EJSO 2001; **27**: 105–108
doi:10.1053/ejso.2000.1039, available online at http://www.idealibrary.com on IDEAL®

EJSO
European Journal of Surgical Oncology

EDUCATIONAL SECTION

The drosophila genome and its oncological implications

D. A. Rew

Consultant Surgeon and Honorary Senior Lecturer, Royal South Hants Cancer Centre, Southampton University Hospitals, Southampton SO14 0YG, UK

The genome of the fruit fly has recently been sequenced, prior to the release of the human genome sequence within the next few years. The fly has some 13 600 genes, compared with the estimated 80 000 genes in the human genome. Some 70% of genes appear to be broadly conserved across eukaryotic species, and some remarkable homologies have been found between 177 genes in the fly and the 289 human genes so far associated with diseases in man. The fruit fly genome is likely to prove an elegant model and a rich source of experimentation for the aetiology and regulation of human cancers.

Key words: Human Genome Project; drosophila genome; cancer.

INTRODUCTION

Drosophila melanogaster is the common fruit fly. Its relative simplicity, short life cycle and breeding capacity have made it the subject of intense genetic studies since 1910 when T. H. Morgan created the first of a huge series of genetic mutants.[1] The fly may seem a long way removed from the problems of surgical oncology, but an act of extraordinary technical, intellectual and computing virtuosity in 1999 led to the sequencing of the entire functional genome of the fly in less than a year. Initial assessments of the component genes have revealed some quite unexpected and fascinating homologies with many of the genes associated with human cancers, which are now open to new forms of experimentation and analysis.

Following the completion of the human genome project, biological science will be dominated by the interpretation and application of the massive data sets thus generated. The lessons learned from experimentation using detailed sequence data from the fruit fly genome will be important in this respect. Much of this work will be focussed on the cause and treatment of human diseases, and on cancer in particular. It is thus appropriate that as cancer specialists we have an awareness of the terminologies, principles, problems and clinical applications thrown up by genome sequencing and interpretation, as illustrated by the drosophila work.

THE CHALLENGE OF GENOME ANALYSIS

Drosophila was the third and largest major eukaryotic genome to be sequenced, alongside the yeast *Saccharomyces cerevisiae* and the worm *Caenorhabditis elegans*. Its sequencing in less than a year was a result of work to test an approach adopted by a commercial and academic consortium of Celera Genomics and the Berkeley, European and Canadian Drosophila Genome Projects. It has developed the arcane science of genome analysis on an industrial scale, with the aim of accelerating the Human Genome Project.[2]

The DNA which comprises the genome consists of huge linear sequences of adenine, cytosine, thymine and guanine (A, C, T and G) bases within a species specific number of chromosomes. The encoded genes are separated by long lengths of base sequence repeats of uncertain or structural function. To complicate matters further, the code for many genes is fragmented rather than continuous. Using standard chemical techniques, automatic sequence machines can read the genetic code at high rates and with an accuracy in excess of 99.9%. Unfortunately, this is not simply a matter of working from one end of the genome to another. The efficacy of sequencers falls off sharply after sequences of 1000 bases. Strategies must thus be devised to sequence much smaller

0748–7983/01/010105 + 04 $35.00/0

fragments of the genome and then to rebuild as a jigsaw puzzle, matching overlapping and matching sequences. Given the size of the genome, this is an exercise in massive computing power. It is further complicated by the fact that the eukaryotic genome contains very long simple base repeat sequences, fragments of which can be difficult to place. For example, the *D. melanogaster* genome occupies 180 million base pairs (Mbps), of which 125 Mbps are gene rich euchromatin spread across four chromosomes.

