Society For Science and Nature

www.scienceandnature.org

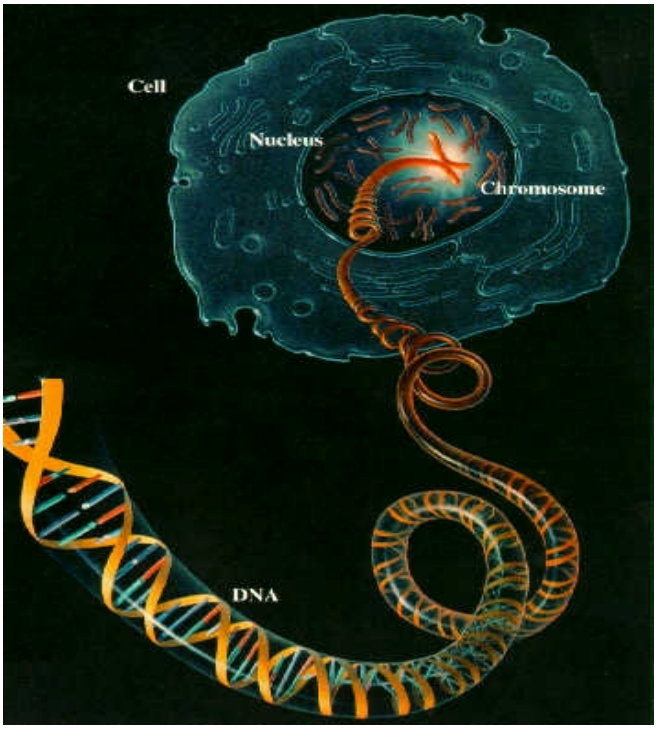
Published on September 20'2011

AMAZING FACTS ABOUT HUMAN DNA AND GENOME

Dr. Shishir Kumar Gangwar and Mr. Birhanu Worabo

Biotechnology is founded upon an ever increasing understanding of the mechanisms that maintain living organisms and allow them to reproduce from generation to generation. At the heart of life is deoxyribonucleic acid, DNA, the long, double helix molecule that carries the hereditary genetic instructions necessary to produce organisms. The genetic composition of an organism – its genotype – in conjunction with environmental influences determines its appearance and physical characteristics – its phenotype. One only has to remember how a human being changes in size, shape and behavior during a 70-80 year lifetime to know that the correlation is not simple. Deoxyribonucleic acid (DNA) and the basic protein called histone are present in the chromosomes.

The genetic instructions for a human being – its genome – are contained on DNA that is around 1.6 meters long but only one fifth of a millionth of a centimeter wide. Every cell of our bodies contains a copy of this DNA divided into 46



parts of discrete length – the chromosomes. These are so highly condensed that they can fit into the cell's nucleus which measures 3-4 millionths of a meter in diameter. Between them, human chromosomes carry some three thousand million units of chemical coding. The units are known as bases and come in four types – adenine, thymine, cytosine and guanine, or A, T, C, and G. It is the sequence of these bases in the DNA molecules which determines the biochemistry of cells and physiology of organisms.

BAC (bacterial artificial chromosome) clones seem to represent human DNA far more faithfully than their YAC or cosmid counterparts and appear to be excellent substrates for shotgun sequence analysis resulting in accurate contiguous sequence data (Principles of gene manipulation and genomics, pp 81)

Minisatellites also called variable number of tandem repeats (VNRT) have been used to carry out the first human DNA fingerprinting (Jeffreys *et al.,* 1985). It is the most commonly used fingerprint markers.

