# ‘Introduction to Genomics’ Chapter 1 (Notes)

A human genome’s base pairs, around the size of $3.2x10^9$, are distributed among 22 paired chromosomes – XX in female and XY in male. A human is defined as follows: phenotype = genotype + environment + life history + epigenetics, where the phenotype is a collection of observable traits (hereditary); the genotype is the nuclear and mitochondrial DNA sequence (leads to pharmacogenomics – personal medicine); the environment is the living surroundings and nutrition; the life history is the integrated total of experiences and the epigenetic factors are the small amount of cells having different DNA.

A common method of generalising genetic effects from those of surroundings and experience is by implementing controlled experiments with genetically identical organisms. Human sibling and twin studies have been the basis for many important decisions however it is still difficult to measure the intelligent quotient IQ of adults due to bias (IQ remains constant in early childhood).

A human genome contains: about 2-3% protein-coding genes (transcribed into messenger RNA), about 3000 non-protein coding genes (exclusive of m-RNAs and in control of gene expression), binding sites for ligands (regulation of transcription) and 10-20% repetitive elements of unknown function.

Functionally, dispersed gene families and tandem gene family arrays have moderately repetitive DNA. This could also be said for short and long interspersed elements (SINEs and LINEs) and pseudogenes – degenerate genes which have mutated beyond functionality – in the case of repetitive elements with no known function. Highly repetitive DNA belongs to minisatellites, microsatellites and telomeres.

Note that in protein synthesis, a synonymous mutation is one which changes a codon to another codon for the same amino acid. The 20 standard amino acids in proteins are as follows: non-polar including glycine G, alanine A, proline P, valine V, isoleucine I, leucine L, phenylalanine F and methionine M; polar including serine S, cysteine C, threonine T, asparagine N, glutamine Q, tyrosine Y and tryptophan W; charged including aspartic acid D, glutamic acid E, lysine K, arginine R and histidine H.

The human genome is believed to have 23000 protein-coding genes containing exons (translatable regions) interrupted by introns (spliced out before translation), the genes appearing on both strands. A gene neighbourhood includes a set of closely linked related genes due to gene duplication (mechanism of evolution) followed by divergence.

Ideally, following a genome sequence determination, the corresponding proteome (the amino-acid sequences of the proteins expressed) can be inferred. However variety is introduced through alternative splicing and RNA editing which describe the relationship between genome sequences and proteins potentially encoded in them.

Sequences are logically one-dimensional. They must adopt three-dimensional structures to perform certain activities due to their native state, reversible denaturation and renature. The following paradigm is used: DNA sequence determines protein sequence, protein sequence determines protein structure and protein structure determines protein function. Predictions of protein structures can thus be written as programs using structure prediction methods, allowing for the creation of library structures of the encoded proteins in any genome.

Cells, included in the biosphere, are classified as prokaryotes if they: do not have a nucleus, have a size of 10μm, have no organised subcellular structure as their internal differentiation, have a fission cell division and have most of their DNA in the form of a single circular molecules with few proteins permanently attached; and as eukaryotes if they: do have a nucleus, have a size of -0.1mm, have their internal differentiation in the form of nuclei, mitochondria, chloroplasts, cytoskeleton, endoplasmic reticulum and Golgi apparatus, have a mitosis cell division and most of their DNA sequestered in the nucleus by being complexed with histones to form chromosomes. Both cell types deal with packaging and cell division (after DNA replication) problems. Plasmids are found in yeast cells (eukaryotic) – yeast artificial chromosomes YACs – which are largely used in genome sequencing.

Humans contain 46 chromosomes (22 pairs) with an additional two X chromosomes for females and an X and a Y for males. Deviations from the normal chromosome complement have unwelcome results.

Genomes are long strings of As, Ts, Gs and Cs, having functional regions such as protein-coding ones, non-protein-coding RNAs and targets of regulatory interactions. Gene identification is easier in prokaryotes as they are smaller and have fewer genes than eukaryotes which are contiguous. There are two basic approaches to identification: priori methods which recognise sequenced patterns within expressed genes and their regions; and 'been there, seen that' methods recognise regions corresponding to previously known genes. Combined approaches are possible. Characteristics to identify eukaryotic genes: the initial 5' exon, internal exons, the final 3' exon and the non-random sequences of all coding regions.

