SPECIAL REPORT

Attack of the genomes

How many genome sequences do you need to characterize a model organism? For *Drosophila*, **Heidi Ledford** finds, a dozen is a good start.

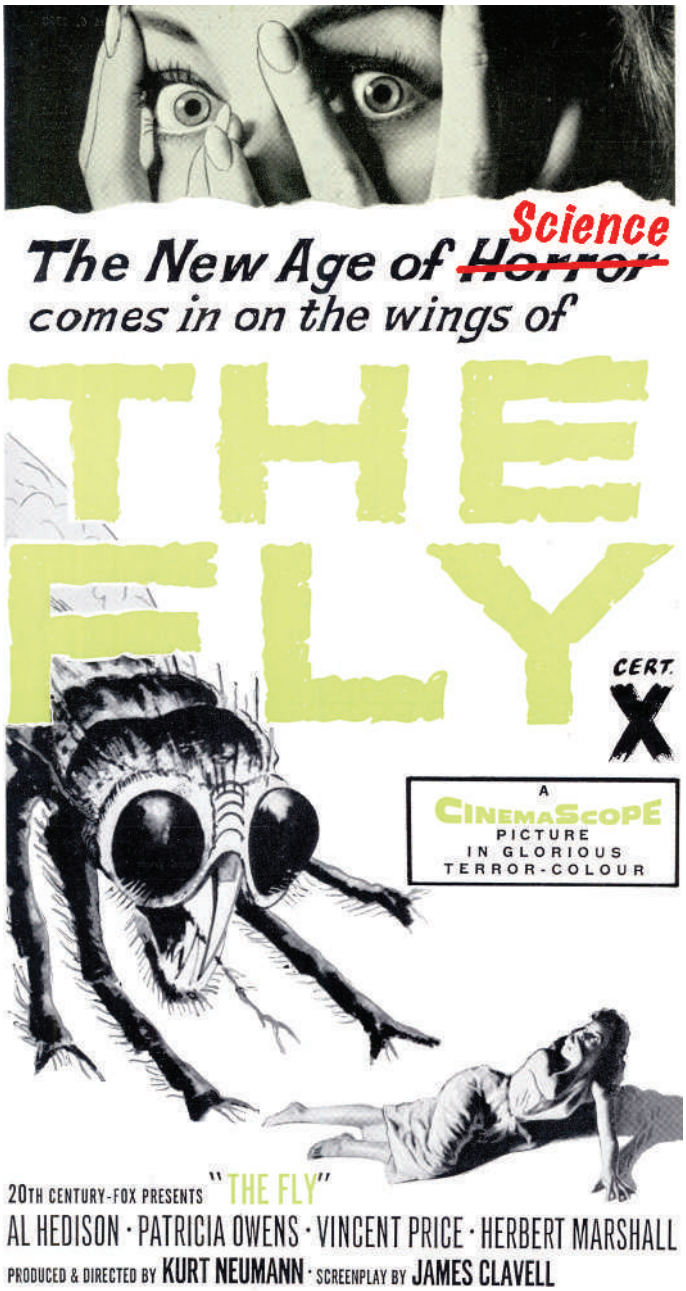
There was a time not so long ago when sequencing a single genome was cause for celebration. If that genome was from a eukaryote, so much the better. A multicellular eukaryote? Then break out the champagne.

The bar has now been raised even higher with the publication in this issue of full genome sequences from, not one, but ten fruitfly species, to add to the two sequenced previously[1,2]. Getting the genome sequence for one's favourite organism is still an achievement, but researchers are realizing that to truly understand how genomes function and evolve, they need points of comparison.

The *Drosophila* research community is not the only one benefiting from comparative genomics. More than 20 vertebrate genomes have been published or are being sequenced, and more are on the way, in projects often funded solely with the aim of using the sequences to improve understanding of the human genome.

Daniel Hartl, a geneticist at Harvard University who studies vineyard yeasts as well as *Drosophila*, says that he has lost count of how many yeast genomes have been sequenced. "I don't even keep track. It's like the Broad Institute sequences one of these before breakfast," he says, referring to the genomics centre in Cambridge, Massachusetts. The genomes of 22 species of yeast have been published, with another four on the way, bringing the number of sequenced fungal genomes to more than 60.

Given the plummeting price and escalating power of sequencing technology, researchers can now afford to be a little greedy. "At one stage even I thought: 'This is ridiculous, we're getting more and more genomes'," says Greg Elgar, a vertebrate genomicist at Queen Mary, University of London. "But it really does give us new insight into evolutionary processes."



THE RONALD GRANT ARCHIVE

By placing these vast repositories of information side by side, sequence-gazers have been able to trace the evolution of genomes. It's no sleepy drama. Chromosomes fragment and rejoin in different orientations, and entire genomes have duplicated themselves. Genes have been tossed from region to region, sometimes coming under the influence of new regulatory elements that alter the time and place of the genes' expression. These chaotic changes only begin to explain the vast phenotypic differences between related species (see 'Dew-loving all-stars'). Many of the genomic changes had been inferred from laboratory experiments and snippets of sequence, says Antonis Rokas, a geneticist at Vanderbilt University in Nashville, Tennessee, but these whole-genome studies provide the full script and confirm the story of genomic rearrangement. Nevertheless, inferring the source of the turmoil remains a challenge. "There are many different paths by which you can create a scrambled genome," says Rokas. "Identifying the most likely path is very challenging."