DNA also helps in regulation of gene expression by selective import of proteins into the nucleus. Proteins responsible for genome structure and organisation are all imported into the nucleus selectively. They include histones, DNA polymerases, RNA Polymerases, transcription factors and splicing factors. These proteins are targeted to the nucleus by specific amino acid sequences called nuclear localisation signals. These signals direct their transport through the nuclear pore complex. (Biotechnology part one ,pp 41)

The sequence of the human genome is 2.9×10^9 base pairs (2.9 Gbp or gigabase pairs) in length. If the sequence were typed onto paper, at about 3000 letters per page, it would fill 1 million pages of text. This extraordinary amount of information is encoded by the sequence of just four bases, cytosine, adenosine, guanine, and thymine. Most people expected the human genome sequence to reveal the actual number of genes found in human beings. In reality the massive amount of sequence needs sophisticated interpretation in order to determine how many genes it contains. The best estimates so far predict only 25,000 genes, but the number may be more or less. Of the identified genes, we only know the function of around 50%. More than 40% of the predicted human proteins are similar in structure to proteins in organisms such as fruit flies or worm (David p.clark,2009 Biotechnology applying the genetic revolution)

Using antisense to regulate gene expression is so widespread in nature that scientists became curious how many potential antisense/sense partners exist in various genomes. Computer algorithms have been devised to search for sequences that could function as antisense. In the human genome, there are a predicted 1600 different partners. The most interesting finding in the antisense field is the realization that small noncoding regulatory RNAs called

microRNAs (miRNAs) inhibit gene expression through an antisense mechanism. Using computer searches, an additional 250 potential microRNAs have been identified in humans, but because these are only about 20 nucleotides long, identifying them conclusively by computer is very difficult. (David p.clark,2009 Biotechnology applying the genetic revolution pp. 129)

The human genome contains only a few percent of coding DNA; thus, using real genes does not produce enough points on the map. A sparse map makes it difficult to order the sequences obtained in the genome sequencing project. Therefore, other markers, including physical markers, are also used on genomic maps.(David p.clark,2009 Biotechnology applying the genetic revolution pp. 232)

The entire human genome could be represented in 10000 YAC clones. In a BAC, human DNA is inserted into a plasmid in an *E. coli* cell. (A plasmid is a small piece of double-stranded DNA found in addition to the main genome, usually but not always circular.) A BAC can carry about 250 000 bp. Despite their smaller capacities, BACs are preferred to YACs because of their greater stability and ease of handling. In a YAC, human DNA is stably integrated into a small extra chromosome in a yeast cell. A YAC can contain up to 106 base pairs.(Arthur m. Lusk,2002 Introduction to bioinformatics ,pp 74)

Most of the human genome is not involved in coding for proteins.

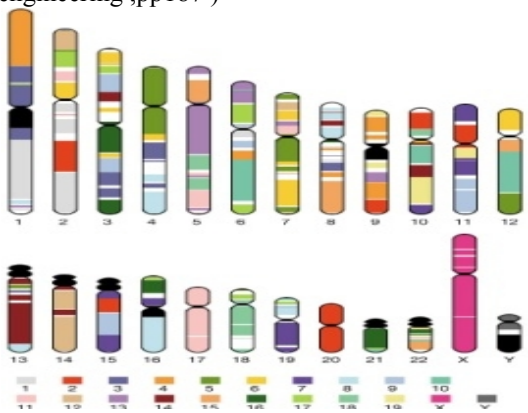
In the human genome, only about **3%** of the total amount of DNA is actually coding sequence. Even when the introns and control sequences are added, the majority of the DNA has no obvious function. This is sometimes termed 'junk' DNA, although this is perhaps the wrong way to think about this apparently redundant DNA.(Dismond T. nicoll ,2008 an introduction to genetic engineering ,pp 29)

Early analysis of human DNA indicated that the 'average' size of a coding region is around 1500 base pairs, and the average size of a gene is 10--15 kbp.Gene density is about one gene per 40--45 kbp, and the intergenic distance is around 25--30 kbp. However, gene structure in eukaryotes can be very complex, and thus using 'average' estimates is a little misleading. .(Dismond T. nicoll ,2008 an introduction to genetic engineering ,pp 29)

In the human DNA, about 40% of the total is either highly or moderately repetitive sequence DNA, which can often cause problems in the cloning and analysis of genes. Of the remaining 60%, which represents unique sequence and low-copy-number sequence elements, only around 3% constitutes the actual coding sequence. This immediately poses a problem in the analysis of the human genome, in that around 97% of the DNA could perhaps be 'avoided' if the genes themselves could be identified for further study. .(Dismond T. nicoll ,2008 an introduction to genetic engineering ,pp 179)