Transposable elements are skittish segments of DNA which move around the genome. Alternative mechanisms of transposition, which are displayed by different types of elements, are retro-transposons (class I) and transposons (class II). Long and short interspersed elements (LINEs and SINEs) are found in the human genome, having the most common LINE L1 appearing about 20,000 times in it as well as having around 300,000 copies of the Alu element – the most common SINE with size 280kb. The total amount of L1 + Alu is 7% of the human genome. Transposable elements' biological effects include sequence broadcasting, altering properties of genes, being an important engine of evolution, causing chromosomal rearrangements and leakage if epigenetic modification.

Genome sequencing projects of non-human species help in research about evolution and the human genome functions. If a region if conserved over the years, then it must have been conserved foo a reason. The projects also have direct application to human welfare and history. Most genome projects target individual species as a major component of public DNA repositories comes from metagenomic data (sequences derived from environmental samples without isolating individual organisms). The first genome to be sequenced was the single-stranded DNA virus, bacteriophage X-174.

High-throughput sequencing deals with the generation of raw data and assembly. Most methods sequence DNA molecules by fragmenting them and partially sequencing the pieces which have a typical length – the read length. De novo sequencing deals with the determination of the complete sequence of the first genome from a species by fragmenting them and partially sequencing them using one/single-end or both/paired-ends with the number of bases reported being the read length. Assembling the genome comes next from the sequences of overlapping fragments. A partial assembly of these into a contiguous

sequence is a contig. The fragments must cover the entire genome with enough replicates to detect errors. The data set's coverage is the ratio of the total number of bases sequenced to the genome length. After the first genome – the reference genome – is available, other genome sequences of the species can be determined with the assembly step being replaced by mapping fragments onto the reference; i.e. re-sequencing.

Exome sequencing deals with the sequencing of the protein-coding regions – the exons – making up about 1% of the entire genome. RNA sequencing deals with the conversion of RNA to complementary DNA and sequencing the results. ChIP sequencing deals with the sequencing of DNA fragments to which certain proteins are known to bind. Methylation-pattern determination deals with the comparison of native DNA sequences.

Humans' genomic sequences differ at around 0.1% of the positions (except for identical siblings). A lot of these variations come from isolated base substitutions/single-nucleotide polymorphisms (SNPs). Clusters of these are called haplotypes. Furthermore, human genome sequencing is done by a number of international organisations such as the International HapMap project and its extension, the 1000-genome project. Several companies even offer personal genome sequencing.

Genomics and proteomics have made great contributions to medicine and surgery such as: prevention of disease, detection and precise diagnosis, discovery and implementation of effective treatment and health care delivery. Analysis of a patient's genes and proteins permits selection of drugs and dosages optimal for individual patients (i.e. pharmacogenomics).

Databases archive the information generated by high-throughput sequencing and present it in a useful form. Sources of biological data include high-throughput streams such as systematic genome sequencing, protein expression patterns, metabolic pathways, protein interaction patterns and regulatory networks as well as the scientific literature/bibliographical databases. The earliest databanks were the Protein Sequence and the Protein Data banks; todays are the National Centre for Biotechnology Information NCBI and the European Bioinformatics Institute. Databanks, apart from archiving and curating, have been active in information retrieval and analysis.

A genome browser is a type of database which deals with full-genome sequences and related information. It is a project designed to organise and annotate genome information, and to present it via web pages together with links to related data. It is similar to an encyclopaedia. It also provides tools for searching and analysis. Two major genome browsers are Ensembl and Santa Cruz Genome Browser.

The divergence of sequences and structures within and between species is referred to as protein evolution. The basic tool for investigating sequence divergence is the multiple sequence alignment and the basic tool for investigating structural divergence is superposition.

DNA sequencing has its ethical, legal and social issues. Questions have arisen about privacy issues, the data that should be included and access. There are two major national repositories of human genome information in the UK, the National DNA Databank NDNAD (for law enforcement and forensics) and the UK BioBank (for medicine). In the US, there is the Combined DNA Index System and the National DNA Index System (CODIS/NDIS for forensics).