Chapter 2

Genome organization & evolution (Part II)

人類基因體 The human genome

- 蛋白質編碼基因 Protein coding gene
- 重複序列 Repeat sequences
- 核糖核酸 RNA
- Single-nucleotide polymorphisms (SNPs) and haplotypes (單核苷酸多態性和單倍型)
- http://www.azalea.org.tw/ (杜鵑花自然科學論壇)

- 2001, Feb. International Human Genome Sequencing Consortium and Celera Genomics published, separately, ***drafts*** of the human genomes
- 2003, the finishing of the genome was announced, with reduced error rate and closure of most gaps
- ~3.2×10^9 bp
  - Coding sequences from less than 5% of the human genome
  - Repeat sequence over 50%
  - The most surprising feature: small number of genes identified.
  - Finding of only about 20,000~30,000 genes

- Human genome
  - Distributed over 22 chromosome pairs plus X and Y chromosomes
  - Range from 279 Mbp down to 48 Mbp
  - X chromosome contains 163 Mbp and Y chromosome only 51 Mbp
- Exons of human protein-coding genes are relatively small
- Introns are relatively long

Protein coding genes

- IHGSC (international human genome sequencing consortium) has estimated ~32,000 genes in all.
- Top categories in a functional classification & structure revealed the most common types: http://www.ebi.ac.uk/proteome/ (pp. 88~90)

Repeat Sequences

- Repeat sequences comprise over 50% of genome
  - Transposable elements or interspersed repeats – include LINES and SINES, long terminal repeats, DNA transposon (跳躍子) fossils (化石)
  - Retroposed pseudogenes
  - Simple `stutters' – repeats of short oligomers (寡聚合物) , include minisatellites & microsatellites
  - Segmental duplications, of blocks of ~10-300 kbp.
  - Blocks of tandem (一前一後) repeats, including gene families.

RNA

- RNA genes in human genome include:
  - 497 transfer RNA genes.
  - Genes for 28S, 5.8S rRNA, & 5S rRNA
  - Small nucleolar(核仁) RNAs
  - Spliceosomal snRNAs (U1~U6).

SNPs: Single-nucleotide polymorphisms (單一核甘酸多型性)

- SNP (pronounced 'snip') is a genetic variation between individuals, limited to a single base pair – substituted, inserted, or deleted.
- Example: Sickle-cell anaemia (鐮刀狀貧血) disease caused by a specific SNP
  - An **A→T** mutation in the β–globin (血紅球蛋白) gene changes a Glu→Val, creating a sticky surface on the haemoglobin molecule that leads to *polymerization of the deoxy form* (去氧形式的聚合作用)

- SNPs are distributed through the genome, occurring on the average every 2000 bp
- Although they arose by mutation, many positions containing SNPs have low mutation rates, and provide stable markers for mapping genes.
- A high-quality, high-density human SNP map contains 1.42 million SNPs
- Not all SNPs are linked to diseases. Many are not within functional regions.

- Strong correlation of a disease with a specific SNP is advantageous in clinical work, because it is relatively easy to test for affected people or carries.
- Treatments of disease caused by defective or absent proteins include:
  - Providing normal protein
  - Life-style adjustments that make the function unnecessary.
  - Gene therapy to replace absent proteins is an active field of research.

Human SNPs database

- **SNP Fact Sheet**: http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml
- SNP consortium: http://snp.cshl.org/
- Human genome variation: https://www.gwascentral.org/

Example – three blood allelles

- A ....gctggtgacccctt...
- B ....gctcgtcaccgcta...
- O ....cgtggt-acccctt...

- A: N-乙醯半乳醣胺(N-Acetylgalactosamine)： 軟骨素中主要成分之一，與葡萄糖胺合用 可以治療退化性關節炎
- B: galactose (半乳醣)
- O: produce no active enzymes

Genetic diversity in anthropology (人類學)

- SNP data are of great utility in anthropology, given clues to historical variations in population size, and migration patterns.
- Degrees of genetic diversity are interpretable in terms of the size of founding population.
- Population-specific SNPs are informative about migrations
  - Mitochondrial(線粒體的) sequences provide information about female ancestors
  - Y chromosome sequences provide information about male ancestors

- 112/10/12

Genetic Diversity & Personal Identification

- Variation in our DNA sequences give us individual fingerprints, useful for identification & for establishment of relationships, including but not limited to question of paternity (親屬).
- Use of DNA analysis as evidence in criminal trials is now well-established.
- For most of us, all our mitochondria are genetically identical, a condition called homoplasmy (同質素).
- However, some individuals contain mitochondria with different DNA sequence; this is heteroplasmy (異值素).
  - Such sequence variation in a disease gene can complicate the observed inheritance pattern of the disease.