SOME INTERESTING FACTS ABOUT OUR GENOME

- The information would fill two hundred 500-page telephone directories.
- Between humans, our DNA differs by only 0.2%, or 1 in 500 bases (letters). (This takes into account that human cells have two copies of the genome.)
- If we recited the genome at one letter per second for 24 hours a day it would take a century to recite the book of life.
- If two different people started reciting their individual books at a rate of one letter per second, it would take nearly eight and a half minutes (500 seconds) before they reached a difference
- A typist typing at 60 words per minute (around 360 letters) for 8 hours a day would take around 50 years to type the book of life.
- Our DNA is 98% identical to that of chimpanzees.
- The vast majority of DNA in the human genome – 97% – has no known function
- The first chromosome to be completely decoded was chromosome 22 at the Sanger Centre Cambridge shire, in December 1999. There are 6 feet of DNA in each of our cells packed into a structure only 0.0004 inches across (it would easily fit on the head of a pin).
- There are 3 billion (3000000000) letters in the DNA code in every cell in your body.
- If the entire DNA in the human body was put end to end it would reach to the sun and back over 600 times (100 trillion × 6 feet divided by 93 million miles = 1200). .(Dismond T. nicoll ,2008 an introduction to genetic engineering ,pp187)



- Ninety per cent of human DNA is non-coding and the minisatellites which contain the repeating units are scattered throughout these areas. Within any particular minisatellite there is a repeating sequence of nucleotides. The sequence that is repeated is relatively in variant with in the minisatellite and consists of approximately 7-30 nucleotides
- In Human DNA, at Least 30% of the Genome Consists of Repetitive Sequences
- On average, Single nucleotide polymorphisms(SNPs) occur every 500–1,000 nucleotides in human DNA
- 10 percent of human DNA consists of sequences present in hundreds of thousands to millions of copies. Much of this highly repetitive DNA consists of SINEs, or Short Interspersed Elements (David P.clark, 2005, molecular biology understanding the genetic revolution pp.82)

- About 7% of human DNA consists of repeats of the300 bp Alu element (An example of a SINE, a particular short DNA sequence found in many copies on the chromosomes of humans and other primates.) From 300,000 to 500,000 copies (per haploid genome) of the Alu element are scattered throughout human DNA. Though apparently useless, they make up 6 to 8 percent of a human's genetic information. They occur singly or in small clusters and the majority are mutated or incomplete. The human Alu element possesses two tandem repeats of this ancestral 130 bp B1 sequence plus an extra, unrelated 31 bp insertion of obscure origin (David P.clark, 2005, molecular biology understanding the genetic revolution pp.83)
- Defects in human DNA repair systems cause assorted health problems. In particular, the higher mutation rates that occur in the absence of DNA repair cause a higher frequency of various forms of cancer.
- Human DNA may be analyzed using small blood samples or a few cells scraped from the inside of the cheek
- The probe is a segment of human DNA that may or may not be from a coding region

The sequence of the human genome still has a few gaps. These are mostly in the highly repetitive and highly condensed heterochromatin, which contains few coding sequences. The total estimated size of the human genome is 3,200 million ($3.2 X10^9$) base pairs of DNA or 3.2 Gigabase pairs (Gbp; 1 Gbp = 10^9 base pairs) of which 2.95 Gb is euchromatin. A typical page of text contains about 3,000 letters. So the human genome would fill about a *million pages*. Most DNA from higher organisms is *non-coding DNA*, including *intergenic regions*, *introns*, *repetitive sequences* and so forth. About 28% of human DNA is transcribed into RNA but, since primary transcripts include introns, only a mere *1.25%* is sequence that actually codes for proteins. On average, the introns are longer in human DNA than in other organisms sequenced so far. There are both AT-rich regions and GC-rich regions in the human genome. Curiously, the zones of GC-rich sequence have a higher density of genes and the introns are shorter. The significance of this is unknown (David P.clark, 2005, molecular biology understanding the genetic revolution pp.684)