PRINCIPLES OF THE DROSOPHILA WHOLE GENOME SHOTGUN (WGS) SEQUENCING TECHNIQUE

WGS sequencing as used on the Drosophila Whole Genome Project is based upon the whole genome fragmentation into segments of 2 kb, 10 kb and 130 kb (one kb = one kilobase = one thousand A, T, C or G bases) using sequence specific enzymatic techniques. Each of the millions of resulting fragments is then adapted for machine reading by standard DNA cloning (replication) techniques using plasmid vectors. Plasmids are circular, double-stranded DNA molecules which live in symbiosis in some yeasts and bacteria, and which replicate independently of the host cell DNA. Sections of DNA of interest such as the fragments of the human genome can be spliced into plasmids to produce large numbers of identical copies of the fragment for machine analysis.[3]

Sufficient DNA sequencing of 500 base pair (bp) fragments is then performed to ensure that each base pair is measured numerous times, so as to permit computer aided reconstruction of the base sequence of each chromosome. This is in turn facilitated using a chromosomal framework or scaffold based upon prior knowledge about the physical structure of the chromosomes and the location of key genes within them. Greater detail on this process is provided by the Celera collaborative group.[4–6]

In contrast to the industrial scale Whole Gene Shotgun sequencing approach of the Celera group,[W1] the public Human Genome Project is a consortium of five major academic sequencing centres each working separately but interdependently in parallel on segments and chromosomes of the human genome. Fragments of the genome are used to create artificial chromosomes in bacteria for replication and analysis.[W2] Completion of a working draft was announced jointly in June 2000. The data is publicly available.

INTERPRETATION OF THE BASE SEQUENCE

Once the entire sequence of an organism is to hand, its interpretation is a major challenge.[7] The genes, or protein encoding sequences, need to be identified from the interspersed non-gene coding DNA, counted, and a putative function assigned. This process is known as annotation.[W3] This task is helped by the existence of massive libraries of gene sequences from various organisms and recorded in massive international gene data banks such as Genebank. Thirteen thousand six hundred genes are predicted within the fruit fly genome sequence. Fortunately, relatively few genes are found to be duplicated within this genome, thus simplifying research.

Comparing and matching homologous gene sequences across species and with known functions and protein products is another major computing task. Classification of the protein products of the documented genes is aided by common regions or domains of proteins which confer specific three-dimensional form and function. Thus, more than a thousand families of proteins can be identified in the fly on the basis of mosaics of identifiable domains.

The successful sequencing of a genome bypasses the need to isolate, characterize and sequence all constituent proteins separately, and allows their further study in normal and experimental mutation form through synthetic manufacture and duplication using standard molecular biology techniques. Human intelligence remains an important factor in deciphering the genome. In the case of *D. melanogaster*, this was substantially expedited by convening a 2-week analytical meeting in late 1999 of fly genome, protein and data-processing experts to interpret the raw data. This process may be a template for the interpretation of the much more complex human genome.

A higher level function not revealed by base sequences alone is the control of the genome, and the information which determines the expression of the phenotype. It is now possible to study temporal and spatial patterns of gene expression across whole cells and tissues using complex sets of gene probes and massive genetic data-processing techniques such as DNA microarrays, thus in effect freezing entire biochemical processes and protein mixtures at predetermined points, to unravel their relationships.

THE FLY GENOME, HUMAN GENES AND HUMAN DISEASE

Considerable homologies have been found between genes in fly and man involved in basic cellular processes, such as transcription factors, their regulatory targets, structural, signalling, chromosomal and transmembrane proteins.[8] Particularly surprising is the conservation of the homeobox genes which determine the orderly progress of embryogenesis in two such structurally diverse organisms.

The most unexpected finding of the fruit fly project has been the extent to which genes known to be

correlated with diseases in man have close homologies in the fly, which are thus open to precise and controlled experimentation.[8] Of the 289 disease-associated genes documented in man, there are 177 counterparts in the fly. A large number of other genes involved in replication, chromosome structure, DNA repair, transcription, translation, metabolism and solute transport also find homologies in man, thus providing a basis for the further understanding of many of the common core processes in the biology of multicellular, metazoan organisms from insect to man.

Rubin *et al.*[1] report that, on the first analysis, 68% of human cancer genes surveyed have Drosophila homologues, including those for multiple endocrine neoplasia type 1, Peutz Jeghers disease, ataxia telangiectasia, retino-blastoma, Bc12 and p53. Neurological diseases such as the muscular dystrophies; endocrine regulators such as insulin and the congenital adrenal syndromes; cardiovascular disorders and conduction defects; malformation syndromes such as Achondroplasia; are also widely represented.