Evolution of Genomes

- Availability of complete information about genomic sequences has redirected research
- A general challenge in analysis of genomes is to identify “interesting events.”
- Synonymous nucleotide substitution: changes in codons that do not alter the amino acid
- Non-synonymous nucleotide substitution: changes in codons that cause mutations in the corresponding protein.

New Field of Comparative Genomics

- What ***genes*** do different phyla （門、類） share? What genes are unique to different phyla? Do the arrangements of these genes in the genome vary from phylum to phylum?
- What homologous ***proteins*** do different phyla share? What proteins are unique to different phyla? Does the integration of the activities of these protein vary from phylum to phylum? Do the mechanisms of control of expression patterns of these proteins vary from phylum to phylum?

- What ***biochemical functions*** do different phyla share?
- What biochemical functions are unique to different phyla?
- Does the integration of these biochemical functions vary from phylum to phylum?
- If two phyla share a function, and the protein that carries out this function in one phylum has a homologue in the other, does the homologous protein carry out the same function?

- Andrade, Ouzounis, Sander, Tamames, and Valencia compare proteins from 3 major domains of life:
  - *Haemophilus influenzae* (嗜血流感桿菌), 1680 genes
  - *Methanococcus jannaschii* (甲烷球菌), 1735 genes
  - *Saccharomyces cerevisiae* (烘培酵母菌), 6278 genes

- Major categories processes involve ***energy, information, and communication and regulation***:

General Function Classes

- Energy
  - Biosynthesis of cofactors (輔因子), amino acids
  - Central and intermediary metabolism (代謝)
  - Energy metabolism
  - Fatty acids and phospholipids (脂肪酸與磷脂值)
  - Nucleotide biosynthesis
  - Transport
- Information
  - Replication
  - Transcription
  - Translation
- Communication and regulation
  - Regulation functions
  - Cell envelope/cell wall
  - Cellular processes

- Analysis of shared functions among all domains of life has led people to ask whether it might be possible to define a minimal organism – an organism with the smallest gene complement consistent with independent life based on the central DNA→RNA→protein dogma.

- The smallest known independent organism is *Mycoplasma genitalium* (黴漿菌), with 468 predicted protein sequences.

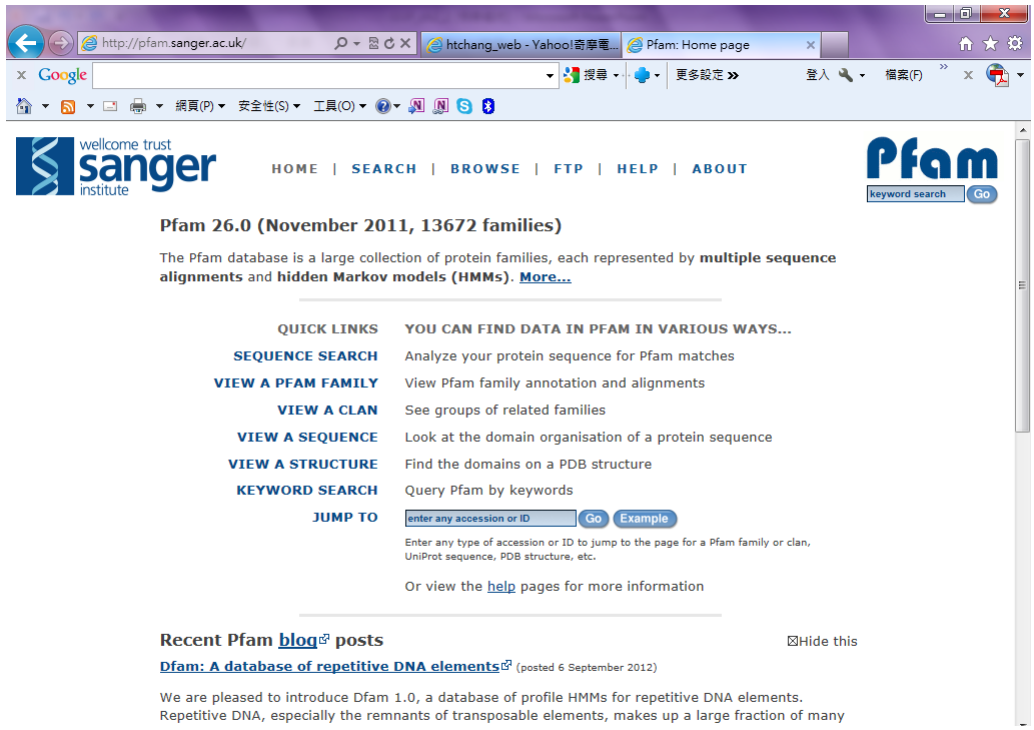
Function Classes in the Proposed Minimal Genome

