacterized LOC105372743			coxsackievirus and adenovirus receptor-like			uncharacterized LOC102723360			microRNA 3648-1		
EXON			EXON NT_187310.1			EXON NT_187314.1			EXON NT_187317.1		
1	4843137	4843168	1	507	621	1	130707	130852	1	108634	108813
2	4844687	4845037	2	8645	8700	2	124524	128287	2	109005	109065
			3	21772	23212	3	118507	122374			
small integral membrane protein 11B			histone-lysine N-methyltransferase 2C			spermatogenesis and centriole associated 1-like			ubiquitin conjugating enzyme E2G 2		
EXON NT_187315.1			EXON NT_029490.5			EXON			EXON		
1	1251	1338	1	161052	161078	1	33218548	33218651	1	31834182	31836028
2	5199	5287	2	161291	161407	2	33216816	33217966	2	31822252	31822287
3	11252	11418	3	166919	167020	3	33201976	33202850	3	31822112	31822157
4	14405	14930	4	167213	168046	4	33196108	33196259	4	31811491	31811609
						5	33195340	33195897	5	31807739	31809628
									6	31802772	31805681
uncharacterized LOC105372730			uncharacterized protein ENSP00000383407-like			chromosome 21 open reading frame 2			inducible T-cell co-stimulator ligand		
EXON NT_113954.1			EXON			EXON			EXON		
1	2362	3146	1	1677194	1677299	1	31373310	31373594	1	31274996	31275196
2	4561	4620	2	1650168	1650208	2	31372426	31372976	2	31272640	31272680
microRNA 3687-1			3	1647068	1647105	3	31371841	31371859	3	31271059	31271409
EXON NT_187317.1			4	1643674	1643760	4	31371016	31371112	4	31269464	31269754
1	108634	108813	5	1636215	1636291	5	31368253	31369996	5	31265472	31265636
2	109005	109065	6	1628205	1628270	6	31367225	31367454	6	31264246	31264281
			7	1625584	1626831	7	31364655	31366206	7	31261019	31263236
						8	31363136	31364518	8	31257183	31257270
uncharacterized LOC105369159			amyloid beta precursor protein			ubiquitin specific peptidase 25			collagen, type VI, alpha 2		
EXON			EXON			EXON			EXON		
1	3104714	3104759	1	13205177	13205320	1	2764217	2764630	1	33132311	33132365
2	3105306	3105590	2	13204756	13205012	2	2797083	2797160	2	33145642	33145783
3	3215783	3215879	3	13200089	13201516	3	2800189	2800333	3	33146171	33146769

4	3228571	3228825	4	13174345	13174652	4	2812096	2812219	4	33146995	33147016
5	3265248	3265307	5	13146171	13146338	5	2825694	2825856	5	33148200	33148265
6	3364796	3364903	6	13124135	13124264	6	2833949	2834035	6	33150064	33150117
7	3401555	3403425	7	13087428	13087540	7	2839313	2839450	7	33150201	33150245
8	3425806	3425901	8	13085192	13085385	8	2843001	2843077	8	33150569	33150595
9	3453520	3453933	9	13076967	13077924	9	2845329	2845402	9	33150843	33150869
10	3454647	3454764	10	13056032	13056234	10	2852890	2853038	10	33150962	33151006
11	3521682	3522087	11	13055267	13055518	11	2858231	2858358	11	33151592	33151645
12	3555911	3555996	12	13043788	13044105	12	2859158	2859253	12	33152066	33152128
13	3571473	3571631	13	13034207	13034376	13	2860396	2860557	13	33152806	33152868
14	3573281	3573361	14	13031552	13031608	14	2861169	2861395	14	33153222	33153311
15	3574020	3574103	15	13016536	13016669	15	2864723	2864793	15	33153980	33154042
16	3622881	3622939	16	13010146	13010220	16	2865593	2865821	16	33154707	33154769
17	3623545	3623732	17	13009262	13009420	17	2867540	2867740	17	33155253	33155315
18	3624429	3624517	18	13004586	13004741	18	2876590	2876732	18	33155748	33155810
19	3628532	3628646	19	12989819	12989947	19	2881855	2881968	19	33156300	33156350
20	3641186	3643967	20	12988782	12988881	20	2883969	2884064	20	33156688	33156723
21	3647642	3647673	21	12979716	12980167	21	2898460	2898638	21	33157067	33157129
22	3647794	3649716	22	12945933	12946154	22	2900458	2900536	22	33158843	33158905
23	3661369	3661589	23	12939216	12939269	23	2904260	2904339	23	33159077	33159112
24	3665019	3665164	24	12931765	12931866	24	2908595	2908718	24	33159458	33159503
25	3665633	3665681	25	12925914	12926060	25	2910188	2912183	25	33159657	33159809
26	3674422	3674875	26	12914742	12915963	26	2912495	2914264	26	33159977	33160429
									27	33160695	33160733
									28	33161773	33167041
uncharacterized LOC101927745			chromosome 21 open reading frame 33			U2 small nuclear RNA auxiliary factor 1			long intergenic non-protein coding RNA 479		
EXON			EXON			EXON			EXON		
1	2436180	2436497	1	31167804	31168030	1	30141643	30141770	1	28749934	28749967
2	2434690	2434741	2	31168294	31168346	2	30131832	30139444	2	28749312	28749475
3	2429655	2429689	3	31170252	31170365	3	30129886	30129935	3	28748426	28748544
4	2429192	2429321	4	31171369	31171488	4	30129630	30129728	4	28748212	28748324
5	2428665	2428698	5	31174442	31174534	5	30128847	30128980	5	28748048	28748128
6	2424770	2424810	6	31177397	31177553	6	30128160	30128755	6	28747743	28747864
7	2404714	2405103	7	31179013	31179915	7	30127148	30127441	7	28745712	28747192
long intergenic non-protein coding RNA 313			radial spoke head 1 homolog (Chlamydomonas)			uncharacterized LOC101927843			long intergenic non-protein coding RNA 320		
EXON			EXON			EXON			EXON		
1	30512256	30512415	1	29530325	29530546	1	9132505	9132651	1	7837187	7837408
2	30510164	30510231	2	29527158	29527271	2	9130721	9130778	2	7820228	7820280
3	30505704	30505886	3	29526950	29527055	3	9088502	9088561	3	7816442	7816638
4	30504625	30505510	4	29518869	29520653	4	9077137	9077300	4	7815762	7815853
5	30499065	30499165	5	29516829	29516900	5	9075366	9075627	5	7815541	7815666

6	30498784	30499008	6	29511483	29511636	6	9052080	9052348	6	7812835	7812988
7	30497532	30497681	7	29510090	29510239	7	9050492	9050602	7	7812610	7812739
8	30496650	30496717	8	29506678	29507062	8	9043136	9043550	8	7791503	7791609
9	30496152	30496453							9	7776782	7778025
phosphatidylinositol glycan anchor biosynthesis class P			long intergenic non-protein coding RNA 308			long intergenic non-protein coding RNA 1425			cysteine/tyrosine-rich 1		
EXON			EXON			EXON			EXON		
1	24107183	24107350	1	9132809	9132904	1	8781177	8781235	1	13607060	13607476
2	24107040	24107092	2	9143638	9143745	2	8782587	8782876	2	13600458	13600560
3	24106626	24106995	3	9145963	9146018	3	8813331	8813358	3	13598851	13599017
4	24106337	24106502	4	9148952	9149090	4	8819165	8819245	4	13516435	13517599
5	24103744	24103817	5	9149170	9149304	5	8820491	8820583	5	13514464	13514621
6	24101454	24101572	6	9150528	9150720	6	8831371	8831607	6	13511769	13512327
7	24099556	24099904							7	13500401	13502826
MIR548X host gene			mitochondrial ribosomal protein L39			ADAM metalloproteinase with thrombospondin type 1 motif 1			chromosome 21 open reading frame 58		
EXON			EXON			EXON			EXON		
1	5793966	5794195	1	12641595	12641681	1	13878417	13879601	1	33355214	33358091
2	5786940	5787220	2	12640641	12640847	2	13877934	13877969	2	33354829	33355017
3	5772116	5772217	3	12637988	12638127	3	13876531	13876877	3	33352204	33353329
4	5720282	5720356	4	12635560	12635659	4	13876050	13876182	4	33351400	33351540
5	5646254	5646485	5	12633991	12634058	5	13875190	13875903	5	33348908	33349739
6	5623719	5623812	6	12631494	12631606	6	13874067	13874754	6	33345648	33345759
7	5604430	5604603	7	12628085	12628150	7	13872631	13873954	7	33336677	33336768
8	5597507	5597564	8	12619848	12627157	8	13870479	13872470	8	33335322	33336346
9	5595457	5595883							9	33334622	33334986
PAX3 and PAX7 binding protein 1			glutamate receptor, ionotropic, kainate 1			ubiquitin specific peptidase 16			poly(rC) binding protein 3		
EXON			EXON			EXON			EXON		
1	19805518	19806050	1	16973575	16974225	1	16058809	16058969	1	32884153	32884287
2	19804006	19804134	2	16728088	16728255	2	16059901	16059963	2	32899958	32899986
3	19798540	19798716	3	16723920	16724177	3	16062065	16062166	3	32905046	32907537
4	19796288	19796509	4	16707175	16707356	4	16064787	16064965	4	32930400	32930554
5	19795251	19795354	5	16689002	16689055	5	16069029	16069132	5	32933791	32933814
6	19793548	19794186	6	16685310	16685483	6	16070463	16070566	6	32934783	32934815
7	19793272	19793461	7	16677018	16677161	7	16071468	16071655	7	32935189	32935305
8	19789405	19790986	8	16632986	16633129	8	16072527	16072622	8	32943547	32943678
9	19785311	19786102	9	16630718	16630762	9	16073218	16073348	9	32945094	32945496
10	19785109	19785268	10	16625304	16625417	10	16073673	16073760	10	32948143	32948286
11	19782691	19782890	11	16623031	16623234	11	16074801	16074879	11	32948488	32951591
12	19779766	19779910	12	16621558	16621781	12	16076205	16076296	12	32951780	32952839
13	19778984	19779105	13	16614887	16617493	13	16076664	16077791	13	32953025	32953675
14	19777870	19777946	14	16611156	16611373	14	16080859	16081513	14	32964109	32964187

15	19777440	19777506	15	16595816	16596041	15	16082953	16083047	15	32964978	32965037
16	19775531	19775814	16	16589244	16589494	16	16084284	16084370	16	32969445	32969497
17	19772365	19772511	17	16587737	16587896	17	16086315	16088150	17	32974222	32974391
18	19771446	19771600	18	16585038	16585668	18	16088258	16088682	18	32975862	32976646
19	19768091	19769259	19	16571125	16571577						
cysteine/tyrosine-rich 1 antisense RNA 1			long intergenic non-protein coding RNA 158			proteasome (prosome, macropain) assembly chaperone 1			uncharacterized LOC102724740		
EXON			EXON			EXON			EXON		
1	13440135	13440320	1	12465370	12465893	1	26217444	26217706	1	26051874	26051902
2	13450701	13450853	2	12463958	12464201	2	26215964	26216070	2	26045280	26045521
3	13504943	13505390	3	12458746	12458901	3	26214477	26214628	3	26042294	26042423
4	13537772	13537894	4	12454900	12454992	4	26214116	26214178	4	26040840	26040976
5	13584511	13586319	5	12430386	12430876	5	26212641	26212839	5	26039852	26040345
6	13602149	13603444	6	12420012	12420530	6	26211627	26211763	6	26025019	26025216
						7	26209638	26209856	7	26022899	26023618
uncharacterized LOC400867			long intergenic non-protein coding RNA 114			protein arginine methyltransferase 2			transmembrane protein 50B		
EXON			EXON			EXON			EXON		
1	25972002	25972621	1	25807641	25807669	1	33669359	33670739	1	20514030	20514203
2	25960217	25960469	2	25803037	25803166	2	33671088	33671182	2	20502979	20503118
3	25957618	25957885	3	25784697	25787643	3	33677478	33677831	3	20501202	20501314
4	25954934	25955031	4	25783430	25784696	4	33678498	33678680	4	20499534	20499601
5	25953253	25953713	5	25781589	25781652	5	33682650	33682811	5	20494605	20494697
6	25951492	25951592	6	25780491	25780608	6	33683767	33683931	6	20490196	20490289
7	25949183	25949340	7	25775513	25776709	7	33685998	33689556	7	20489919	20489976
8	25947361	25947609	8	25775076	25775260	8	33691015	33696128	8	20482974	20485055
9	25936093	25936157	9	25774508	25774606	9	33697575	33697746	9	20473406	20473489
10	25911483	25914564	10	25773091	25773521	10	33698487	33699316	10	20466678	20468843
long intergenic non-protein coding RNA 1423			family with sequence similarity 207 member A			uncharacterized LOC101930100			WRD1700		
EXON			EXON			EXON NT_113953.1			Exon		
1	25380087	25380268	1	31973889	31974061	1	78937	79136	1	29530325	29530546
2	25367508	25367613	2	31974189	31974378	2	79280	79736	2	29527158	29527271
3	25367154	25367224	3	31977876	31978029	3	79787	79874	3	29526950	29527055
4	25366618	25366669	4	31994292	31994360	4	81153	81690	4	29518869	29520653
5	25365913	25365980	5	32001226	32001373	5	97421	97445	5	29516829	29516900
6	25360544	25360737	6	32007389	32007457	6	108144	108232	6	29511483	29511636
			7	32010872	32011181				7	29510090	29510239
									8	29506678	29507062
beta-site APP-cleaving enzyme 2			chondrolectin			DNA (cytosine-5-)-methyltransferase 3-like			PBX/knotted 1 homeobox 1		
EXON			EXON			EXON			EXON		
1	28201749	28201821	1	4951532	4951592	1	31295932	31296408	1	30008702	30008856
2	28201993	28202767	2	5037776	5038479	2	31295346	31295458	2	30027392	30027421

