Genome Resources

Maj Gen (R) Suhaib Ahmed, HI (M)

The human genome comprises DNA sequences mostly contained in the nucleus. A small portion is also present in the mitochondria. The nuclear DNA is present in chromosomes. Each chromosome is a single coiled up molecule of DNA and a supporting protein backbone. The human genome comprises 23 pairs of chromosomes including 22 autosomes and one pair of sex chromosomes.

Deoxyribonucleic acid (DNA)

DNA is a double stranded molecule in which each strand is made of individual units called nucleotides. The nucleotides consist of a deoxyribose (5-carbon sugar), a nitrogen containing base attached to the sugar, and a phosphate group (Fig. 1.1). There are four different nucleotides each differing by the nitrogenous base. The bases are either purines (adenine and guanine) or pyrimidines (cytosine and thymine). The four bases are abbreviated as A, G, C and T. The deoxyribose sugar has 5 carbon atoms numbered 1', 2', 3', 4', and 5'. A hydroxyl group on the 5' and 3' carbons is attached to a phosphate group to form the DNA backbone. The 5' end of the strand is the starting point of DNA molecule while the 3' end is the finishing point. During the synthesis of a new strand nucleotides are added at the 3' end and the strand elongates from 5' to the 3' end.

The two strands of DNA run in opposite directions and wind around each other to form a right handed spiral. The strand that runs from left to right (→) is called the forward strand whereas the strand running from right to left (←) is called the reverse strand (Fig 1.2). The nitrogenous bases of the nucleotides face towards the interior of the helix. The two strands are kept together through hydrogen bonds between purines and pyrimidines. Adenine (A) forms two hydrogen bonds with thymine (T) and cytosine (C) forms three hydrogen bonds with guanine (G) on the opposite strands. The G-C bond is stronger than the A-T bond. This makes the G-C rich areas of DNA more stable than the rest. The synthesis of DNA starts with separation of the two strands followed by addition of nucleotides on each strand from 5' end to the 3' end.

DNA sequence

The nucleotide composition of a DNA molecule is called its "sequence". By convention and for convenience the DNA sequence of only the forward strand is written (Fig 1.3). The sequence is written from left to right starting from the 5' end and going towards the 3' end. Only the abbreviations of the nucleotides i.e. G, A, T and C are used. Occasionally other abbreviations are also used e.g. Y for pyrimidines, R for purines, and N for any nucleotide.

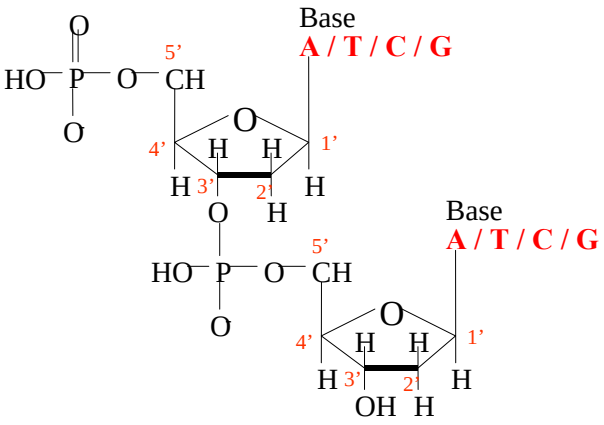

Fig 1.1. The chemical structure of DNA molecule.

Forward strand →
5′-CTGGACTTGACTGATACGTGCATTAGCATGTCCCAGTGCTAGCTAGATCAT-3′
3′-GACCTGAACTGACTATGCACGTAATCGTACAGGGTCACGATCGATCTAGTA-5′
← Reverse strand

Fig 1.2. Double stranded DNA. The forward and the reverse strands run in opposite directions.

5′-CGGAGCTGCGCGATGGCGAACTCAAGGAGCACATCAGCCGCGTCCACGCCGCC
AACTACGGTGTTTACGGTGCCCGCAAAGTGTGGCTAACCCTGAACCGTGAGGCATC
GAGGTGGCCAGATGCACCGTCGAACGGCTGATGACCAAACTCGGCCTGTCCGGGAC
CACCCGCGGCAAAGCCCGCAGGACCACGATCGCTGATCCGGCCACAGCCCGTCCCG
CCGATCTCGTCCAGCGCCGCTTCGGACCACCAGCACCTAACCGGCTGTGGGTAG-**3′**

Fig. 1.3. Mycobacterium tuberculosis insertion sequence (IS6110) (GenBank accession: AE000516.2).