Work is already in progress to assess the consequences of the expression of the equivalent human gene sequences in the fly, and strategies to counterbalance their effects. For example, the human spinocerebellar ataxia type 3 gene product also causes neurodegeneration in the fly, but its effects can be reversed by co-expression of the human HSP70 gene. The relative simplicity of the fly model, the rapid breeding cycle and large numbers of offspring allow for huge productivity increases for *in vivo* experimentation. This should be particularly useful in the precise study of the interaction of sets and regulatory networks of genes and the consequences of their precise mutations, including those which lead to neoplastic changes.

Another interesting feature of the genetics of multicellular organisms is the way in which different genes are singly or multiply expressed within the genome. Thus, there appears to be only one p53 gene in the fly, but some 700 transcription factor genes. Some 70% of the genes in all three organisms thus far studied in detail have commonality between species, an observation which seems likely to extend to man, while some 30% have no obvious commonality across species. It will prove possible with time to learn more of the evolution of gene function in more complex organisms, through duplication, multiplication and progressive mutation of 'simpler' genes in the more complex genomes of mammals. Now that the principles of whole genome sequencing are so clearly established, we may anticipate the partial and whole genome sequencing of many other index organisms, such as of small mammals, which will give us even more powerful insights into the mechanisms of human aging and disease.

We have all grown up alongside the quiet and unglamorous work of the drosophila biologists, whose meticulous labours have now brought a rich reward to clinical science. A similar exercise on the small flowering plant *Arabidopsis thaliana* has given scientists the key to plant genetics and hence the mechanisms to manipulate the production of rare and complex compounds of medicinal value.[W1] The human genome, now revealed in full through the endeavours of the public and commercial human genome projects, is much more ambitious in scale, with approximately 80 000 genes to be sequenced, annotated and understood. The literature of the next few decades will be filled with the product of these researches, and the biological databases will be massively expanded, as we approach a digital documentation and representation of the biological world. These will be fascinating times for oncologists, as the fly genome finds resonances in cancer research, and as we apply lessons from the various genome projects to the challenges of massive data sets in clinical management and research.

Key Points

1. The fruit fly is an important experimental model for human genetics.
2. The fly genome has been fully sequenced in the past year as a preamble to the commercial sequencing of the human genome.
3. About 70% of genes appear to be strongly conserved across the diverse organisms thus far sequenced, including important regulatory genes in the human cell. Thirty percent of genes appear to be unique to individual species.
4. Of the 289 human diseases for which the aetiological gene has been identified, 177 find strong homologies in the fruit fly.
5. Sixty-eight percent of known human cancer genes have drosophila homologues. The fly thus promises to be an important experimental model for human cancer genetics.

ACKNOWLEDGEMENTS

The material for this article was drawn largely from reviews and papers in the issue of *Science* of 24th March 2000, vol. 287, announcing and mapping the drosophila genome, to which further reference should be made for greater technical detail.

REFERENCES

1. Rubin GM, Lewis EB. A brief history of Drosophila's contribution to genome research. *Science* 2000; **287**: 2216–8.
2. Jasny BR, Bloom FE. Editorial: Flying to new heights. *Science* 2000; **287**: 2157.
3. Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J. *Molecular Biology of the Cell.* Third Edition. New York: Scientific American Books, WH Freeman and Co, 1995.

4. Adams MD, Celinker SE, Holt RA, *et al*. The genome sequence of Drosophila Melanogaster. *Science* 2000; **287**: 2185–95.
5. Myers EW, Sutton GG, Delcher AL, *et al*. A whole genome assembly of Drosophila. *Science* 2000; **287**: 2196–204.
6. Rubin GM, Yandell MD, Wortman JR, *et al*. A whole genome assembly of Drosophila. *Science* 2000; **287**: 21.
7. Brenner S. Perspectives: Genomics: The end of the beginning. *Science* 2000; **287**: 2173–4.
8. Kornberg TB, Krasnow MA. The drosophila genome sequence: implications for biology and medicine. *Science* 2000; **287**: 2218–20.

Website references

W1. http://www.celera.com/ The Celera corporate website providing further information on the human genome and Arabidopsis genome sequencing projects.
W2. http://www.sanger.ac.uk/HGP/ (An overview of the public Human Genome Project provided by the Sanger Centre, Cambridge, UK).
W3. http://ensembl.ebi.ac.uk/ The Ensembl database of human genomic organization provided by the European Bioinformatics Institute.