3	28245161	28247392	3	5038548	5040538	3	31294987	31295031	3	30038514	30038624
4	28260458	28260546	4	5062064	5062163	4	31293824	31293903	4	30041683	30041810
5	28263142	28265016	5	5279025	5279494	5	31293629	31293741	5	30044245	30044416
6	28271705	28271921	6	5280067	5280161	6	31292715	31292886	6	30047260	30047430
7	28276011	28276139	7	5290701	5291010	7	31290259	31290346	7	30051100	30052244
8	28277568	28277702	8	5291162	5291319	8	31288809	31288897	8	30052325	30052422
9	28280150	28280255	9	5294392	5294478	9	31285815	31285890	9	30055495	30055623
10	28284944	28285093	10	5296983	5297085	10	31285002	31285141	10	30059063	30059139
11	28291350	28291518	11	5300146	5301562	11	31283218	31283303	11	30062894	30063148
12	28309563	28316726				12	31280531	31280757	12	30064082	30068123
protein O-fucosyltransferase 2			heat shock transcription factor 2 binding protein			uncharacterized LOC102724951			MCM3AP antisense RNA 1		
EXON			EXON			EXON NT_187313.1			EXON		
1	32321933	32322088	1	30693578	30693685	1	52772	52973	1	33263417	33263555
2	32319214	32320120	2	30692253	30692512	2	55241	55275	2	33263716	33264719
3	32317544	32317719	3	30690779	30690929	3	55375	56191	3	33268703	33269645
4	32316354	32316651	4	30678481	30678584	4	65098	65318	4	33270416	33270518
5	32314270	32314298	5	30667464	30667613	5	78783	80726	5	33275026	33277191
6	32312295	32312361	6	30664514	30664646	6	83949	84202	6	33277599	33277942
7	32311425	32311602	7	30648022	30648139	7	84962	85508	7	33278743	33279057
8	32311209	32311334	8	30644079	30644181	8	85584	85638	8	33281082	33281423
9	32304031	32304211	9	30626417	30626520	9	87276	87925	9	33281827	33281874
10	32298120	32301905	10	30563384	30564154	10	89158	91143	10	33283169	33285893
formiminotransferase-cyclodeaminase			pyridoxal (pyridoxine, vitamin B6) kinase			ribosomal RNA processing 1B			MX dynamin-like GTPase 2		
EXON			EXON			EXON			EXON		
1	33189662	33189759	1	30753289	30753573	1	30693740	30693986	1	28396215	28396247
2	33188341	33188524	2	30763092	30763202	2	30704076	30704158	2	28397461	28401498
3	33187099	33187227	3	30766331	30766839	3	30706493	30706557	3	28404277	28404541
4	33186084	33186197	4	30768256	30768315	4	30708062	30708147	4	28411028	28411347
5	33185750	33185929	5	30770865	30773053	5	30708828	30708889	5	28411981	28412173
6	33184580	33184717	6	30775657	30775766	6	30709226	30709355	6	28414209	28414343
7	33184311	33184442	7	30775857	30775963	7	30710464	30710528	7	28416602	28416756
8	33180458	33180519	8	30777916	30777999	8	30710925	30711106	8	28424757	28424895
9	33180010	33180139	9	30780271	30780317	9	30717471	30717565	9	28429779	28429977
10	33179609	33179770	10	30783187	30783272	10	30718745	30718842	10	28431805	28431883
11	33176301	33176933	11	30784692	30784737	11	30719962	30719981	11	28433089	28433211
12	33172223	33173455	12	30786710	30786821	12	30720324	30721127	12	28433388	28433529
13	33171431	33171526	13	30787775	30787911	13	30721708	30722432	13	28435057	28436320
14	33170971	33171265	14	30789890	30789956	14	30724480	30724632	14	28436826	28439065
15	33170454	33170686	15	30790143	30796499	15	30725631	30725694	15	28440936	28441190
						16	30727382	30730271	16	28442183	28443134
pituitary tumor-transforming 1			uncharacterized LOC441058			receptor interacting serine/threonine			transmembrane protease, serine 3		

interacting protein						kinase 4					
EXON			EXON	NT_187319.1		EXON			EXON		
1	31907694	31908095	1	293743	293918	1	28801052	28801298	1	29430134	29431038
2	31900617	31903269	2	286694	286866	2	28800192	28800384	2	29429516	29429660
3	31900567	31900591	3	285284	285459	3	28800110	28800150	3	29424119	29424610
4	31899587	31899639	4	282530	282580	4	28790717	28791008	4	29423121	29423341
5	31895355	31895463	5	281924	282038	5	28785289	28785437	5	29422595	29422718
6	31890385	31890556	6	281031	281218	6	28783346	28783395	6	29419601	29419726
7	31889402	31889448	7	280592	280831	7	28780964	28781033	7	29418162	29418205
8	31883777	31886902	8	278951	279959	8	28780805	28780963	8	29416068	29417943
			9	265412	265499	9	28779951	28780054	9	29414309	29414404
			10	256105	256162	10	28778074	28778332	10	29410733	29410878
			11	254705	255161	11	28773560	28776225	11	29409908	29410060
									12	29406029	29406971
1-acylglycerol-3-phosphate O-acyltransferase 3			ubiquitin associated and SH3 domain containing A			C2 calcium-dependent domain containing 2			cystathionine-beta-synthase like		
EXON			EXON			EXON			EXON		
1	30899427	30899537	1	29438054	29438250	1	28987562	28988082	1	40645	40782
2	30911992	30913558	2	29440500	29440553	2	28982297	28982474	2	37686	37761
3	30916736	30916807	3	29443614	29443800	3	28976339	28976437	3	36405	36621
4	30938083	30938211	4	29444308	29444780	4	28960764	28960882	4	32929	33035
5	30959200	30959643	5	29447216	29447414	5	28956164	28956277	5	30663	30797
6	30981122	30981250	6	29447602	29447715	6	28953048	28953225	6	30036	30115
7	30982466	30982521	7	29450634	29450803	7	28952297	28952419	7	29807	29941
8	30988742	30989434	8	29452593	29452801	8	28948790	28948913	8	29623	29692
9	30993826	30994051	9	29460889	29461012	9	28946524	28946632	9	28320	28411
10	30994824	30994970	10	29466295	29466394	10	28943651	28943715	10	27373	27498
11	30999121	31002307	11	29469024	29469701	11	28941852	28942994	11	26731	26815
12	31003310	31003471	12	29471680	29471788	12	28941184	28941358	12	24861	24966
13	31004845	31004998	13	29476644	29476788	13	28939916	28940029	13	23646	23723
14	31005580	31005682	14	29477346	29477610	14	28935148	28935941	14	23254	23388
15	31012238	31012313	15	29478726	29478835	15	28933245	28933554	15	22565	22673
16	31015180	31015379	16	29481249	29481872	16	28929391	28929582	16	21223	21307
17	31016496	31021786				17	28927891	28929320	17	18282	18403
						18	28919302	28923536	18	17611	18067
ribosomal RNA processing 1			salt inducible kinase 1			WD repeat domain 4			family with sequence similarity 3 member B		
EXON			EXON			EXON			EXON		
1	30823729	30823954	1	30461202	30461314	1	29913599	29913781	1	28338442	28338502
2	30825542	30825624	2	30460215	30460384	2	29913184	29913467	2	28350552	28350760
3	30826864	30826921	3	30459599	30459715	3	29910894	29910959	3	28350926	28351090
4	30827511	30827596	4	30456166	30456229	4	29907743	29907883	4	28353650	28355170
5	30829381	30829442	5	30455830	30455991	5	29897632	29897788	5	28356780	28356896

6	30831553	30831743	6	30455435	30455559	6	29896474	29896586	6	28357115	28357258
7	30831823	30831887	7	30455202	30455325	7	29893854	29893914	7	28372570	28372693
8	30832099	30832292	8	30454426	30454649	8	29889874	29889972	8	28378668	28378726
9	30833762	30833841	9	30454051	30454197	9	29888754	29888818	9	28379878	28380599
10	30834709	30834806	10	30453545	30453670	10	29887761	29887944	10	28381205	28381292
11	30835054	30835073	11	30453213	30453429	11	29886447	29886516	11	28382784	28382916
12	30836466	30836752	12	30452452	30452733	12	29883418	29884434	12	28391300	28391919
13	30837408	30838294	13	30451735	30451966	13	29877625	29880162			
			14	30448707	30451309	14	29877272	29877473			
dopey family member 2			collagen, type VI, alpha 1			PR domain containing 15			transient receptor potential cation channel, subfamily M, member 2		
EXON			EXON			EXON			EXON		
1	23198727	23198785	1	33015941	33016139	1	28913462	28913674	1	31384355	31384467
2	23198900	23199063	2	33016826	33016955	2	28913188	28913332	2	31387448	31388057
3	23233262	23233443	3	33018461	33018661	3	28912887	28913052	3	31388840	31388928
4	23234523	23234693	4	33020718	33020877	4	28911726	28911805	4	31398306	31398474
5	23242907	23243096	5	33021136	33021264	5	28911464	28911551	5	31400946	31401126
6	23245745	23245843	6	33021347	33021367	6	28908948	28909226	6	31403313	31403535
7	23246166	23246289	7	33021691	33021711	7	28905686	28905813	7	31410025	31410205
8	23248273	23248382	8	33021802	33021846	8	28901492	28901569	8	31411904	31411965
9	23248634	23248748	9	33023276	33023329	9	28897427	28897494	9	31413189	31413389
10	23253564	23253684	10	33023800	33023844	10	28896117	28896199	10	31416910	31417012
11	23257423	23257542	11	33023944	33023970	11	28895755	28895943	11	31425096	31425217
12	23259757	23259859	12	33024450	33024476	12	28894519	28894564	12	31425464	31425817
13	23261878	23262069	13	33024570	33024614	13	28893784	28893877	13	31429606	31429743
14	23264642	23265326	14	33024965	33025018	14	28893223	28893420	14	31431939	31432068
15	23266996	23267267	15	33025171	33025233	15	28891368	28891521	15	31433488	31433633
16	23271454	23271606	16	33026202	33026264	16	28888758	28889010	16	31434451	31434563
17	23272793	23272893	17	33026355	33026409	17	28881282	28881383	17	31435873	31436089
18	23273957	23274147	18	33026555	33026590	18	28873815	28874045	18	31439331	31439452
19	23279240	23280869	19	33026940	33027002	19	28872126	28872255	19	31440097	31440229
20	23281709	23281820	20	33028359	33028421	20	28870660	28870841	20	31440786	31440957
21	23282572	23282760	21	33031613	33031675	21	28870305	28870399	21	31448083	31448266
22	23285354	23285476	22	33031892	33031954	22	28869629	28869716	22	31452119	31452300
23	23287964	23288101	23	33032313	33032363	23	28868697	28868756	23	31452615	31452747
24	23294869	23296107	24	33032590	33032625	24	28862358	28862525	24	31456462	31456611
25	23297738	23297842	25	33033089	33033151	25	28860152	28860246	25	31457837	31457924
26	23297940	23298006	26	33033345	33033410	26	28857510	28857641	26	31459044	31459131
27	23304204	23304349	27	33033849	33033884	27	28856095	28856229	27	31459721	31460019
28	23311213	23311292	28	33034523	33034559	28	28855259	28855422	28	31460852	31460928
29	23312167	23312276	29	33034951	33034959	29	28854287	28854366	29	31461202	31461303
30	23312401	23312547	30	33035389	33035578	30	28853774	28853893	30	31469323	31469409

31	23314477	23314538	31	33036153	33036262	31	28849897	28850028	31	31471254	31471359
32	23315675	23315803	32	33036410	33036593	32	28844346	28849154	32	31473259	31473360
33	23322206	23322342	33	33036719	33036902	33	28838726	28838806	33	31474981	31475097
34	23322948	23323003	34	33037312	33037341	34	28836904	28837113	34	31475884	31477273
35	23323237	23323398	35	33037583	33039241	35	28832417	28835914			
36	23326296	23326425									
37	23327512	23328466									
OTTHUMP00000067855			tetratricopeptide repeat domain 3			collagen, type XVIII, alpha 1			pericentrin		
EXON			EXON			EXON			EXON		
1	20676592	20676905	1	24107463	24107665	1	32439329	32439433	1	33358314	33358474
2	20679820	20681198	2	24115788	24118523	2	32439571	32439665	2	33359080	33359402
3	20683108	20684238	3	24119982	24121593	3	32489681	32491073	3	33360569	33360781
4	20752989	20753048	4	24122025	24122538	4	32502434	32502978	4	33368589	33368960
5	20755370	20755462	5	24122991	24123078	5	32508087	32508173	5	33380320	33380400
6	20756780	20756843	6	24124425	24124478	6	32509668	32509727	6	33380915	33381190
7	20769236	20770925	7	24125485	24125818	7	32510543	32510672	7	33381649	33381704
8	20775009	20775799	8	24127501	24127569	8	32511603	32511679	8	33383204	33383378
9	20776518	20776964	9	24128197	24128282	9	32511942	32512157	9	33383876	33384012
10	20783992	20784514	10	24129542	24129636	10	32512458	32512518	10	33385621	33385732
11	20786002	20786098	11	24130773	24130835	11	32512519	32512545	11	33387296	33387518
12	20789489	20789589	12	24135006	24135535	12	32514094	32514156	12	33388179	33388260
13	20796115	20796178	13	24141930	24142638	13	32514262	32514348	13	33389644	33389818
14	20800067	20800204	14	24155429	24155534	14	32514659	32514712	14	33391166	33391383
15	20801905	20802020	15	24156009	24156171	15	32514892	32515050	15	33397672	33398126
16	20806231	20806515	16	24157175	24157220	16	32516155	32516217	16	33400776	33401331
17	20808921	20809070	17	24158811	24158934	17	32516987	32517013	17	33415886	33416032
18	20809160	20809300	18	24160272	24160335	18	32518309	32519099	18	33420024	33420175
19	20815653	20815740	19	24163195	24163255	19	32521053	32521184	19	33422934	33423076
20	20816186	20816325	20	24166874	24166958	20	32521639	32521701	20	33423391	33423623
21	20828533	20828660	21	24169572	24169706	21	32522610	32522636	21	33424862	33425024
22	20831571	20831800	22	24172826	24172906	22	32523678	32523713	22	33425356	33425568
23	20834000	20834121	23	24174753	24174865	23	32524467	32524538	23	33431457	33431686
24	20836622	20836636	24	24178717	24178837	24	32525028	32525063	24	33432206	33432322
25	20845167	20845414	25	24181673	24181795	25	32525417	32525506	25	33432427	33432447
26	20848105	20848264	26	24182738	24182839	26	32526727	32526756	26	33433782	33433988
27	20851423	20851806	27	24184270	24184362	27	32526879	32526905	27	33435743	33435913
28	20852459	20852664	28	24185012	24185076	28	32527355	32527417	28	33436523	33436675
29	20853433	20853515	29	24186085	24186221	29	32527693	32527767	29	33445381	33446259
30	20857679	20857845	30	24187143	24187515	30	32528737	32528763	30	33447029	33447184
31	20857845	20861055	31	24190847	24191098	31	32529054	32529107	31	33450261	33451031
32	20863160	20863937	32	24193891	24193937	32	32529550	32529624	32	33452396	33452498
33	20868499	20868616	33	24194994	24195050	33	32530692	32530760	33	33456162	33456316

