Comparative Genomics II

1. Background

Two major questions of comparative genomics

a) What is conserved? - What are the common requirements for eukaryotic life?
b) What is different? - What makes each species unique?

2. Comparison of eukaryotic model organisms

When the Drosophila genome was completed in 2000, it was possible to look at gene conservation across three major eukaryotic model organisms, *D. melanogaster* (fly), *C. elegans* (worm), and *S. cerevisiae* (yeast). Although the human genome had not yet been completed, many genes were already known from human, mouse, and other mammals and these could also be compared.

Overall, the greatest proportion of shared genes was between mammals and Drosophila, with about 50% of the Drosophila genes giving a significant BLAST ($E < 10^{-10}$) match to mammalian genes.

About 35% of the worm genes and 37% of the yeast genes matched a mammalian gene.

b) Human disease genes in model organisms

Human genes known to be associated with disease from the OMIM (Online Mendelian Inheritance in Man) database were used as queries for BLAST searches of the fly, worm, and yeast genomes. A BLAST cut-off of $E < 10^{-6}$ was used to define significant hits.

Of 289 human disease genes:
230 (80%) were found in the fly
212 (73%) were found in the worm
120 (42%) were found in yeast

Conclusion:
Model organisms, especially Drosophila, can be very useful for studying human disease.

By 2007, the complete genomes of 12 different Drosophila species had been sequenced. Although these are all from the same genus, the sequence divergence among the species is about the same as that among mammals. About 7,000 genes had single-copy orthologs in all 12 species and almost all of these showed evidence for expression and lacked transposable element insertions. This may represent the "core" Drosophila genome. Another 5,000 genes showed homology across all species but were not single copy. That is, they were multi-gene families with multiple paralogs in different species. The number of predicted "unique" or "lineage specific" genes in the different species ranged from hundreds to thousands, but many of these lacked evidence for expression, so it is difficult to determine how many are real, functional genes.

3. Comparison of plant genomes

The first plant genome to be sequenced was that of the mustard weed, *Arabidopsis thaliana*, in 2000. This species has a small genome and is the major model system for plant genetics.

The 125 Mb Arabidopsis genome has about 25,000 genes, which is more than the fly (14,000) or the worm (19,000), and about the same as human.

Compared to fly and worm, Arabidopsis has more genes that are present as paralogs in the genome. A higher percentage of the Arabidopsis genes are part of multi-gene families. This suggests that gene duplication has played an important role in plant genome evolution.

In 2002, the genomes of two strains of rice, *Oryza sativa*, were sequenced. This allowed the first comparative genomic analysis in plants. Rice has a genome size of 430 Mb. The number of rice genes is around 40,000-50,000, depending on the method used for prediction.

80-85% of the predicted Arabidopsis genes had a homolog in rice.
Only ≈50% of the predicted rice genes had a homolog in Arabidopsis.

Plant specific genes?
Of the genes shared by Arabidopsis and rice, 30.5% had a homolog in yeast, worm, or fly.
Of the rice genes with no homolog in Arabidopsis, 2.4% had a homolog in yeast, worm, or fly.

Are Arabidopsis genes a subset of rice genes, similar to a plant minimum genome, while rice has evolved additional new genes?

4. Horizontal gene transfer
Are there bacterial genes in the human genome?
Were genes transferred directly from bacteria to humans (or other vertebrates)?

In the IHGP human genome Nature paper, they claim "hundreds of human genes appear likely to have resulted from horizontal gene transfer from bacteria at some point in the vertebrate lineage".

At the time, they identified 223 human proteins that had significant homology to bacterial proteins, but no match in yeast, worm, fly, plant, or any other (non-vertebrate) eukaryote. Possibilities:

a) Contamination?
   Unlikely – 35 genes were tested in humans by PCR and are real. Many have orthologs in other vertebrates.
b) Genes present in common ancestor of eukaryotes, but lost in yeast, worm, fly, plant, *etc*.
   Requires many independent cases of gene loss, but is possible.
c) Could be transfer from humans to bacteria. 113 of the genes are found in many diverse bacteria species, so would require many independent gene transfers.
d) The authors prefer the scenario that the genes were transferred directly from bacteria to vertebrates (at least 113 of the genes).

Since then, a number of criticisms and alternate explanations have been published:

The general finding is that when more non-vertebrate eukaryote genomes are searched, homologs to these genes can be found. This supports independent gene loss and argues against bacteria to human horizontal gene transfer. Furthermore, phylogenetic analysis of the

sequences can be used to test the hypothesis. If there was horizontal transfer, the human genes should be more similar in sequence to the bacterial genes than to other eukaryotes. This does not appear to be the case.

At present, the number of potential bacterial genes in the human genome has dropped below 40, and will likely decrease as more diverse eukaryotic genomes are sequenced.

5. Human/Mouse comparison

After human, the next vertebrate genome to be sequenced was that of the mouse. A comparison of the human and mouse genomes revealed that at least 98% of the genes had homologs between the two species. On a small scale, the order of genes was well conserved, but on a larger scale, there were many re-arrangements.

This means that mouse and humans genes are found in conserved blocks (that is, the same genes in same order = "synteny"). However, the chromosomal locations of these blocks are not well conserved. For example, gene blocks from mouse chromosome 16 are spread over 6 different human chromosomes.

These types of re-arrangements are common among vertebrate species.