- Translation, including protein synthesis
- DNA replication
- Recombination and repair – a second functions of essential proteins involved in DNA replication
- Transcription apparatus
- Chaperone (伴隨)-like proteins
- Intermediary metabolism – the glycolytic (醣解) pathway
- No nucleotide, amino acid, or fatty acid biosynthesis
- Protein-export machinery (外傳機制)
- Limited repertoire of metabolite transport proteins

Protein functional class	Number of families appearing in all known genomes
Translation, including ribosome structure	53
Transcription	4
Replication, recombination, repair	5
metabolism	9
Cellular process: (chaperones, secretion, cell division, cell wall biosynthesis)	9

Web resource box page 108: databases of aligned gene families:

- **Pfam: Protein families database**
- http://pfam.sanger.ac.uk/
- **COG: Clusters of orthologous groups**
- http://www.ncbi.nlm.nih.gov/COG/
- **HOBACGEN: Homologous Bacterial Genes Database**
- http://pbil.univ-lyon1.fr/databases/hobacgen.html
- **HOVERGEN: Homologous Vertebrate Genes Database**
- http://pbil.univ-lyon1.fr/databases/hovergen.html
- **TAED: The Adaptive Evolution Database**
- http://www.sbc.su.se/~liberles/TAED3.0/



Please pass the genes: horizontal gene transfer

- A clear example of lateral or horizontal gene transfer – a bacterium picking up a gene from the soil in which it was growing, which an organism of another species had deposited there.
- In general, horizontal gene transfer is the acquisition of genetic material by one organism from another, by natural rather than laboratory procedures, through some means other than descent from a parent during replication or mating.

- Evidence for horizontal transfer includes
  (1) discrepancies (不一致性) among evolutionary trees constructed from different genes
  (2) direct sequence comparisons between genes from different species
- Phenomenon of horizontal gene transfer is not limited to prokaryotes. Both eukaryotes and prokaryotes are chimeras (嵌合體).

- Nor is gene transfer necessarily limited to ancient ancestors. Numerous bacterial genes may have entered the human genome. Conversely, at least 8 human genes appeared in the *M. tuberculosis* (肺結核菌) genome.
- Although evidence for the importance of horizontal gene transfer is overwhelming (勢不可擋的), it was dismissed for a long time as rare and unimportant.
- The evolutionary tree as an organizing principle of biological relationship is a deeply ingrained(根深蒂固的) concept: scientists display favor in their commitment to trees, even when trees are ***not*** an appropriate model of a network of relationships.

Comparatives genomics of eukaryotes

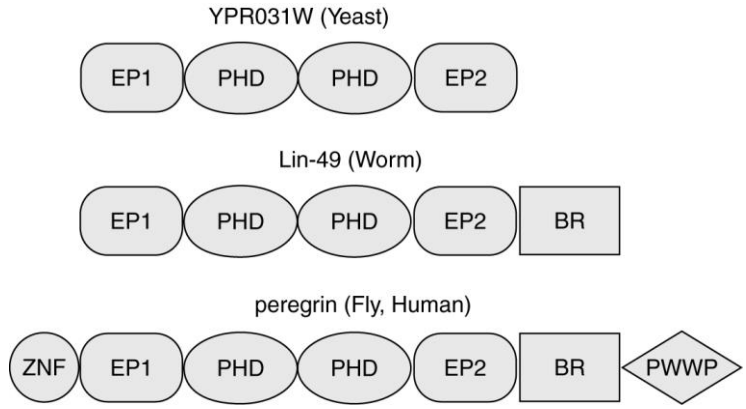
- Comparison of genomes of yeast, fly, worm, and human revealed 1308 groups of proteins that appear in all four.
- These form a conserved core of proteins for basic functions, including metabolism, DNA replication and repair, and translation.

Distribution of probable homologues of predicted human proteins

Vertebrates (脊椎動物) only	22%
Vertebrates and other animals	24%
Animals and other eukaryotes	32%
Eukaryotes and prokaryotes	21%
No homologues in animals	1%
Prokaryotes only	1%

- To create new proteins
  - Inventing new domains is an unusual event
  - Create different combinations of existing domains in increasingly complex ways is far more common
  - A common mechanism is by accretion (合生) of domains at the ends of modular proteins (Fig. 2.4)
    - This process can occur independently, and take different courses, in different phyla.

Fig. 2.4



ZNF: C_2H_2 type zinc finger
EP1, EP2 = enhancer of polycomb 1 and 2
PHD = plant homeodomain
BR = Bromo domain
PWWP = domain containing sequence motif Pro-Trp-Trp-Pro