34	20870633	20870825	34	24196182	24196255	34	32531242	32531284	34	33460023	33460163
35	20871179	20872690	35	24198243	24198407	35	32531791	32531853	35	33461814	33461987
36	20890928	20891049	36	24199742	24200807	36	32538203	32538246	36	33462587	33462782
37	20892878	20892984	37	24201747	24201812	37	32538608	32538748	37	33464202	33464424
38	20899343	20899526	38	24206787	24206936	38	32539326	32539470	38	33464699	33464849
39	20901425	20901523	39	24216966	24217105	39	32539550	32539623	39	33465721	33466407
40	20909546	20909713	40	24219898	24220271	40	32540027	32540158	40	33470096	33470340
41	20916435	20916647	41	24221241	24221337	41	32540211	32540938	41	33471171	33471273
42	20917742	20917863	42	24222687	24222787	42	32540983	32541370	42	33472356	33472529
43	20919233	20919315	43	24225526	24225616	43	32541436	32541649	43	33474275	33474394
44	20919631	20919714	44	24226304	24226405	44	32541692	32541785	44	33475047	33475276
45	20920479	20920652	45	24229867	24230228	45	32543548	32543793	45	33476689	33476765
46	20922344	20923497	46	24231762	24231888	46	32544256	32544453	46	33478002	33478140
			47	24232074	24232217	47	32545303	32545418	47	33478886	33479013
			48	24232894	24237299	48	32546380	32547912	48	33479476	33479961
synaptojanin 1			listerin E3 ubiquitin protein ligase 1			T-cell lymphoma invasion and metastasis 1			minichromosome maintenance complex component 3 associated protein		
EXON			EXON						EXON		
1	19762138	19762232	1	16026956	16027148	EXON			1	33320343	33320400
2	19762029	19762131	2	16020923	16021126	1	18594111	18594169	2	33318260	33319586
3	19760964	19761109	3	16020331	16020429	2	18593119	18593169	3	33317807	33318030
4	19736153	19736239	4	16018884	16019114	3	18498175	18498227	4	33314689	33314767
5	19734030	19734297	5	16016508	16016560	4	18373435	18373614	5	33314185	33314329
6	19729249	19729474	6	16015311	16015491	5	18310924	18311100	6	33311719	33311909
7	19728420	19728503	7	16005463	16005636	6	18300202	18301175	7	33309378	33309517
8	19722498	19722559	8	16004744	16004934	7	18285934	18286381	8	33307580	33307777
9	19721170	19721266	9	16003658	16003793	8	18279680	18279852	9	33306753	33307021
10	19719940	19720109	10	16001794	16002165	9	18259918	18260142	10	33304317	33306076
11	19718230	19718311	11	16000562	16001371	10	18257598	18257783	11	33301174	33302254
12	19715688	19715840	12	16000057	16000098	11	18251745	18251891	12	33300117	33301144
13	19712837	19713002	13	15998769	15998958	12	18246823	18247664	13	33299513	33299715
14	19710524	19710547	14	15995509	15995725	13	18244237	18244407	14	33298309	33298409
15	19707532	19707723	15	15994709	15994898	14	18237100	18237204	15	33295472	33295603
16	19704480	19704564	16	15993650	15993889	15	18229416	18229497	16	33294985	33295098
17	19700625	19700765	17	15992578	15992731	16	18221193	18221279	17	33293131	33293457
18	19700135	19700327	18	15991524	15991668	17	18216613	18216837	18	33290981	33291178
19	19699105	19699263	19	15990954	15991140	18	18199154	18199257	19	33288968	33289036
20	19691908	19692064	20	15987409	15987568	19	18188439	18188618	20	33288584	33288718
21	19691195	19691312	21	15986352	15986456	20	18187258	18187326	21	33286622	33288325
22	19690879	19691094	22	15981656	15981798	21	18186828	18186953	22	33285721	33286313
23	19687480	19687558	23	15980344	15980479	22	18181087	18181195	23	33282125	33282979
24	19684376	19684538	24	15979999	15980143	23	18175517	18175696	24	33280499	33281352

25	19680585	19680794	25	15978575	15978788	24	18175310	18175428	25	33278999	33279389
26	19678170	19679981	26	15977859	15978096	25	18170125	18170233	26	33277657	33277914
27	19677602	19677649	27	15977454	15977528	26	18165082	18165140	27	33276994	33277123
28	19676287	19676325	28	15975412	15975598	27	18164405	18164507	28	33275003	33275209
29	19676088	19676158	29	15970718	15970889	28	18161629	18163720	29	33271021	33271171
30	19673863	19673971	30	15969301	15969521	29	18161257	18161344	30	33269317	33269618
31	19673100	19673317	31	15966662	15966856	30	18158654	18158886			
32	19669053	19669076	32	15965347	15965514	31	18152608	18155556			
33	19662951	19665993	33	15962336	15964702						
URB1 ribosome biogenesis 1 homolog (S. cerevisiae)			Down syndrome cell adhesion molecule			bromodomain and WD repeat domain containing 1			disco interacting protein 2 homolog A		
EXON			EXON			EXON			EXON		
1	19426961	19427195	1	27880811	27881331	1	26347635	26347978	1	33493134	33493414
2	19419737	19419876	2	27742646	27742964	2	26347433	26347491	2	33496663	33496714
3	19418505	19418656	3	27727002	27727148	3	26347264	26347293	3	33496904	33496955
4	19417614	19417746	4	27673277	27675091	4	26347033	26347092	4	33497258	33497420
5	19412637	19412733	5	27664850	27664959	5	26330629	26332774	5	33503338	33503498
6	19411379	19411514	6	27663732	27663833	6	26330457	26330555	6	33518949	33519020
7	19409590	19409675	7	27655447	27655592	7	26329935	26330095	7	33524792	33524911
8	19407839	19407964	8	27652699	27653060	8	26328003	26328224	8	33531180	33531299
9	19406699	19406823	9	27649570	27649675	9	26314340	26314440	9	33532774	33533025
10	19402595	19402790	10	27403291	27403437	10	26312935	26313005	10	33538505	33538681
11	19400810	19400947	11	27387657	27387935	11	26311443	26311543	11	33543449	33543568
12	19397348	19397521	12	27381862	27382137	12	26310365	26310405	12	33545609	33545806
13	19396084	19396213	13	27373311	27373607	13	26308566	26308664	13	33563284	33563375
14	19395199	19395315	14	27372293	27372568	14	26304475	26304625	14	33566319	33566429
15	19393988	19394100	15	27346273	27346551	15	26304091	26304225	15	33567716	33567839
16	19391729	19391848	16	27330247	27330366	16	26299083	26299211	16	33568196	33568305
17	19389641	19389757	17	27311844	27314777	17	26298652	26298877	17	33568777	33568879
18	19389051	19389189	18	27310289	27310462	18	26292679	26292864	18	33571416	33571480
19	19388125	19388295	19	27246544	27246623	19	26289837	26290020	19	33571638	33571731
20	19386902	19387098	20	27223234	27223430	20	26284988	26285081	20	33572675	33572794
21	19384896	19385114	21	27222083	27222179	21	26281893	26282024	21	33572927	33574183
22	19383496	19383675	22	27221323	27221451	22	26272671	26272765	22	33575948	33576087
23	19381148	19382003	23	27213119	27213286	23	26270787	26270976	23	33579329	33579465
24	19379566	19379767	24	27201410	27201480	24	26266565	26266690	24	33580073	33580498
25	19378762	19378948	25	27178683	27178923	25	26266369	26266476	25	33581107	33581399
26	19375658	19375716	26	27176750	27176896	26	26263504	26263628	26	33583963	33584302
27	19372899	19373092	27	27168046	27168201	27	26262692	26262774	27	33584735	33584936
28	19371596	19371706	28	27158387	27158520	28	26259278	26259389	28	33585826	33585935
29	19371286	19371349	29	27127913	27128066	29	26258600	26258661	29	33586016	33586096
30	19368355	19368526	30	27121362	27121479	30	26253594	26253997	30	33587103	33588484

31	19367509	19367611	31	27119794	27119957	31	26252697	26252852	31	33588767	33588888
32	19359421	19359581	32	27118100	27118198	32	26252344	26252464	32	33589014	33589125
33	19358683	19358794	33	27114344	27114532	33	26251285	26251369	33	33589711	33591945
34	19356670	19356776	34	27112879	27113169	34	26249429	26249554	34	33592415	33593298
35	19355993	19356136	35	27109229	27109405	35	26247673	26247745	35	33594914	33594975
36	19354723	19354832	36	27097061	27097091	36	26246858	26246899	36	33595477	33595997
37	19353409	19353606	37	27089917	27090032	37	26244978	26245121	37	33597975	33598124
38	19351868	19352109	38	27086150	27086299	38	26244187	26244339	38	33599905	33600079
39	19350658	19351257	39	27078270	27078467	39	26240300	26240466	39	33600752	33600875
40	19345211	19349291	40	27076563	27076865	40	26236517	26236737	40	33601562	33604205
			41	27046606	27047578	41	26234411	26234578			
						42	26232955	26233854			
						43	26219670	26231607			
v-ets avian erythroblastosis virus E26 oncogene homolog			OTTHUMP00000109038			MX dynamin-like GTPase 1			adenosine deaminase, RNA-specific, B1		
EXON			EXON			EXON			EXON		
1	25695850	25695972	1	24024350	24024437	1	28454707	28455009	1	32108761	32108985
2	25694714	25694859	2	24014976	24015156	2	28456030	28456287	2	32151125	32151230
3	25619036	25619140	3	24014564	24014820	3	28456948	28457103	3	32152756	32152811
4	25609854	25609939	4	24000636	24000848	4	28460223	28460455	4	32154986	32155151
5	25575327	25575743	5	23996228	23996362	5	28460462	28461496	5	32162594	32162765
6	25532554	25532696	6	23981525	23982479	6	28461946	28464113	6	32166407	32166498
7	25519671	25521496	7	23980201	23980346	7	28464725	28464800	7	32167751	32167874
8	25508219	25508313	8	23973024	23973186	8	28465906	28466367	8	32168907	32169024
9	25504967	25506272	9	23970641	23971584	9	28470029	28470221	9	32178823	32179154
10	25479596	25479819	10	23964443	23964625	10	28471207	28471344	10	32179189	32181397
11	25461953	25461956	11	23931052	23931323	11	28473886	28474040	11	32186800	32186900
12	25457602	25457753	12	23917434	23918605	12	28475079	28475217	12	32205802	32205876
13	25437698	25437901	13	23894906	23894936	13	28475908	28476106	13	32209922	32210856
14	25436749	25436829	14	23840228	23840476	14	28477980	28478058	14	32214522	32214636
15	25434264	25434840	15	23838176	23838457	15	28479640	28479762	15	32216777	32216945
16	25433183	25433352	16	23827506	23829733	16	28480192	28480333	16	32217554	32217731
17	25432804	25432986	17	23820831	23822181	17	28483329	28483487	17	32218666	32218785
18	25426142	25426636	18	23801410	23801477	18	28485359	28485435	18	32219108	32219283
19	25425851	25425907	19	23799204	23799364	19	28486813	28487061	19	32238747	32238928
20	25425187	25425234	20	23793919	23794033	20	28492720	28493406	20	32255028	32255206
21	25414219	25418115	21	23790734	23790947				21	32256210	32260755
22	25401453	25401840	22	23785080	23788609						
cystathionine-beta-synthase			lanosterol synthase (2,3- oxidosqualene-lanosterol cyclase)			PWP2 periodic tryptophan protein homolog (yeast)					
EXON			EXON			EXON					
1	30110932	30111053	1	33262626	33263016	1	31141519	31141626			
2	30110417	30110554	2	33261744	33261882	2	31143176	31143288			

