**The *Drosophila melanogaster* genome**

1. Background

Published in 2000 (*Science* 287: 2185-2195), largest genome completed at the time.
Total genome = 180 Mb, 5 chromosomes (X, 2, 3, 4, Y).
Sequenced amount =120 Mb euchromatin.
Heterochromatin (including Y chromosome) was not sequenced.
Done by Craig Venter's Company (Celera Genomics) by WGS method in collaboration with publicly funded Drosophila Genome Projects (clone based) in Europe and America

Why Drosophila? From paper …

a) Historical importance (genes are on chromosome, linearly arranged, *etc*.)
b) Large research community (≈5000 people worldwide) = high profile
c) Powerful research tools (2500 known genes, many genetic tools)
d) Modest genome size (180 Mb total)

This is a good test for WGS method on complex eukaryote, before human genome attempted

2. Genome Sequencing Strategy

see **Table 2** from Science paper

some sequence (≈29 Mb was already available from public projects

Sequencing Strategy:

1. Genomic DNA broken into 3 size classes (2 kb, 10 kb, 130 kb) and cloned into plasmids (2 kb, 10 kb) or BACs (130 kb)
2. Inserts sequenced with forward and reverse primers to get "paired reads" for over 70% of clones
3. total reads =1,903,468 from 2 kb; 1,278,386 from10 kb; 19,738 from 130 kb = 3,201,592 total reads. Ave. read length = 551 bp.
4. Tot. seq ≈ 1,700,000,000 bp; ≈ 12.5x coverage; assembled by computer
5. Closure of gaps left to public projects

Sequencing was done in 4 months using 300 ABI 3700 automated sequencers (96 capillary) and liquid handling robots and ≈ 50 people.

3. Assembly Strategy : Celera Assembler

a) Screener: mask all known repetitive sequences so they are not used in alignment.
b) Overlapper: pairwise comparison of all reads for overlaps; require at least 40 bp of overlap with < 6% mismatch in unmasked sequence.
c) Unitigger: build unique contigs of overlapping fragments supported by paired reads
d) Scaffolder: combine unitigs with paired BAC read data to make "scaffolds", where unitigs are ordered and oriented and approx. size of gap is known.
e) Repeat resolution: fill-in sequence around masked repeats. Rocks = contigs supported by at least 2 mate pairs; Stones = contigs supported by 1 mate pair plus overlap with another mate pair supported contig. Pebbles = best overlap tiling across gaps without mate pair support
f) Consensus: collapse overlaps into single sequence using highest quality sequence reads.

Processor time:
First step in assembly: Pairwise comparison of 3.2 million reads for overlaps = n (n-1) / 2 ≈ $(3.2 \times 10^6)^2 / 2 = 5 \times 10^{12}$
There are ≈ $3.2 \times 10^7$ seconds in a year. So if you did 1 pairwise comparison per second it would take > 100,000 years. The Celera supercomputer was capable of 32 million comparisons per second. Still this would require ≈ 48 hours for initial comparison. The human genome (27 million reads) would take ≈100 days. This can be shortened with parallel processing.

Celera had a custom supercomputer built by Compaq - the fastest non-military computer in the world. According to the paper, the complete Drosophila assembly required less than 1 week of processor time.

Celera made 3 assemblies (joint, 12.5x WGS, 6.5x WGS). The first include all shotgun + public data; the second only shotgun data; the third only about half of the shotgun data (as a test of assembly for 6.5x coverage, which was expected for human)

4. Annotation
*ab initio* prediction - Genscan and Genie were used to predict ORFs. Genie incorporated Drosophila specific parameters (intron signals, codon usage, EST data). Genscan was not customized for Drosophila. Thus, Genie appears to give the better prediction. Genscan predicts many more genes that are probably not real.

Experimental identification - EST sequences from BDGP (Berkeley Drosophila Genome Project) plus database information from known gene and protein sequences (≈ 2500).

Annotation "jamboree" - over 40 Drosophila researchers from around the world met at Celera for two weeks to find and classify as many genes as possible.

Total prediction: ≈ 13,600 genes, fewer then *C. elegans* (the worm, which has ≈20,000 genes)
About 8,000 of the genes (60%) were of unknown function
Gene functional classifications given in **Table 6** of paper

5. Further reading
- see *Nature Genetics* 2000, vol 24: 327-328. "Fly meets shotgun: shotgun wins"

Conclusion: Shotgun method works for eukaryote with relatively large genome and repetitive DNA, but even 12.5x coverage leaves many gaps and much "finishing" work.

Latest release of the Drosophila genome (as of 2016) = Release 6.010, includes complete euchromatic sequence of all major chromosome arms, plus about 9 Mb of heterochromatic sequence, including improved scaffolds of the Y chromosome and the "dot" chromosome (Chr. 4). Also includes improved gene annotations.
Number of protein encoding genes = 13,907.

For latest version and updates, see:
Berkeley Drosophila Genome Project (http://www.bdgp.org/)
Flybase (http://www.flybase.org/)