Over half the human genome consists of repeated sequences. Some 45% is parasitic DNA (SINEs—13%; LINEs—20%; defunct retroviruses—8% and DNA-based transposons—3%). Repeats of just a few bases (microsatellites, VNTRs, etc.) account for 3% and duplications of large genome segments for 5%. Much of the genome resembles a retro-element graveyard, with only scattered outcrops of human information. The junk DNA tends to accumulate near the ends and close to the centromeres of the chromosomes. (David P.clark, 2005, molecular biology understanding the genetic revolution pp.684)

How many genes the human genome contains may seem to be a simple question; however, computer algorithms to find genes are far from perfect, especially when surveying DNA that is mostly non-coding. Although the best estimates are probably around 30,000 to 40,000 genes, analysis of the same human genome sequence has resulted in estimates of from 25,000 to 70,000 genes. Many predicted genes could be inactive pseudo genes and conversely, many genes may be overlooked especially if they consist of small exons interrupted by many long introns. Furthermore, different computer analyses may assign a particular exon sequence to different genes. Although all protein encoding genes must be transcribed, unfortunately the converse is not true. Non-gene sequences are transcribed relatively frequently and thus the presence of a transcript does not confirm the existence of a genuine gene. Determining the exact number of human genes will require very detailed analysis using a combination of laboratory and computer methods.

Are humans really more complex than other organisms? The revelation that humans only have around 25,000 genes rather than the previously estimated 100,000 upset many people. [Actually, the estimate of 100,000 human genes was little more than a guess based on inflated self-importance.] The lowly nematode worm *Caenorhabditis,* with approximately 18,000 genes, therefore has half as much genetic information as humans. The mice, and presumably most of our fellow mammals, have essentially the same number of genes as humans. Those who apparently feel that human pride depends on having more genetic information than other organisms have retreated behind the claim that humans have more gene products than other organisms. This claim is based on the observation that alternative splicing generates multiple proteins from single genes and is more frequent in higher animals. Even so, most genes do not undergo alternative splicing and there is no particular reason to believe that humans indulge in more alternative splicing than other mammals—especially the chimpanzee with which we share some 98.5% of our DNA sequence. Quibbling about which animal has most genes is in any case now moot. Sequencing has revealed that the genome of the rice plant contains 40,000 to 50,000 genes—some 10,000 more than humans. So it is the flowering plants that represent the peak of evolution, not us (David P. Clark, 2005, molecular biology understanding the genetic revolution pp.684-5)

More than 90% of the identifiable domains that make up human proteins are related to those of worms and flies. Most novel genes include previously evolved domains and thus appear to result from the re-shuffling of ancient modules.

Comparing the sequence of the human genome with other genomes originally suggested that just *over two hundred human genes were apparently borrowed from bacteria during relatively recent evolution.*

Several thousand human genes produce non coding RNA (rRNA, tRNA, snRNA, snoRNA, etc., such non-coding RNA genes lack open reading frames and are often short .There are about 500 human tRNA genes—fewer than in the worm *Caenorhabditis*! The human rRNA genes for 18S, 28S and 5.8S rRNA are found as cotranscribed units. Tandem repeats of these are found on the short arms of chromosomes 13, 14, 15, 21 and 22 giving a total of about 200 copies of these rRNA genes. The 5S rRNA gene is found separately, but also in tandem repeats, the longest

cluster being on chromosome 1, near the telomere of the long, q-arm. There are 200–300 genuine 5S rRNA genes and at least 500 pseudogenes.

Different human individuals differ by approximately one base change every 1,000–2,000 bases. This amounts to around 2.5 million SNPs over the whole genome. About 60,000 known SNPs fall within the exons of genes. Therefore, the genetic diversity in the human population is much smaller than would be expected. Despite their smaller population, chimpanzees show much more genetic diversity than humans. The most likely explanation is that after splitting off from chimps about 5 million years ago, humans went through a genetic bottleneck. Modern humans probably emerged about 100,000 years ago from a small initial population; therefore, the genetic diversity in the beginning was very low (David P.clark, 2005, molecular biology understanding the genetic revolution pp.684-90)