3	30109962	30110122	3	33256822	33256930	3	31147945	31148039
4	30107456	30107533	4	33256046	33256167	4	31148371	31148470
5	30106177	30106393	5	33253668	33253764	5	31148817	31148960
6	30104173	30104228	6	33250742	33250767	6	31149456	31149591
7	30102701	30102807	7	33250581	33250716	7	31149883	31150112
8	30101970	30102596	8	33249877	33249985	8	31152012	31152155
9	30100435	30100569	9	33249372	33249490	9	31152955	31153025
10	30099808	30099887	10	33247930	33248027	10	31153579	31153715
11	30099579	30099713	11	33247217	33247244	11	31154540	31154688
12	30099395	30099464	12	33244880	33244936	12	31154823	31154970
13	30098092	30098183	13	33243746	33243817	13	31155144	31155294
14	30097145	30097270	14	33242443	33242493	14	31156369	31156549
15	30096503	30096587	15	33241620	33241769	15	31158772	31158919
16	30094633	30094738	16	33240864	33240960	16	31160203	31160311
17	30093418	30093495	17	33240028	33240133	17	31161092	31161157
18	30093026	30093160	18	33230394	33230459	18	31162124	31162342
19	30092337	30092445	19	33229868	33229948	19	31162439	31162589
20	30090995	30091079	20	33226073	33228853	20	31162744	31162818
21	30087383	30088175	21	33223316	33225427	21	31164789	31165373
			22	33222638	33223013			
uromodulin like 1			periodic tryptophan protein 2		thrombospondin-type laminin G			
			homolog		domain and EAR repeats			
EXON			EXON	NT_187300.1		EXON		
1	29097151	29097406	1	144623	144734	1	31745625	31745772
2	29105509	29105584	2	142961	143073	2	31724737	31724798
3	29110197	29110439	3	138210	138304	3	31715768	31716355
4	29118276	29118437	4	137779	137878	4	31700381	31701119
5	29119483	29119666	5	137289	137432	5	31688405	31688851
6	29120722	29120818	6	136658	136793	6	31625460	31626697
7	29122486	29122672	7	136137	136366	7	31613647	31614796
8	29124490	29124630	8	134091	134234	8	31607972	31608008
9	29127596	29127714	9	133222	133292	9	31601977	31602197
10	29128063	29128216	10	132532	132896	10	31591982	31592952
11	29133118	29133372	11	131559	131707	11	31584549	31585697
12	29135900	29136009	12	131277	131424	12	31573173	31574387
13	29136358	29136470	13	129698	131103	13	31567877	31568115
14	29138060	29138279	14	125935	127474	14	31565235	31565325
15	29143754	29143891	15	124199	125154	15	31563990	31564146
16	29145072	29145902	16	123902	124120	16	31562644	31562775
17	29147126	29147310	17	123655	123805	17	31561484	31561710
18	29147765	29148022	18	123426	123500	18	31559845	31560031

19	29150065	29150177	19	120871	121455	19	31556075	31556304
20	29153303	29153516				20	31543391	31543578
21	29154711	29155143				21	31542699	31543103
22	29155279	29155416				22	31538972	31539073
23	29157023	29157342				23	31532084	31534128
24	29160537	29160682						
25	29161198	29161434						
26	29161864	29162023						
27	29162196	29162337						
28	29163905	29163989						
29	29171631	29171833						
30	29176288	29177187						

6.1 Sequence Analysis Tools:

6.1.1 Microsoft Excel: Microsoft Excel was used for statistical analysis of data of compressed sequence. Functions of Microsoft excel used were:

MIN, AVERAGE, SUM,

6.1.2 Notepad ++: Notepad ++ was used to designed an algorithm for compression and decompression of DNA sequence. Functions of Notepad++ used were:

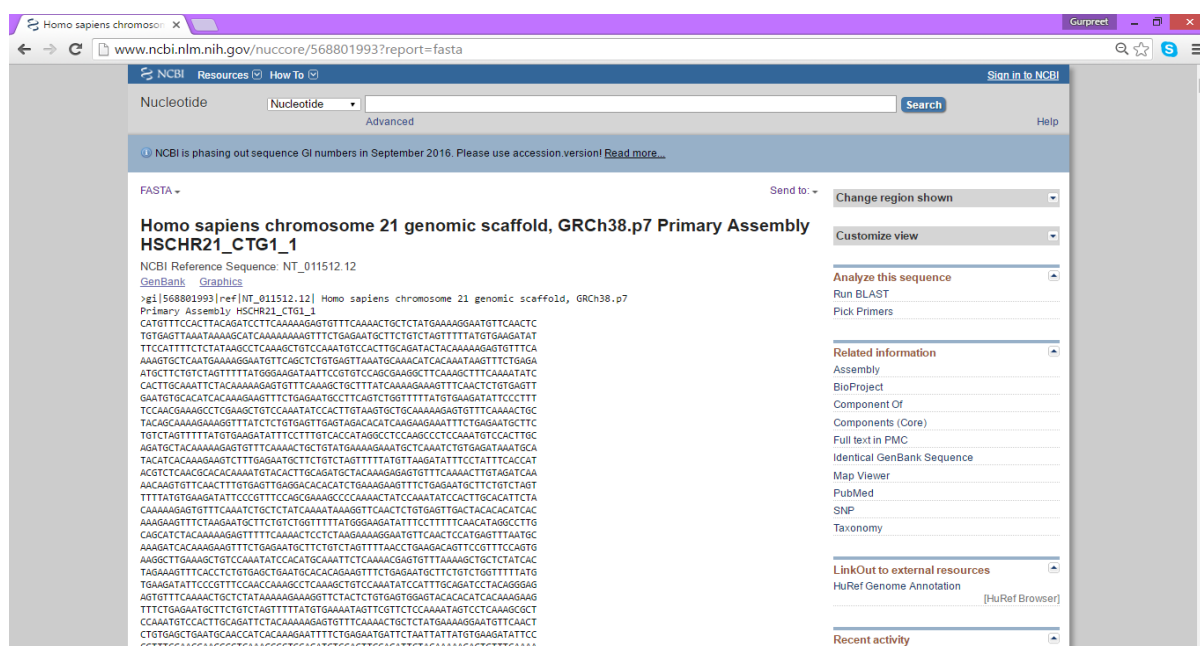
Record a Macro, Save Current recorded Macro, Run a Macro Multiple Times.

6.1.3 SPSS: SPSS tool was used to analyse the data by doing two sample t-test, ANOVA and Tukey's test.

6.2 Methods:

Nucleotide sequences in the FASTA format were downloaded from the Nucleotide database of NCBI website (www.ncbi.nlm.nih.gov/). Nucleotide sequence of chromosome 21 of *Homo sapiens* were downloaded which contain 726 genes. Genes which contain introns and exons were taken, which are 304. Pseudo-genes and predicted genes were excluded. Exon and intron sequences were downloaded and stored in separate files for each gene. Sequences of promoter (length- 200 nucleotide bases), 3'UTR (length- 800 nucleotide bases), 5' UTR (length- 200 nucleotide bases) and intergenic regions were also downloaded. Exons and introns of the same gene were fused together with 'J' in between. All exons of the same gene were stored in the same document. Similarly, all introns of the same gene were stored in same document.

Fig 6.1 Screenshot showing the *Homo sapiens* chromosome 21 sequence.



A macro was recorded and run to remove the space in the sequence. Macros were recorded in the notepad++, which was run until the end of the sequence.

Run Length Encoding compression algorithm was used to compress the sequence. RLE was used to compress homopolymeric nucleotide repeat, dinucleotide repeat and tri-nucleotide repeat. Lower limit to compress the DNA sequence for homopolymeric nucleotide repeat, dinucleotide repeat and tri-nucleotide repeat is 4, 3 and 2 respectively.

To set the upper limit of repeat units, the probabilities of occurring homopolymeric nucleotide repeat, dinucleotide repeat and tri-nucleotide repeat in a DNA sequence of 10^7 nucleotide bases were calculated using the following formulae:

For homopolymeric nucleotide repeat:

$$[P(A)]^n \times N = 1$$

$$[P(A)]^n = 1/N$$

Apply natural log on both sides:

$$n \ln P(A) = \ln(1/N)$$

$$n_A = [\ln(1/N)] / \ln P(A)$$

Where $N = 10^7$, $P(A) = 1/4$

$$n_A = (\ln 1/10^7) / (\ln 1/4)$$

$$n_A = 11.62$$

$$n_A \sim 12$$

For di-nucleotide repeat:

$$[P(AA)]^n \times N = 1$$

$$[P(AA)]^n = 1/N$$

Apply natural log on both sides:

$$n \ln P(AA) = \ln(1/N)$$

$$n_{AA} = [\ln(1/N)] / \ln P(AA)$$

Where $N = 10^7$, $P(A) = 1/16$

$$n_{AA} = (\ln 1/10^7) / (\ln 1/16)$$

$$n_{AA} = 5.81$$

$$n_{AA} \sim 6$$

For tri-nucleotide repeat:

$$[P(AAA)]^n \times N = 1$$

$$[P(AAA)]^n = 1/N$$

Apply natural log on both sides:

$$n \ln P(AAA) = \ln(1/N)$$

$$n_{AAA} = [\ln(1/N)] / \ln P(AAA)$$

Where $N = 10^7$, $P(A) = 1/64$

$$n_{AAA} = (\ln 1/10^7) / (\ln 1/64)$$

$$n_{AAA} = 3.87$$

$$n_{AAA} \sim 4$$

Where A – homopolymeric repeat, AA- dinucleotide repeat, AAA- trinucleotide repeat, P- probability

Therefore the upper limit* to compress the DNA sequence for homopolymeric nucleotide repeat, dinucleotide repeat and tri-nucleotide repeat is 12, 6 and 4 respectively.

**This is an expected length of repeats based on probabilistic model. Longer repeats may be encountered in the sequence. So the efficiency of compression may reduce in such cases. However since we are going to use same algorithm for comparing all the sequences and the sample size is reasonably large. This will mitigate the unwanted fluctuations in special cases of very long repeats in the sequence.*

Homopolymeric nucleotide repeat, dinucleotide repeat and tri-nucleotide repeat considered for compression were:

Table2. Nucleotide Repeats to be considered for compression

Homopolymeric nucleotide repeat:	GCT or CTG or TGC
As	GTA or TAG or AGT
Gs	GTC or TCG or CGT
Ts	GTT or TTG or TGT
Cs	AAC or ACA or CAA
Di- nucleotide repeat:	AAT or ATA or TAA
GA or AG	ACC or CCA or CAC
GT or TG	ACT or CTA or TAC
GC or CG	ATC or TCA or CAT
AT or TA	ATT or TTA or TAT
AC or CA	CCT or CTC or TCC
TC or CT	CTT or TTC or TCT
Tri- nucleotide repeats	
GGA or GAG or AGG	
GGC or GCG or CGG	
GGT or GTG or TGG	
GAA or AAG or GAA	
GAC or ACG or CGA	
GAT or ATG or GAT	
GCA or CAG or AGC	
GCC or CCG or CGC	

The macros were recorded separately for homopolymeric nucleotide repeat as N, for dinucleotide repeat as NN and for tri-nucleotide repeat as NNN.

Repeats were replaced in the following manner:

Table3. Replacement of nucleotide repeats using RLE

S. No.	Repeats	Replacement
1. Homopolymeric nucleotide repeat (N)		
(i)	AAAAAAAAAAAAA	[A12
(ii)	AAAAAAAAAAAAA	[A11
(iii)	AAAAAAAAAAAAA	[A10
(iv)	AAAAAAAAAAAAA	[A9
(v)	AAAAAAAAAAAAA	[A8
(vi)	AAAAAAAAAAAAA	[A7
(vii)	AAAAAAAAAAAAA	[A6
(viii)	AAAAAA	[A5
(ix)	AAAA	[A4
(i)	GGGGGGGGGGGGG	[G12
(ii)	GGGGGGGGGGGGG	[G11
(iii)	GGGGGGGGGGGGG	[G10
(iv)	GGGGGGGGGGGGG	[G9
(v)	GGGGGGGGGGGGG	[G8
(vi)	GGGGGGGGGGGGG	[G7
(vii)	GGGGGGGGGGGGG	[G6
(viii)	GGGGG	[G5
(ix)	GGGG	[G4
(i)	CCCCCCCCCCCCC	[C12
(ii)	CCCCCCCCCCCCC	[C11
(iii)	CCCCCCCCCCCCC	[C10
(iv)	CCCCCCCCCCCCC	[C9
(v)	CCCCCCCCCCCCC	[C8
(vi)	CCCCCCCCCCCCC	[C7
(vii)	CCCCCCCCCCCCC	[C6
(viii)	CCCCC	[C5
(ix)	CCCC	[C4
(i)	TTTTTTTTTTTTT	[T12
(ii)	TTTTTTTTTTTTT	[T11
(iii)	TTTTTTTTTTTTT	[T10
(iv)	TTTTTTTTTTTTT	[T9
(v)	TTTTTTTTTTTTT	[T8
(vi)	TTTTTTTTTTTTT	[T7

(vii)	TTTTTT	[T6
(viii)	TTTTT	[T5
(ix)	TTTT	[T4
2. Di- nucleotide repeats (NN)		
(i)	GAGAGAGAGAGA	[GA6
(ii)	GAGAGAGAGA	[GA5
(iii)	GAGAGAGA	[GA4
(iv)	GAGAGA	[GA3
(i)	ATATATATATAT	[AT6
(ii)	ATATATATAT	[AT5
(iii)	ATATATAT	[AT4
(iv)	ATATAT	[AT3
(i)	ACACACACACAC	[AC6
(ii)	ACACACACAC	[AC5
(iii)	ACACACAC	[AC4
(iv)	ACACAC	[AC3
(i)	GTGTGTGTGTGT	[GT6
(ii)	GTGTGTGTGT	[GT5
(iii)	GTGTGTGT	[GT4

(iv)	GTGTGT	[GT3
(i)	GCGCGCGCGCGC	[GC6
(ii)	GCGCGCGCGC	[GC5
(iii)	GCGCGCGC	[GC4
(iv)	GCGCGC	[GC3
(i)	TCTCTCTCTCTC	[TC6
(ii)	TCTCTCTCTC	[TC5
(iii)	TCTCTCTC	[TC4
(iv)	TCTCTC	[TC3
3. Tri- nucleotide repeat (NNN)		
(i)	GGAGGAGGAGGA	[GGA4
(ii)	GGAGGAGGA	[GGA3
(iii)	GGAGGA	[GGA2
(i)	GGCGGCGGCGGC	[GGC4
(ii)	GGCGGCGGC	[GGC3
(iii)	GGCGGC	[GGC2
(i)	GGTGGTGGTGGT	[GGT4
(ii)	GGTGGTGGT	[GGT3

(iii)	GGTGGT	[GGT2
(i)	GAAGAAGAAGAA	[GAA4
(ii)	GAAGAAGAA	[GAA3
(iii)	GAAGAA	[GAA2
(i)	GACGACGACGAC	[GAC4
(ii)	GACGACGAC	[GAC3
(iii)	GACGAC	[GAC2
(i)	GATGATGATGAT	[GAT4
(ii)	GATGATGAT	[GAT3
(iii)	GATGAT	[GAT2
(i)	GCAGCAGCAGCA	[GCA4
(ii)	GCAGCAGCA	[GCA3
(iii)	GCAGCA	[GCA2
(i)	GCCGCCGCCGCC	[GCC4
(ii)	GCCGCCGCC	[GCC3
(iii)	GCCGCC	[GCC2