Ribonucleic Acid (RNA)

RNA in a cell is seen as messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). The basic structure of all three types of RNA molecules is similar. It exists as single stranded molecule. The components of RNA are the same as of DNA with two

major differences i.e. uracil (U) replaces thymine (T) and ribose sugar replaces deoxyribose. Uracil is a pyrimidine that is structurally similar to thymine, and can base-pair with adenine.

DNA and RNA secondary structures

A three dimensional shape of a segment of single stranded DNA or RNA is often called "secondary structure". It is formed by hydrogen bonding between the adjacent nitrogenous bases. DNA is a double stranded molecule; however, during in-vitro experiments single stranded DNA is often encountered. PCR primers are a typical example. Single stranded DNA and RNA may give rise to helices, loops and hairpin structures (Fig. 1.4). DNA strands rich in G-C content are especially prone to develop secondary structures.

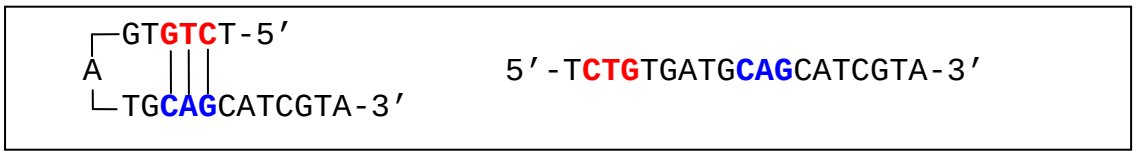

Fig 1.4 Example of hairpin secondary structure of a short sequence of DNA.

The genome resources

DNA analysis like Polymerase Chain Reaction (PCR) requires a basic knowledge of the sequence to be analyzed. The DNA sequence of interest can be obtained from a number of resources. Lack of knowledge about such resources is an important reason for underutilization of PCR in diagnostic pathology.

The most common source of genomic information is the National Centre for Biotechnology Information (NCBI) website. This may be accessed at the web address http://www.ncbi.nlm.nih.gov/

Information at the NCBI website is contained in databases including PubMed, GenBank (nucleotide and protein sequences), protein structures, complete genomes, taxonomy and others. Searching of the major databases is done through "Entrez" which is an integrated, text-based search engine and retrieval system.

GenBank® is a genetic sequence database at the National Institute of Health (NIH), Bethesda USA. It is an annotated collection of all publicly available DNA sequences. As of August 2009 there are approximately 106,533,156,756 bases in 108,431,692 sequence records in the traditional GenBank divisions and 148,165,117,763 bases in 48,443,067 sequence records in the Whole Genome Shotgun (WGS) projects division. The GenBank contains data of over 300,000 species including 1000 complete genomes of bacteria and archaea.

The information is submitted to GenBank by individual researchers and project groups all over the world. The database is increasing by about 1700 specie records every month.

About 12% of the sequences in the GenBank are of human origin. The files in the GenBank are sorted into 'divisions' such as bacteria (BCT), viruses (VRL), primates (PRI) and rodents (ROD) etc.

The complete release notes for the current version of GenBank are available at the NCBI website. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration which includes the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

When a new sequence is received the GenBank staff assigns it a unique accession number. Each sequence is also given a unique NCBI "gi" identifier number. A third identifier is the version of the sequence. The older as well as the newer versions of the sequences are kept in the database. For example complete nucleotide sequence of the human beta globin complex can be accessed by entering the accession no NG_000007.3 or the gi number 28380636 in the Entrez life sciences search engine. The version of the sequence, if any, is identified by the digit after a decimal in the accession number. In the accession number NG_000007.3, 3 indicates the version. The sequence records can be viewed in the GenBank, FASTA or Graphic format. The nucleotide sequence can be downloaded in the FASTA format which is the most convenient format for use in primer designing etc.

How to get the accession number of the sequence of interest?

The accession number of a sequence is usually obtained from the scientific publications. Since GenBank database is also linked to the scientific literature via PubMed and PubMed Central.

Basic Local Alignment Search Tool (BLAST)

BLAST is a fundamental and most frequently performed function of the GenBank. It is a family of search tools that allows comparison and finding similarities between various sequences. BLAST allows comparison of a sequence data with any of the known sequences available in the GenBank. For example the primer BLAST feature allows searching of the entire human, bacterial, viral or parasitic genomes to know whether a newly designed primer would cross anneal with a homologus sequence elsewhere or not.

Searching the human β-globin gene (GenBank NG_000007.3)

1. Go to the NCBI site (http://www.ncbi.nlm.nih.gov/)
2. Enter the accession number or simply type "beta globin gene" and press go.
3. The search will display the list of related links.
4. Clicking on the link of interest will open the page related to beta globin gene. The sequence of the gene can be visualized in FASTA or graphic format (Fig 1.5).
5. The FASTA format displays the desired sequence in a new window (Fig 1.6) from where it may be cut and pasted somewhere else.