(i)	GCTGCTGCTGCT	[GCT4
(ii)	GCTGCTGCT	[GCT3
(iii)	GCTGCT	[GCT2
(i)	GTAGTAGTAGTA	[GTA4
(ii)	GTAGTAGTA	[GTA3
(iii)	GTAGTA	[GTA2
(i)	GTCGTCGTCGTC	[GTC4
(ii)	GTCGTCGTC	[GTC3
(iii)	GTCGTC	[GTC2
(i)	GTTGTTGTTGTT	[GTT4
(ii)	GTTGTTGTT	[GTT3
(iii)	GTTGTT	[GTT2
(i)	AACAACAACAAC	[AAC4
(ii)	AACAACAAC	[AAC3
(iii)	AACAAC	[AAC2
(i)	AATAATAATAAT	[AAT4
(ii)	AATAATAAT	[AAT3

(iii)	AATAAT	[AAT2
(i)	ACCACCACCACC	[ACC4
(ii)	ACCACCACC	[ACC3
(iii)	ACCACC	[ACC2
(i)	ACTACTACTACT	[ACT4
(ii)	ACTACTACT	[ACT3
(iii)	ACTACT	[ACT2
(i)	ATCATCATCATC	[ATC4
(ii)	ATCATCATC	[ATC3
(iii)	ATCATC	[ATC2

(i)	ATTATTATTATT	[ATT4
(ii)	ATTATTATT	[ATT3
(iii)	ATTATT	[ATT2
(i)	CCTCCTCCTCCT	[CCT4
(ii)	CCTCCTCCT	[CCT3
(iii)	CCTCCT	[CCT2
(i)	CTTCTTCTTCTT	[CTT4
(ii)	CTTCTTCTT	[CTT3
(iii)	CTTCTT	[CTT2

Compression scheme can be done in the following steps:

Table4. Compression Scheme

S. No.	Steps
1	N
2	NN
3	NNN
4	N+NN
5	N+NNN
6	NN+NNN
7	NNN+NN
8	NN+N
9	NNN+N
10	N+NN+NNN
11	N+NNN+NN
12	NN+NNN+N
13	NN+N+NNN
14	NNN+NN+N
15	NNN+N+NN

Scheme with the Lowest PCR was selected. Lowest PCR means the highest compressibility. We got highest compression with NN+NNN+N i.e. dinucleotide repeat compression followed by tri-nucleotide repeat compression followed by homopolymeric nucleotide repeat compression. All the 15 above mentioned permutations of compression algorithms were used to compress the sequences. Thus 15 different compressed lengths were obtained for each sequence. The lowest compressed length was selected using MS Excel function 'Min' to calculate the PCR values.

7.1 Compression of intergenic sequences, exons and non-coding sequences of genes of human Chromosome 21

Compression of all the sequences was done for different type of sequences. We took the compression scheme where we got maximum compression. After that the PCR was calculated as follow:

$$\text{PCR} = \frac{\text{Minimum Compressed Length of Sequence}}{\text{Total Length of Sequence}} \times 100$$

Table5. PCR value of all the sequences.

GENE	EXONS (Varying)	PROMOTER (200 bp)	5'UTR	3'UTR	INTRONS	INTERGENIC
1-acylglycerol-3-phosphate O-acyltransferase 3	95.70	87.50	84.00	95.50	95.14	95.44
ADAM metalloproteinase with thrombospondin type 1 motif 1	95.82	96.50	93.00	94.75	95.41	95.18
ADAM metalloproteinase with thrombospondin type 1 motif 5	94.89	92.50	97.00	95.75	95.51	94.87
adenosine deaminase, RNA-specific, B1	96.01	95.50	89.00	95.38	95.50	95.54
alpha-crystallin A chain	96.42	96.50	96.00	94.50	96.17	
amyloid beta precursor protein	95.52	95.00	99.00	93.38	95.20	95.25
apoptosis associated transcript in bladder cancer	95.59	94.00	96.00	95.75	95.09	95.40
ATP binding cassette subfamily G member 1	95.95	93.00	97.00	97.13	95.50	96.20
ATP synthase, H ⁺ transporting, mitochondrial F1 complex, O subunit	95.98	98.00	95.00	96.13	95.13	95.04
ATP synthase, H ⁺ transporting, mitochondrial Fo complex subunit F6	96.50	97.50	96.50	95.13	95.46	94.73
autoimmune regulator	95.84	96.50	98.00	94.75	94.64	91.43
B melanoma antigen family member 2	96.69	98.00	98.00	95.00	94.48	NA
B3GALT5 antisense RNA 1	96.11	97.00	95.50	94.38	95.82	NA
BACH1 intronic transcript 2	97.26	96.00	94.00	95.75	94.99	95.18
Beta-1,3-galactosyltransferase 5	97.43	97.00	94.00	94.50	95.58	95.76
beta-site APP-cleaving enzyme 2	96.08	87.50	94.00	97.38	95.64	95.16
bromodomain and WD repeat domain containing 1	95.77	95.50	89.00	94.38	94.33	96.61
BRWD1 antisense RNA 1	95.08	93.50	96.00	92.75	96.14	94.92
BTB and CNC homology 1, basic leucine zipper	95.72	97.00	90.50	94.00	95.12	95.85

transcription factor 1						
BTG family member 3	95.36	93.00	93.00	94.75	99.06	NA
C2 calcium-dependent domain containing 2	95.51	98.00	96.00	92.50	95.51	95.28
C21orf62 antisense RNA 1	95.97	83.00	72.00	95.88	95.28	96.34
C21orf91 overlapping transcript 1	96.78	97.50	96.50	96.25	95.20	95.16
CAF-1 subunit B	96.68	91.50	96.00	89.25	94.59	95.98
carbonyl reductase 1	96.85	89.00	98.00	96.75	98.45	95.13
carbonyl reductase 3	95.82	97.00	94.50	94.75	95.00	NA
CBR3 antisense RNA 1	96.29	95.00	94.50	94.50	95.03	95.35
chaperonin containing TCP1, subunit 8 (theta)	96.90	96.50	100.00	94.50	94.35	96.07
chloride intracellular channel 6	94.27	94.00	98.00	94.88	95.30	95.54
CHODL antisense RNA 1	97.83	96.00	98.50	96.00	95.19	95.13
chondrolectin	95.76	98.00	95.00	94.63	94.91	93.87
chromosome 21 open reading frame 2	95.85	96.00	93.00	96.75	95.39	95.61
chromosome 21 open reading frame 33	96.76	97.50	96.00	95.75	95.95	95.07
chromosome 21 open reading frame 58	95.31	98.50	96.50	93.63	95.22	95.43
chromosome 21 open reading frame 59	96.30	89.50	95.50	96.88	95.80	95.19
chromosome 21 open reading frame 62	95.44	97.50	97.50	95.25	95.70	95.46
chromosome 21 open reading frame 91	94.55	91.00	96.50	94.75	95.01	95.34
claudin 14	96.89	93.00	97.50	96.38	95.60	95.52
COL18A1 antisense RNA 1	96.90	96.00	97.50	94.50	94.63	94.66
COL18A1 antisense RNA 2	96.01	94.50	95.00	95.88	95.44	96.05
collagen, type VI, alpha 1	96.78	96.50	87.00	95.88	95.10	95.12
collagen, type VI, alpha 2	97.11	91.50	90.50	97.75	94.92	96.63
collagen, type XVIII, alpha 1	95.04	93.00	90.00	95.75	95.00	
coxsackie virus and adenovirus receptor	94.65	96.50	94.00	93.50	94.90	95.21
coxsackie virus and adenovirus 'receptor	95.76	98.00	95.00	94.63	94.91	93.87
coxsackievirus and adenovirus receptor-like	95.97	92.50	91.50	96.63	94.76	NA
CPI-B (OTTHUMP00000109464) cystatin-B(liver thiol proteinase inhibitor) stefin B(stefin-B)	95.96	98.50	97.50	92.50	95.37	95.30
crystallin alpha A	96.68	95.50	95.50	94.25	95.11	96.92
crystallin zeta like 1	95.95	95.50	96.50	95.25	94.79	NA

cystathionine-beta-synthase	97.28	93.50	97.50	97.38	95.64	94.10
cystathionine-beta-synthase like	96.86	91.00	96.50	93.00	95.41	95.10
cysteine-tyrosine-rich 1	95.64	96.50	97.00	95.13	95.41	96.31
cysteine-tyrosine-rich antisense RNA 1	96.46	95.00	95.50	92.75	95.40	94.20
D21S2088E	96.00	98.00	95.50	94.38	94.55	94.67
DIP2A intronic transcript 1	95.56	96.50	96.00	95.88	94.24	95.29
disco interacting protein 2 homolog A	99.12	96.50	96.00	96.25	95.12	95.14
DNA (cytosine-5-)- methyltransferase 3-like	96.66	97.00	95.00	91.13	95.18	95.32
DnaJ heat shock protein family (Hsp40) member C28	95.45	94.00	93.00	93.13	95.17	94.80
dopey family member 2	97.06	97.50	96.50	93.00	95.45	95.25
Down syndrome cell adhesion molecule	95.86	97.50	97.50	95.75	95.25	95.25
Down syndrome critical region 10 (non-protein coding)	95.36	97.50	99.00	96.00	93.36	95.76
Down syndrome critical region 3	96.53	98.00	95.50	94.88	95.33	94.78
Down syndrome critical region 4	96.21	96.50	97.00	96.75	95.18	NA
Down syndrome critical region 8	96.89	97.50	94.50	95.25	95.40	95.50
Down syndrome critical region 9 (non-protein coding)	97.22	97.00	97.00	94.63	94.88	96.70
downstream neighbor of SON	93.80	95.50	96.00	96.88	93.87	NA
DSCAM antisense RNA 1	94.35	97.50	96.50	95.75	96.65	95.05
DSCAM intronic transcript 1	96.11	96.00	95.50	96.38	93.90	93.06
dual specificity tyrosine-(Y)- phosphorylation regulated kinase 1A	95.37	92.00	84.50	94.00	94.78	95.66
ES1 protein homolog, mitochondrial	96.64	96.50	96.00	95.00	95.82	NA
eva-1 homolog C (C. elegans)	95.48	98.00	95.50	93.88	95.37	95.29
family with sequence similarity 207 member A	96.41	79.00	98.50	96.13	95.64	96.91
family with sequence similarity 3 member B	96.37	97.00	97.50	94.38	95.59	94.75
formiminotransferase- cyclodeaminase	97.20	92.00	94.50	96.25	94.78	95.57
GA binding protein transcription factor alpha subunit	94.95	99.00	94.50	93.88	94.54	95.24
glutamate receptor, ionotropic, kainate 1	96.34	94.50	94.00	90.75	94.63	95.14
GRIK1 antisense RNA 1	98.28	96.00	97.00	96.00	95.83	95.11
GRIK1 antisense RNA 2	95.71	88.00	98.00	92.38	95.44	95.97
hCG_401127	96.31	95.50	93.00	97.25	95.01	95.07

hCG_401147	95.55	96.00	94.50	96.63	95.32	95.56
heat shock protein family A (Hsp70) member 13	95.76	96.50	97.50	96.00	94.04	95.04
heat shock transcription factor 2 binding protein	97.05	98.00	95.00	95.50	94.88	NA
high mobility group nucleosome binding domain 1	95.08	93.50	96.00	92.50	96.14	94.92
histone-lysine N-methyltransferase 2C	96.19	95.50	96.00	94.63	94.44	NA
homeobox protein cut-like 1	95.50	95.00	97.00	95.50	95.01	95.05
ICOS ligand	96.44	94.50	94.00	96.63	95.54	NA
IL10RB antisense RNA 1 (head to head)	97.01	94.00	93.00	94.50	96.25	NA
inducible T-cell co-stimulator ligand	97.05	98.00	90.50	96.38	95.56	95.28
integrin subunit beta 2	96.62	97.50	97.00	96.25	95.03	71.43
interferon (alpha, beta and omega) receptor 1	95.63	91.00	93.50	94.88	94.94	NA
interferon (alpha, beta and omega) receptor 2	95.72	95.50	93.50	96.63	95.16	93.16
interferon gamma receptor 2 (interferon gamma transducer 1)	94.24	95.00	89.50	95.63	95.28	97.31
interleukin 10 receptor subunit beta	95.80	98.00	92.00	95.25	95.47	92.51
ITGB2 antisense RNA 1	96.14	95.50	96.50	94.75	94.85	94.94
junctional adhesion molecule 2	95.50	95.00	93.00	94.50	95.00	95.44
junctional adhesion molecule 4	96.81	97.00	99.00	96.13	95.53	95.63
lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase)	96.59	97.00	95.00	93.38	95.70	NA
Leber congenital amaurosis 5-like	95.27	97.00	92.50	94.88	95.29	NA
leucine rich repeat containing 3	95.97	92.00	90.50	96.75	94.54	95.11
lipase, member I	96.55	96.50	93.00	93.63	94.73	94.54
listerin E3 ubiquitin protein ligase 1	96.34	96.00	95.50	96.38	94.76	94.86
long intergenic non-protein coding RNA 111	98.61	98.00	95.00	95.38	95.82	96.11
long intergenic non-protein coding RNA 112	98.09	97.50	97.00	95.25	94.54	NA
long intergenic non-protein coding RNA 113	98.53	96.50	96.50	96.25	95.29	94.79
long intergenic non-protein coding RNA 114	96.69	97.50	90.50	96.38	94.87	96.00
long intergenic non-protein coding RNA 1423	98.66	97.00	95.00	94.13	96.07	96.18
long intergenic non-protein coding RNA 1424	96.49	93.00	98.00	95.25	95.23	99.00