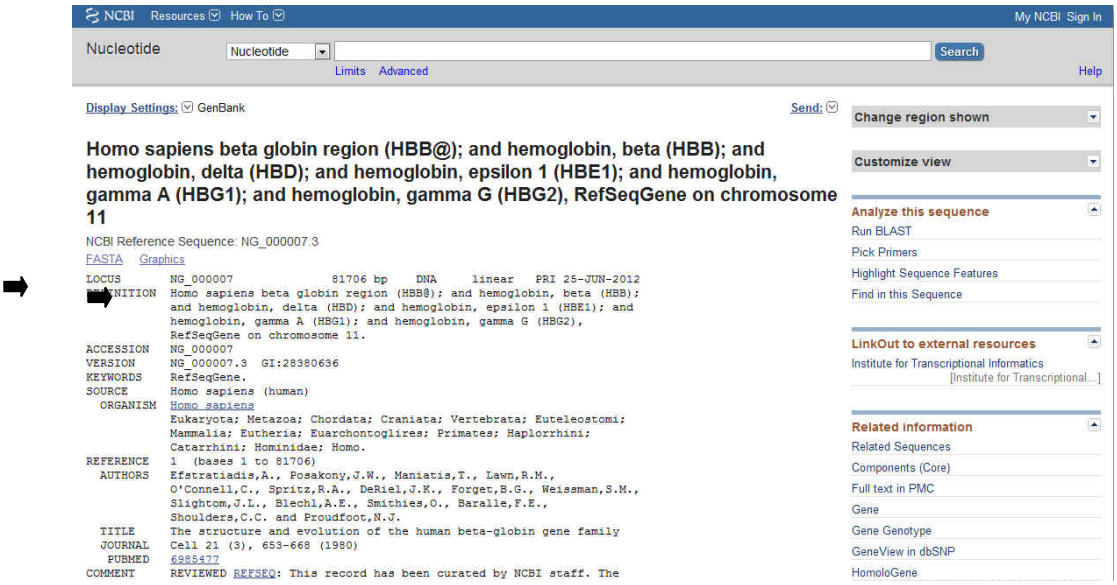

Fig. 1.5. NCBI Reference sequence of human beta globin region. The sequence record can be opened in graphic or FASTA format by clicking on the related links (arrow).

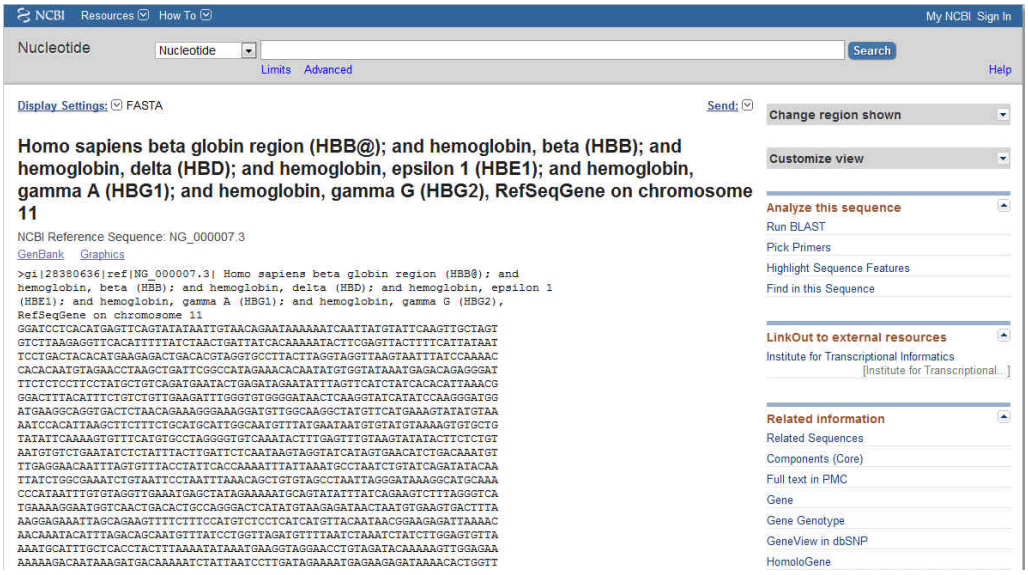

Fig. 1.6. Sequence of the human beta globin region in FASTA format. The sequence can be cut and pasted anywhere for further use.

Bibliography

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J Sayers EW (2010) GenBank. Nucleic Acids Res 38 (Database issue): 46-51.