long intergenic non-protein coding RNA 1425	97.97	97.50	94.00	96.00	94.69	94.84
long intergenic non-protein coding RNA 1426	96.61	95.00	95.50	93.13	95.15	95.74
long intergenic non-protein coding RNA 1436	94.22	96.00	95.50	96.13	96.28	95.25
long intergenic non-protein coding RNA 1547	96.43	97.50	96.00	95.88	95.05	NA
long intergenic non-protein coding RNA 1548	94.62	96.00	96.50	94.25	95.88	95.06
long intergenic non-protein coding RNA 1549	95.96	97.00	98.00	94.63	94.32	94.69
long intergenic non-protein coding RNA 158	95.81	95.50	99.50	95.88	95.20	95.15
long intergenic non-protein coding RNA 159	97.49	96.50	95.50	96.00	95.25	95.35
long intergenic non-protein coding RNA 160	97.12	97.50	95.50	94.00	95.98	96.28
long intergenic non-protein coding RNA 161	96.78	95.50	98.00	95.25	95.24	95.35
long intergenic non-protein coding RNA 162	97.41	90.00	98.50	93.00	95.68	95.19
long intergenic non-protein coding RNA 163	96.29	97.50	99.00	95.13	94.87	95.60
long intergenic non-protein coding RNA 189	95.91	96.00	95.50	97.00	95.16	95.54
long intergenic non-protein coding RNA 307	97.88	98.00	86.00	94.75	90.00	95.11
long intergenic non-protein coding RNA 308	98.07	96.00	97.00	93.13	94.90	93.46
long intergenic non-protein coding RNA 310	97.13	99.00	96.00	95.25	96.34	95.57
long intergenic non-protein coding RNA 313	97.15	97.50	98.50	96.38	93.78	95.89
long intergenic non-protein coding RNA 314	96.71	94.00	97.50	95.25	94.86	94.86
long intergenic non-protein coding RNA 316	96.92	97.50	97.50	95.88	95.86	95.51
long intergenic non-protein coding RNA 317	96.08	94.50	97.50	98.13	94.16	95.39
long intergenic non-protein coding RNA 319	96.38	97.50	91.50	97.75	95.03	95.74
long intergenic non-protein coding RNA 320	96.09	97.00	97.00	95.63	94.83	94.91
long intergenic non-protein coding RNA 322	96.67	96.00	97.00	95.75	96.05	94.27
long intergenic non-protein coding RNA 323	95.61	78.00	96.50	96.13	96.31	95.44
long intergenic non-protein coding RNA 479	96.26	94.00	94.50	94.25	95.47	NA

long intergenic non-protein coding RNA 649	96.55	97.50	98.00	95.88	94.41	95.38
long intergenic non-protein coding RNA 945	97.17	96.00	96.00	96.50	96.34	95.09
LRRC3 antisense RNA 1 (head to head)	96.05	95.00	98.00	92.50	95.39	
MAP3K7 C-terminal like	96.73	98.00	97.00	90.00	95.09	95.72
MCM3AP antisense RNA 1	96.03	94.50	88.50	94.13	95.85	95.28
melanocortin 2 receptor accessory protein	97.48	96.00	97.50	95.25	95.45	NA
microRNA 3648-1	92.95	94.00	89.50	93.25	94.76	NA
microRNA 3687-1	92.95	98.00	89.50	93.25	94.76	NA
minichromosome maintenance complex component 3 associated protein	95.96	98.00	96.00	94.75	95.37	NA
MIR155 host gene	95.50	99.00	99.00	94.50	95.02	94.19
MIR548X host gene	97.79	91.00	93.50	96.13	94.94	94.76
mir-99a-let-7c cluster host gene	96.24	98.00	92.00	94.75	94.67	NA
MIS18 kinetochore protein A	93.92	96.00	97.50	97.13	94.53	93.18
mitochondrial ribosomal protein L39	95.34	95.50	93.00	94.25	95.05	95.45
mitochondrial ribosomal protein S6	95.42	96.00	88.00	95.38	95.41	95.35
MORC family CW-type zinc finger 3	95.76	98.00	97.50	96.13	94.62	94.47
MX dynamin-like GTPase 1	97.08	96.50	93.00	95.38	95.57	96.25
MX dynamin-like GTPase 2	96.57	95.50	97.00	95.75	95.81	94.76
N-6 adenine-specific DNA methyltransferase 1 (putative)	97.06	99.00	93.50	94.25	94.48	94.90
NADH-ubiquinone oxidoreductase subunit V3	96.08	95.50	94.00	94.25	94.94	95.26
neural cell adhesion molecule 2	95.90	97.50	94.00	95.38	94.58	95.04
nuclear receptor interacting protein 1	95.16	93.00	94.00	93.13	94.88	94.69
OTTHUMP00000067855	95.70	95.00	94.50	94.75	95.01	95.93
OTTHUMP00000109038	96.64	96.00	92.00	95.25	95.32	95.23
OTTHUMP00000109349 (polypeptide P1.B) trefoil factor 3	96.03	98.50	92.50	96.75	98.99	95.26
PAX3 and PAX7 binding protein 1	94.98	98.00	98.50	91.88	94.89	NA
PAXBPI antisense RNA 1	95.90	96.50	89.50	87.25	95.13	NA
PBX-knotted 1 homeobox 1	95.85	94.50	92.50	92.63	95.18	96.17
pericentrin	96.91	95.50	95.50	96.38	95.28	96.45
periodic tryptophan protein 2 homolog	96.90	97.00	95.00	94.38	95.99	NA
phosphatidylinositol glycan anchor biosynthesis class P	96.38	96.00	93.50	94.25	94.74	NA
phosphodiesterase 9A	96.22	87.00	88.00	92.75	95.39	95.67

phosphofructokinase, liver	96.39	95.00	92.00	94.88	95.00	NA
phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase, phosphoribosylaminoimidazole synthetase	96.80	97.00	91.50	95.75	94.66	NA
pituitary tumor-transforming 1 interacting protein	94.85	96.50	97.00	94.88	95.78	96.18
poly(rC) binding protein 3	96.14	96.50	94.50	95.63	96.12	95.88
potassium channel, inwardly rectifying subfamily J, member 15	96.33	99.00	93.50	95.50	95.72	95.42
potassium channel, inwardly rectifying subfamily J, member 6	96.08	96.50	93.00	94.88	95.60	95.39
potassium channel, voltage gated subfamily E regulatory beta subunit 1	95.98	93.50	90.50	92.50	95.41	96.17
potassium channel, voltage gated subfamily E regulatory beta subunit 2	95.64	97.50	98.50	94.38	94.41	96.75
POTE ankyrin domain family, member D	96.60	98.50	93.50	93.75	94.75	95.52
PR domain containing 15	96.47	91.00	94.50	97.00	94.62	96.40
proteasome (prosome, macropain) assembly chaperone 1	96.05	97.50	94.50	92.50	95.40	94.07
protein arginine methyltransferase 2	95.51	91.00	93.50	93.50	94.55	95.31
protein C-ets-2	95.84	97.00	98.00	95.00	95.61	95.49
protein O-fucosyltransferase 2	96.43	96.50	97.50	95.38	95.25	
Purkinje cell protein 4	95.34	90.50	96.50	93.75	95.54	95.84
putative uncharacterized protein C21orf125	96.68	97.50	97.00	97.38	95.03	NA
PWP2 periodic tryptophan protein homolog (yeast)	97.24	97.50	91.50	95.50	96.09	95.16
pyridoxal (pyridoxine, vitamin B6) kinase	96.32	94.50	94.00	96.25	95.05	95.38
radial spoke head 1 homolog (Chlamydomonas)	96.72	98.00	96.00	96.88	95.79	NA
receptor interacting serine- threonine kinase 4	97.07	94.50	97.00	96.88	95.79	94.87
regulator of calcineurin 1	96.22	96.00	96.00	92.75	95.72	95.67
ribosomal RNA processing 1	96.92	94.50	94.00	94.88	96.06	95.33
ribosomal RNA processing 1B	96.11	94.50	97.50	96.38	95.48	95.09
rippy transcriptional repressor 3	94.09	97.50	98.00	96.13	95.43	94.92
runt-related transcription factor 1	95.18	94.00	97.00	93.63	95.39	95.51

RWD domain containing 2B	95.48	95.50	94.50	91.25	94.71	94.77
S100 calcium binding protein B	96.26	90.00	97.50	95.38	95.41	95.46
salt inducible kinase 1	94.52	96.50	92.50	95.63	95.00	95.36
SAM domain, SH3 domain and nuclear localization signals 1	95.89	96.50	95.00	93.63	95.11	95.69
SAMSN1 antisense RNA 1	96.67	92.00	97.00	94.25	95.21	96.58
serine-threonine-protein kinase SIK1	96.64	96.00	91.50	96.50	95.66	96.57
SET domain containing 4	95.83	96.00	94.50	98.63	95.66	94.74
SH3 domain binding glutamate-rich protein	95.42	95.50	91.50	94.50	94.69	95.40
single-minded family bHLH transcription factor 2	95.64	95.50	92.00	94.63	95.41	91.23
small integral membrane protein 11A	94.94	98.50	96.50	95.63	96.14	95.64
small integral membrane protein 11B	93.68	98.50	97.00	94.13	96.07	NA
small ubiquitin-like modifier 3	96.51	98.00	97.50	94.25	94.74	NA
solute carrier family 19 (folate transporter), member 1	96.47	96.00	88.00	94.00	94.98	94.96
solute carrier family 37 (glucose-6-phosphate transporter), member 1	95.92	96.50	97.00	95.25	95.73	95.83
solute carrier family 5 (sodium-myo-inositol cotransporter), member 3	95.41	96.00	92.00	94.25	95.11	NA
SON DNA binding protein	96.16	91.50	94.50	95.25	94.48	93.85
spermatogenesis and centriole associated 1-like	94.89	96.00	94.00	94.38	95.13	95.83
SR-related CTD-associated factor 4	94.51	86.50	96.00	94.38	94.98	94.86
superoxide dismutase 1, soluble	98.34	93.50	92.50	95.38	99.15	93.85
synaptojanin 1	96.26	93.50	95.50	96.25	98.87	NA
T-cell lymphoma invasion and metastasis 1	96.05	94.50	95.00	95.75	95.24	NA
t-complex 10-like	96.63	98.00	98.00	95.75	95.85	95.39
tetratricopeptide repeat domain 3	95.47	95.00	94.50	94.50	94.94	95.80
thrombospondin-type laminin G domain and EAR repeats	96.07	97.00	96.00	96.25	95.52	95.68
trafficking protein particle complex 10	96.46	96.50	96.50	94.63	95.26	NA
transient receptor potential cation channel, subfamily M, member 2	97.10	92.00	97.00	94.13	95.54	95.26
transmembrane phosphatase with tensin homology	96.35	91.00	90.50	97.50	94.74	NA

transmembrane protease, serine 15	97.01	96.00	98.50	94.75	94.78	94.85
transmembrane protease, serine 2	96.60	99.00	97.00	95.63	95.44	95.31
transmembrane protease, serine 3	96.76	93.50	88.50	95.63	95.84	94.71
transmembrane protein 50B	95.43	97.00	96.50	96.00	94.98	95.09
trefoil factor 1	97.36	96.50	98.00	95.00	95.37	96.09
trefoil factor 2	96.89	94.50	87.00	96.25	95.94	96.38
TRPM2 antisense RNA	96.55	98.00	95.00	96.63	96.04	95.44
tryptophan rich basic protein	96.53	95.00	93.00	94.00	95.31	94.33
TSPEAR antisense RNA 1	97.93	98.50	94.00	95.38	95.93	NA
TSPEAR antisense RNA 2	97.61	95.50	95.00	96.13	96.38	95.52
TTC3 antisense RNA 1	97.05	95.50	96.50	96.63	94.44	96.00
U2 small nuclear RNA auxiliary factor 1	95.33	97.00	96.00	96.25	95.48	NA
U2 small nuclear RNA auxiliary factor 1-like 5	95.49	97.00	96.00	95.00	94.66	95.51
ubiquitin associated and SH3 domain containing A	96.91	97.00	96.50	95.13	95.55	96.00
ubiquitin conjugating enzyme E2G 2	95.77	95.00	97.00	91.25	95.45	NA
ubiquitin specific peptidase 16	95.97	95.00	94.00	94.75	94.74	96.41
ubiquitin specific peptidase 25	96.23	96.00	87.00	95.38	94.48	94.89
UMODL1 antisense RNA 1	97.01	97.50	92.50	94.75	95.99	95.61
uncharacterized LOC100129027	97.13	93.50	96.00	94.50	96.19	95.08
uncharacterized LOC100133286	95.98	98.50	96.00	94.25	95.06	94.25
uncharacterized LOC100505727	96.63	96.00	94.00	95.00	95.58	95.66
uncharacterized LOC100506403	94.85	98.00	97.50	96.88	95.42	95.43
uncharacterized LOC101927745	96.60	94.00	94.00	96.25	95.17	95.21
uncharacterized LOC101927843	97.24	98.00	97.50	96.13	94.75	NA
uncharacterized LOC101927869	96.66	97.00	92.50	97.13	94.74	94.78
uncharacterized LOC101927973	95.18	94.50	92.00	97.75	95.00	95.03
uncharacterized LOC101928107	96.11	94.00	95.00	93.00	95.13	95.51
uncharacterized LOC101928233	97.13	93.00	89.00	95.63	94.98	95.88
uncharacterized LOC101928255	96.04	95.50	98.50	95.50	98.33	95.33
uncharacterized LOC101928269	97.59	89.00	93.50	97.13	95.25	NA

uncharacterized LOC101928284	96.75	95.50	96.00	94.25	95.74	95.32
uncharacterized LOC101928398	95.63	94.00	96.00	96.25	95.94	95.57
uncharacterized LOC101928399	95.99	93.00	99.00	96.13	94.02	95.22
uncharacterized LOC101928435	97.06	94.50	95.00	76.38	96.74	95.83
uncharacterized LOC101928796	95.50	96.50	94.50	94.75	94.33	94.88
uncharacterized LOC101930094	98.22	95.50	97.50	85.75	96.27	95.74
uncharacterized LOC101930100	97.71	99.00	97.50	93.00	95.24	NA
uncharacterized LOC102723360	95.42	97.00	98.50	95.75	95.21	NA
uncharacterized LOC102723380	97.86	94.50	94.50	95.25	94.68	93.32
uncharacterized LOC102724188	96.83	93.50	99.50	94.25	94.40	94.70
uncharacterized LOC102724219	95.37	95.50	96.00	94.13	94.42	NA
uncharacterized LOC102724502	95.83	87.00	97.50	95.00	95.92	95.58
uncharacterized LOC102724740	97.08	91.00	93.50	97.38	95.65	95.44
uncharacterized LOC102724951	95.99	95.50	94.50	83.88	95.02	NA
uncharacterized LOC105369159	96.03	96.50	94.50	94.88	94.68	94.98
uncharacterized LOC105369301	97.02	94.00	96.50	94.38	95.51	95.67
uncharacterized LOC105369306	95.82	96.50	96.50	93.50	92.60	94.99
uncharacterized LOC105369308	99.18	98.00	93.50	95.88	95.87	94.92
uncharacterized LOC105372730	97.99	95.50	99.00	94.38	96.11	NA
uncharacterized LOC105372743	95.04	94.00	98.00	93.88	96.05	94.32
uncharacterized LOC105372746	97.33	97.50	91.50	96.25	94.20	94.57
uncharacterized LOC105372753	97.04	97.00	97.00	96.13	92.86	95.00
uncharacterized LOC105372760	97.78	97.00	96.50	95.00	95.14	95.21
uncharacterized LOC105372763	97.01	95.50	96.50	96.75	95.31	94.76
uncharacterized LOC105372776	97.36	95.50	94.50	96.00	94.95	95.17

uncharacterized LOC105372779	97.59	95.00	96.00	96.25	95.27	95.95
uncharacterized LOC105372784	95.83	97.00	98.50	96.50	95.92	95.58
uncharacterized LOC105372788	95.77	94.00	97.00	96.13	94.44	NA
uncharacterized LOC105372791	97.45	96.00	94.50	95.13	96.15	NA
uncharacterized LOC105372795	96.38	96.00	94.00	97.88	94.74	NA
uncharacterized LOC105372801	97.21	97.50	96.50	95.25	96.28	96.17
uncharacterized LOC105372804	94.19	96.50	97.00	93.50	94.09	94.82
uncharacterized LOC105372809	97.27	92.50	97.00	95.38	95.43	95.41
uncharacterized LOC105372831	97.08	96.50	93.50	95.25	96.46	NA
uncharacterized LOC105372832	98.70	96.00	97.00	92.75	94.90	96.16
uncharacterized LOC105372833	95.79	98.50	95.50	95.00	88.11	NA
uncharacterized LOC105372840	97.44	92.50	95.00	94.13	95.14	95.20
uncharacterized LOC105372842	96.41	92.50	96.50	94.13	96.24	95.10
uncharacterized LOC105377138	95.27	96.00	99.00	95.13	95.74	96.04
uncharacterized LOC105379493	97.57	96.50	95.50	94.75	94.86	NA
uncharacterized LOC105379499	96.81	88.50	95.50	95.75	94.56	NA
uncharacterized LOC284825	95.69	96.00	93.50	97.00	94.70	95.04
uncharacterized LOC339622	95.66	93.50	92.50	97.00	94.86	94.99
uncharacterized LOC400867	95.97	87.00	97.00	95.63	95.21	95.86
uncharacterized LOC441058	96.96	93.50	90.50	95.13	95.44	NA
uncharacterized LOC642852	95.95	95.00	96.00	96.38	95.23	95.38
uncharacterized protein ENSP00000383407-like	96.27	94.50	97.50	97.00	95.38	95.23
URB1 ribosome biogenesis 1 homolog (S. cerevisiae)	97.03	90.50	94.50	96.25	95.71	95.21
uromodulin like 1	97.39	94.50	97.00	96.50	95.68	95.49
v-ets avian erythroblastosis virus E26 oncogene homolog	96.44	97.00	96.50	93.13	95.23	95.42
WD repeat domain 4	96.76	95.50	96.00	95.50	95.50	95.09
WRD1700	96.72	98.00	96.00	96.88	95.86	NA
ybeY metallopeptidase (putative)	96.15	96.00	96.50	95.25	95.80	96.06
zinc finger and BTB domain containing 21	95.10	91.50	96.00	94.50	95.83	96.08
ZNF295 antisense RNA 1	95.84	94.00	95.50	92.38	95.89	95.47

7.2 Validation of lossless nature of the compression algorithm

Decompression and alignment:

After designing the algorithm for compression of DNA sequence, the decompression algorithm was designed to check where the compression is lossy or lossless. After decompression the original sequence and decompressed sequence was aligned using pairwise sequence alignment method.

Fig 7.1 Screenshot showing the Pairwise sequence Alignment results.

```
#=====  
#  
# Aligned_sequences: 2  
# 1: originalsequence  
# 2: decompressedsequence  
# Matrix: EDNAFULL  
# Gap_penalty: 16  
# Extend_penalty: 4  
#  
# Length: 48502  
# Identity: 48502/48502 (100.0%)  
# Similarity: 48502/48502 (100.0%)  
# Gaps: 0/48502 ( 0.0%)  
# Score: 242510  
#  
#=====  
  
originalsequence 1 GGCGGCGCACCTCGCGGGTTTCGCTATTTATGAAAAATTTCCGGTTTAA 50  
|||||  
decompressed 1 GGCGGCGCACCTCGCGGGTTTCGCTATTTATGAAAAATTTCCGGTTTAA 50  
  
originalsequence 51 GGCGTTTCCGTTCTTCTTCGTCATAACTTAATGTTTTATTTAAATACC 100  
|||||  
decompressed 51 GGCGTTTCCGTTCTTCTTCGTCATAACTTAATGTTTTATTTAAATACC 100  
  
originalsequence 101 CTCTGAAAAGAAAGGAAACGACAGGTGCTGAAAGCGAGGCTTTTGGCCT 150  
|||||  
decompressed 101 CTCTGAAAAGAAAGGAAACGACAGGTGCTGAAAGCGAGGCTTTTGGCCT 150  
  
originalsequence 151 CTGTCGTTTCCTTTCTCTGTTTTGTCCGTGGAATGAACAATGGAAGTCA 200  
|||||  
decompressed 151 CTGTCGTTTCCTTTCTCTGTTTTGTCCGTGGAATGAACAATGGAAGTCA 200
```

We took bacteriophage lambda genome sequence for checking the lossless compression. We got 100% identity of decompressed sequence with original sequence.

7.3 Analyzing means of PCR of introns and exons:

For analysing PCR means of exons and introns we had done two sample t- test. One tailed t-test was done to test that statistically there is no difference in the means of exons and introns.

Two sample t – test for one-tailed hypothesis:

$$H_0: \mu \geq 0$$

$$H_A: \mu < 0$$

$$\alpha = 0.05$$

Sample	N	Mean	Std. deviation	Std. Error Mean	F	p-value
Exon	301	96.2794	0.95733	0.5518	107.103	~0.0000
Intron	301	95.286	2.89891	0.16765		

$$P \text{ value} < 10^{-20}$$

The results indicate that the Exons are significantly more resistant to compression as compare to introns. This indicates that there is significant difference in the compressibility of introns and exons.

7.4 Comparing the PCR of coding sequences and non-coding sequence using ANOVA:

ANOVA test was done to compare the PCR of exons, introns, 5'UTR, 3'UTR, promoter and intergenic sequences. It was done to analyse the difference within and between the group samples.

ANOVA TEST:

$$H_0: \mu_{\text{exon}} = \mu_{\text{promoter}} = \mu_{5'\text{UTR}} = \mu_{3'\text{UTR}} = \mu_{\text{introns}} = \mu_{\text{intergenic}}$$

$$H_A: \mu_{\text{exon}} \neq \mu_{\text{promoter}} \neq \mu_{5'\text{UTR}} \neq \mu_{3'\text{UTR}} \neq \mu_{\text{introns}} \neq \mu_{\text{intergenic}}$$

PCR	SS	DF	MS	F	p-value
BETWEEN GROUPS	354.89	5	70.978	15.866	~0.000
WITHIN GROUPS	7775.119	1738	4.474		
TOTAL	8130.009	1743			

$$F = 15.866$$

$$F_{0.05(1),5,1738} = 0.000$$

Reject H_0

ANOVA was rejected. To find out where the difference lies in all sequence samples we further did Tukey's test.

Tukey's Test:

To find means those are significantly different from each other.

$$H_0: \mu_{\text{exon}} = \mu_{\text{promoter}} = \mu_{5'\text{UTR}} = \mu_{3'\text{UTR}} = \mu_{\text{introns}} = \mu_{\text{intergenic}}$$

$$H_A: \mu_{\text{exon}} \neq \mu_{\text{promoter}} \neq \mu_{5'\text{UTR}} \neq \mu_{3'\text{UTR}} \neq \mu_{\text{introns}} \neq \mu_{\text{intergenic}}$$

SAMPLE	Avg. PCR	RANK	N
3'UTR	94.95	1	301
5'UTR	94.99	2	301
INTERGENIC	95.19	3	239
INTRONS	95.26	4	301
PROMOTER	95.30	5	301
EXONS	96.28	6	301

Sample	SE
Equal Sample	0.122
Unequal Sample	0.13

comparison	d	SE	q	q 0.05, 5,1738	Conclusion
1 vs 4	1.33	0.122	10.912	4.03	Rejected
1 vs 3	1.3	0.122	10.541	4.03	Rejected
1 vs 6	1.09	0.13	8.372	4.03	Rejected
1 vs 5	1.02	0.122	8.388	4.03	Rejected
1vs 2	0.98	0.122	7.994	4.03	Rejected
2 vs 4	0.36	0.122	2.917	4.03	Accepted
2 vs 3	0.31	0.122	2.546	4.03	Accepted
2 vs 6	0.11	0.13	0.869	4.03	Accepted
2 vs 5	0.05	0.122	0.394	4.03	Accepted
5 vs 6	0.06	0.13	0.500	4.03	Accepted
5 vs 3	0.26	0.122	2.153	4.03	Accepted
5 vs 4	0.31	0.122	2.524	4.03	Accepted
6 vs 3	0.20	0.13	1.520	4.03	Accepted
6 vs 4	0.24	0.13	1.868	4.03	Accepted
3 vs 4	0.05	0.122	0.371	4.03	Accepted

Overall Conclusion:

$$\mu_{\text{exon}} \neq \mu_{\text{promoter}} = \mu_{\text{introns}} = \mu_{\text{intergenic}} = \mu_{5'\text{UTR}} = \mu_{3'\text{UTR}}$$

This result indicated that the PCR of exons i.e. coding sequences was significantly higher than all other sequences while PCR value of all other sequences is not significantly different from each other at 0.05 level of significance. This show that compression of coding sequences is significantly lower than all other type of sequences while non-coding sequence of any type do not differ significantly from each other in their compressibility.

DNA as the genetic material has the following features: (i) It can store genetic information (ii) Able to generate its replica through the process of replication (iii) It should have its own mechanism to decode the genetic information into functional molecules such as Proteins.

In prokaryotes, most of the DNA sequence is coding. Prokaryotic DNA rarely contains any intron or non-coding sequence. Lower eukaryotic DNA contains very few genes and the number of introns per gene is less. In higher eukaryotes e.g. mammals have ~2% coding sequence. Most of the genes contain introns and the number of intron per gene is more. Genes that contain introns are called split genes. Repetitive sequence is mainly found in intergenic region i.e. the region between two adjacent genes which do not contain any gene.

Coding sequences primarily genes encode for proteins. The process begins with transcription. Promoters are the sequences which are recognized by RNA Polymerase and transcription machinery resulting in transcription of coding part of gene. Promoters contain regulatory sequences upstream the coding region such as TATA element in eukaryotes. The primary transcript not only consists of introns but also certain sequences in the 5' and 3' ends of the genes in addition to exons. These sequences found at extreme ends of the primary transcript are known as 5' UTR and 3' UTR. 5'UTR sequence is present upstream the gene or coding region. It contains a sequence where ribosome binds and begins translation. 3'UTR present immediately after the stop codon; and helps in the termination of translation and in Post Translation Modifications.

In simpler organisms almost the entire DNA consists of unique sequences. Unique sequences mostly contain coding sequence and not contain repetitive sequences. In higher organisms there can be large amounts of repetitive DNA. There are two types of repetitive DNA: tandem repeats and dispersed repeats. The length of the non-repetitive DNA component tends to increase as the complexity of organisms increase. Large amount of DNA present in the plants and animals indicates the presence of repetitive DNA. Most genes are present in non-repetitive DNA. This indicates that genetic complexity is proportional to the amount of non-repetitive DNA.

Genome complexity can also be analysed using **reassociation kinetics**. An organism's DNA can be heated in solution until it melts, and then cooled to allow DNA strands to reassociate forming double-stranded DNA. This is typically done after shearing the DNA to form many fragments a few hundred bases in length. The larger and more complex an organism's genome is, the longer it will take for complementary strands to bump into one another and hybridize.

Complexity of sequence can be measured in the form of compressibility of sequence known as Kolmogorov complexity. Kolmogorov complexity is based on algorithmic information theory considering objects as individual symbol strings, whereas the measures effective measure complexity,

excess entropy, predictive information or thermodynamic depth relate objects to random variables and are ensemble based.

Nature of DNA sequence is not random because DNA sequence contains tandem repeats, ORF, Transposons, Pseudo-genes and duplicate genes. Coding Region contains ORF and Non- Coding Region contains tandem repeats, dispersed repeats, transposons and Pseudo-genes. Coding region of DNA sequence contains the information for synthesizing proteins whereas non coding region does not carry any information necessary to make proteins. There is not any fundamental difference in coding and non-coding DNA sequence except the ORF in coding region. Exons and coding regions are less complex as compare to non-coding regions i.e. introns and intergenic regions as they contain repetitive DNA sequence. Therefore the compressibility of coding is expected to be less than the non- coding region of DNA.

In humans, more than 98% portion of the DNA is non-coding, which contain introns and intergenic region. Introns and intergenic regions contain repetitive DNA sequences. Therefore these sequences can be compressed. Compression can be used to find out the relatedness of coding and non-coding sequences. The percent compression ratio of coding and non-coding regions of DNA must be different. Therefore, the compression technique can be used to test the non- randomness of different type of DNA sequence.

Compression is a tool for different genome comparison and for studying various properties of Genome. Compressibility can be used as a measure of information content, for the purpose of making inferences about sequences such as relatedness of the two sequences. For compressing the DNA sequence Lossless compression process can be used.

Run-length encoding (RLE) is a form of lossless data compression in which runs of data are stored as a single data value and count. This is most useful on data that contains many such runs. For DNA sequence, RLE can be used to compress the homopolymeric nucleotide repeat, di-nucleotide repeat, tri-nucleotide repeat and tandem repeats. It is not useful with files that don't have many runs as it could greatly increase the file size.

We designed RLE compression algorithm for homopolymeric, dinucleotide and tri-nucleotide repeat in the Notepad++. It is a type of lossless compression algorithm and to prove it we designed a decompression algorithm using Notepad++. For the validation of lossless nature of algorithm we took bacteriophage lambda genome and compress that genome sequence. After compression we run the decompression algorithm. To ensure 100% identity we aligned the decompressed sequence using Pairwise Sequence Alignment Method. As a result we got 100% identity after alignment of decompressed sequence with original sequence.

To find out the significant difference between the coding and non-coding sequence of DNA we took the sequence of Exons as coding sequence and Introns as non-coding sequence from Split genes of Chromosome 21 of *Homo sapiens*. Then we compress the exons and introns of same gene separately and calculated the PCR value for both of them. PCR value of exons is more as compare to introns. Then to compare the means we run Two-sample one-tailed t-test. We got p value of $>10^{20}$ which is extremely significant.

After getting highly significant results, we were interested in comparing the other non-coding sequences with coding sequences. Then we calculate the PCR for promoter, 3'UTR, 5'UTR and intergenic sequence. To compare means we run one way ANOVA using SPSS. After the rejection of ANOVA null hypothesis, we perform Tukey's test to find out where the inequality lies in all samples. In Tukey's test we got that the coding sequence i.e. exons are different from all non-coding sequences i.e. from introns, 3' UTR, 5' UTR, promoter and intergenic sequence.

So we have found a novel tool with the help of which we can measure the information content of DNA sequence. Low Compressibility of coding DNA when compared to non-coding DNA is just a beginning to show the application of information theory in genomics.

To find out the significant difference between the coding and non-coding sequence of DNA we took the sequence of Exons as coding sequence and Introns as non-coding sequence. After compression of exons and introns of same gene separately, we calculated the PCR value for both of them. PCR value of exons is more as compare to introns. Then to compare the means we run Two-sample one-tailed t-test. We got p value of $>10^{20}$ which is indicates that the exons are highly resistant to compressibility because of high complexity. After getting highly significant results, we were interested in comparing the other non-coding sequences with coding sequences. Then we calculate the PCR for promoter, 3'UTR, 5'UTR and intergenic sequence. To compare means we run one way ANOVA using SPSS. After the rejection of ANOVA null hypothesis, we perform Tukey's test to find out where the difference lies in all samples. In Tukey's test we got that the coding sequence i.e. exons are different from all non-coding sequences i.e. from introns, 3' UTR, 5' UTR, promoter and intergenic sequence. So we have found a novel tool with the help of which we can measure the information content of DNA sequence.

REFERENCE

- Bose, T., Mohammed, M. H., Dutta, A., & Mande, S. S. (2012). BIND—An algorithm for loss-less compression of nucleotide sequence data. *Journal of biosciences*, 37(4), 785-789.
- Cao, M. D., Dix, T. I., Allison, L., & Mears, C. (2007, March). A simple statistical algorithm for biological sequence compression. In *Data Compression Conference, 2007. DCC'07* (pp. 43-52). IEEE.
- Chen, X., Kwong, S., & Li, M. (2000, April). A compression algorithm for DNA sequences and its applications in genome comparison. In *Proceedings of the fourth annual international conference on Computational molecular biology* (p. 107). ACM.
- Cox, A. J., Bauer, M. J., Jakobi, T., & Rosone, G. (2012). Large-scale compression of genomic sequence databases with the Burrows–Wheeler transform. *Bioinformatics*, 28(11), 1415-1419.
- Emmert-Streib, F. (2010). Statistic complexity: combining Kolmogorov complexity with an ensemble approach. *PLoS One*, 5(8), e12256.
- Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H. S., ... & Soeda, E. (2000). The DNA sequence of human chromosome 21. *Nature*, 405(6784), 311-319.
- Janin, L., Schulz-Trieglaff, O., & Cox, A. J. (2014). BEETL-fastq: a searchable compressed archive for DNA reads. *Bioinformatics*, 30(19), 2796-2801.
- Kaur, M., Garg, E. U., & Sabo, T. (2015). A Review of Various Data Compression Techniques to form a New Technique for Text Data Compression. *reason*, 1(5).
- Li, P., Wang, S., Kim, J., Xiong, H., Ohno-Machado, L., & Jiang, X. (2013). DNA-COMPACT: DNA COM pression Based on a P attern-A ware C ontextual Modeling T echnique. *PloS one*, 8(11), e80377.
- Menconi, G., & Marangoni, R. (2006). A compression-based approach for coding sequences identification. I. Application to prokaryotic genomes. *Journal of Computational Biology*, 13(8), 1477-1488.
- Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biol*, 3(3), 1-10.
- Mohammed, M. H., Dutta, A., Bose, T., Chadaram, S., & Mande, S. S. (2012). DELIMINATE—a fast and efficient method for loss-less compression of genomic sequences Sequence analysis. *Bioinformatics*, 28(19), 2527-2529.
- Nalbantoglu, Ö. U., Russell, D. J., & Sayood, K. (2009). Data compression concepts and algorithms and their applications to bioinformatics. *Entropy*, 12(1), 34-52.
- Nelson, M., & Gailly, J. L. (1996). *The data compression book* (Vol. 2). New York: M&t Books.

- Rajarajeswari, P., & Apparao, A. (2011). Dnabit compress—genome compression algorithm. *Bioinformation*, 5(8), 350.
- Rajeswari, P. R., & Apparao, A. (2010). GenBit compress—algorithm for repetitive and non-repetitive DNA sequences. *J Theor Appl Inf Technol*, 11(1), 25-29.
- Primrose, S. B., & Twyman, R. (2013). *Principles of gene manipulation and genomics*. John Wiley & Sons.
- Primrose, S. B., & Twyman, R. (2013). *Principles of gene manipulation and genomics*. John Wiley & Sons.
- Afify, H., Islam, M., & Wahed, M. A. (2011). Dna lossless differential compression algorithm based on similarity of genomic sequence database. *arXiv preprint arXiv:1109.0094*.
- Rajeswari, P. R., & Apparao, A. (2010). GENBIT Compress-Algorithm for repetitive and non-repetitive DNA sequences. *International Journal of Computer Science and Information Technology*, 2, 25-29.

ORIGINALITY REPORT

8%

SIMILARITY INDEX

7%

INTERNET SOURCES

6%

PUBLICATIONS

0%

STUDENT PAPERS

PRIMARY SOURCES

1

www.mdpi.com

Internet Source

2%

2

Frank Emmert-Streib. "Statistic Complexity: Combining Kolmogorov Complexity with an Ensemble Approach", PLoS ONE, 08/26/2010

Publication

<1%

3

algoritmo.com

Internet Source

<1%

4

heliland.20m.com

Internet Source

<1%

5

Dianat, Noushin, Hélène Dubois-Pot-Schneider, Clara Steichen, Christophe Desterke, Philippe Leclerc, Aurélien Raveux, Laurent Combettes, Anne Weber, Anne Corlu, and Anne Dubart-Kupperschmitt. "Generation of functional cholangiocyte-like cells from human pluripotent stem cells and HepaRG cells", Hepatology, 2014.

Publication

<1%

6

Mohammed, M. H., A. Dutta, T. Bose, S. Chadaram, and S. S. Mande. "DELIMINATE-- a fast and efficient method for loss-less

<1%

compression of genomic sequences:
Sequence analysis", Bioinformatics, 2012.

Publication

7	bioinfoc.ch Internet Source	<1 %
8	Lecture Notes in Computer Science, 2015. Publication	<1 %
9	www.assignmenthelp.net Internet Source	<1 %
10	itworldnet.blogspot.com Internet Source	<1 %
11	www.smallbusinesslogos.co.uk Internet Source	<1 %
12	www.ijsrp.org Internet Source	<1 %
13	airccse.org Internet Source	<1 %
14	www.jatit.org Internet Source	<1 %
15	edoc.hu-berlin.de Internet Source	<1 %
16	www.tutorvista.com Internet Source	<1 %
17	article.sapub.org Internet Source	<1 %

18	www1.spms.ntu.edu.sg Internet Source	<1 %
19	cseweb.ucsd.edu Internet Source	<1 %
20	www.microarray.org Internet Source	<1 %
21	Zanders, Edward D.. "Cell surface and secreted proteins", Human Drug Targets, 2015. Publication	<1 %
22	Popitsch, N., and A. von Haeseler. "NGC: lossless and lossy compression of aligned high-throughput sequencing data", Nucleic Acids Research, 2012. Publication	<1 %
23	www.jscse.com Internet Source	<1 %
24	www.nslj-genetics.org Internet Source	<1 %
25	bioseq.cn Internet Source	<1 %
26	dsr-conferences.com Internet Source	<1 %
27	Berrabah, Wahiba, Pierrette Aumercier, Céline Gheeraert, Hélène Dehondt, Emmanuel Bouchaert, Jérémy Alexandre,	<1 %

Maheul Ploton, Claire Mazuy, Sandrine Caron, Anne Tailleux, Jérôme Eeckhoutte, Tony Lefebvre, Bart Staels, and Philippe Lefebvre. "The glucose sensing O-GlcNacylation pathway regulates the nuclear bile acid receptor FXR", Hepatology, 2013.

Publication

28

allamapparao.org

Internet Source

<1 %

29

www.biomedcentral.com

Internet Source

<1 %

30

www.scribd.com

Internet Source

<1 %

31

www.hipc.org

Internet Source

<1 %

32

Li, Pinghao, Shuang Wang, Jihoon Kim, Hongkai Xiong, Lucila Ohno-Machado, and Xiaoqian Jiang. "DNA-COMPACT: DNA COMpression Based on a Pattern-Aware Contextual Modeling Technique", PLoS ONE, 2013.

Publication

<1 %

33

biorxiv.org

Internet Source

<1 %

34

Park, Kee Hong, Junghee Jung, Jung-Hee Lee, and Yoon-Ho Hong. "Blood Transcriptome Profiling in Myasthenia Gravis Patients to Assess Disease Activity: A Pilot

<1 %

RNA-seq Study", Experimental Neurobiology, 2016.

Publication

35	svn.bigcat.unimaas.nl Internet Source	<1 %
36	www.dsfs.usra.edu Internet Source	<1 %
37	digitalcommons.trinity.edu Internet Source	<1 %
38	ja.wikipedia.org Internet Source	<1 %
39	X. Chen. "DNACompress: fast and effective DNA sequence compression", Bioinformatics, 12/01/2002 Publication	<1 %
40	nar.oxfordjournals.org Internet Source	<1 %
41	journal.frontiersin.org Internet Source	<1 %
42	www.badfigure.org Internet Source	<1 %
43	Fujita, Tomoko, Kitipong Soontrapa, Yoshiya Ito, Keiko Iwaisako, Catharina Sagita Moniaga, Masataka Asagiri, Masataka Majima, and Shuh Narumiya. "Hepatic stellate cells relay inflammation signaling from sinusoids to parenchyma in mouse	<1 %

models of immune-mediated hepatitis",
Hepatology, 2015.

Publication

44

R. Giancarlo. "Textual Data Compression In
Computational Biology: A synopsis",
Bioinformatics, 02/27/2009

Publication

<1 %

45

Melissa D. Nickell. "Genomics of mature and
immature olfactory sensory neurons", The
Journal of Comparative Neurology, 2012

Publication

<1 %

46

Paul D Bozyk. "Mesenchymal stromal cells
from neonatal tracheal aspirates
demonstrate a pattern of lung-specific gene
expression", Stem Cells and Development,
02/22/2011

Publication

<1 %

47

www.cnsgenomics.com

Internet Source

<1 %

48

www.theglobaltutors.co.uk

Internet Source

<1 %

49

Shibuya, K.. "Initial characterization of an
uromodulin-like 1 gene on human
chromosome 21q22.3", Biochemical and
Biophysical Research Communications,
20040709

Publication

<1 %

50

www.rcsb.org

Internet Source

<1 %

51

cdn.intechopen.com

Internet Source

<1 %

52

www.ijeemc.com

Internet Source

<1 %

53

www.stat.ucla.edu

Internet Source

<1 %

54

www.origene.com.cn

Internet Source

<1 %

55

ebi1.uniprot.org

Internet Source

<1 %

56

Ghaly, A.E., M.A. Kamal, and R. Cote. "Effect of Temperature on the Performance of Limestone/Sandstone Filters Treating Landfill Leachate", American Journal of Environmental Sciences, 2007.

Publication

<1 %

57

www.spssforum.com

Internet Source

<1 %

58

crunch.unibas.ch

Internet Source

<1 %

59

www.ebi.ac.uk

Internet Source

<1 %

60

courses.cs.vt.edu

Internet Source

<1 %

61	intl.pnas.org Internet Source	<1%
62	www.ijarcsse.com Internet Source	<1%
63	brilliantpublicschool.com Internet Source	<1%
64	mrc.ac.ir Internet Source	<1%
65	L M Patnaik. "Efficient Compression of non-repetitive DNA sequences using Dynamic Programming", 2006 International Conference on Advanced Computing and Communications, 12/2006 Publication	<1%

EXCLUDE QUOTES ON
EXCLUDE OFF
BIBLIOGRAPHY

EXCLUDE MATCHES < 8 WORDS

This document was created with Win2PDF available at <http://www.win2pdf.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.
This page will not be added after purchasing Win2PDF.