The immediate benefit of comparing genome sequences is the increased precision with which researchers can reveal the sequences that have been carefully preserved over time, implying that they have an important role in the organism[3]. Alternatively, these comparisons can pinpoint sequences that differ in just one species or a group of species. Subsequent lab experiments can determine whether and how those sequences — and not all of them are protein-coding genes — yielded a behavioural or morphological trait unique to that group, translating the genomes' most mysterious bits. "Sequences don't come with an index," says Hartl. "We don't really know what the sequences mean."

Even the relatively easy sequences to sort out — those that code for protein and are traditionally thought of as genes — can be challenging to nail down. The *Drosophila melanogaster* genome, sequenced in 2003 (ref. 4), nearly a century after Thomas Hunt

NATURE PODCAST
New lease of life for fruitfly genetics.
www.nature.com/nature/podcast

Morgan characterized his first white-eyed mutant, has a stellar reputation for 'annotation', as members of the community have supplemented computer-driven predictions with expert annotation of potential genes. Yet analysis of the 12 *Drosophila* genomes surprised observers by revealing hundreds of protein-coding genes in *D. melanogaster* that had been either misannotated or missed completely.

Still more challenging has been to determine which regions of the genome regulate when and where genes are expressed — giving the stage directions for building a living animal. Little is understood about how these regions work and evolve. When the mouse genome sequence was completed[5], there were hopes that human regulatory sequences could be fished out by looking for all the 'conserved' non-coding sequences — that is, those that have remained unchanged over thousands of years of evolution. "One of the big surprises in the genomics community was just how hard that was," says Andrew Clark, a geneticist at Cornell University in Ithaca, New York.

"Sequences don't come with an index. We don't really know what they mean."

The proteins that drive gene expression typically bind to small sequence motifs that can be as short as ten nucleotides. Such motifs will appear by chance many times in the genome, making it difficult to sort out those that are conserved for a purpose, especially as some regulatory elements are located thousands of nucleotides away from the genes they regulate. Regulation of gene expression can also be affected by a gene's location in the genome, and by chemical modifications to the DNA that surrounds it.

"When the genomes started coming out, a lot of people thought they could track the regulatory code just by comparing sequences," says Nicolas Gompel, a developmental biologist at the Institute of Developmental Biology at Marseille-Luminy, France. "That would have been really nice, but unfortunately it doesn't work," he says. "You do find patterns, but they're not necessarily relevant." Large-scale projects such as ENCODE, which is designed to use comparative genomics to hunt for human regulatory sequences, and modENCODE, which aims to do the same in fruitflies and nematodes, aspire to improve these predictions.

Sometimes a little shuffling can help prediction programs to home in on regulatory sequences. When researchers compared known regulatory elements directly upstream of genes in *D. melanogaster* with the same regions in *D. pseudoobscura*, they found that short motifs of 8–10 nucleotides had been conserved, but that the order of the motifs had been jumbled. "Motifs that had been previously characterized were all scrambled," says Hartl, possibly because the tiny regions had been duplicated by chance elsewhere, leaving the original motif free to decay. Annotation programs now compare genomes and look for short sequences that have been conserved but shuffled as a hallmark of these regulatory elements.

Other researchers hope to use the *Drosophila* genomes to identify puzzling cases in which the genetic regulation of a pathway has changed, yet the output of the pathway has remained the same. Researchers have found a few isolated cases of this 'transcriptional rewiring' in yeast and *Drosophila*, leading some to speculate that it might be a trend rather than a trivial chance occurrence. Comparative sequence data are crucial for uncovering these examples, says Alexander Johnson of the University of California, San Francisco. "Most 'evo-devo' studies would miss this type of circuit change," he says, because classical studies start with the overt differences between species and work backwards.

Understanding genetic circuitry is precisely the kind of area that will benefit from *Drosophila* sequences, says Elgar, who has been studying transcriptional rewiring in vertebrates. Genomic studies require careful follow-up with wet lab experiments; *D. melanogaster* has played a starring role in experimental research for more than a century and is up to the job. "I don't work on *Drosophila* but sometimes I wish I did," Elgar says. "Now everybody's going to want to join the fly community." ■

1. *Drosophila* 12 Genomes Consortium *Nature* **450,** 203–218 (2007).
2. Stark, A. *et al. Nature* **450,** 219–232 (2007).
3. *Nature* **449,** 10–11 (2007).
4. Adams, M. D. *et al. Science* **287,** 2185–2195 (2000).
5. Mouse Genome Sequencing Consortium *Nature* **420,** 520–562 (2002).

For more on the *Drosophila* genomes see articles starting on page 183.

Dew-loving all-stars

The *Drosophila* species that have had their genomes sequenced differ quite a bit physically. Here is a small sample.

N. GOMPEL; P. O'GRADY

Drosophila melanogaster
Genome size: 117 million bases
Chromosomes: 4
Of interest because: This fly has redefined itself and genetics several times during the past century. Gene mapping was invented on the huge polytene chromosomes in its salivary glands. And *D. melanogaster* is the only *Drosophila* species that can be reliably manipulated by genetic engineering.

D. grimshawi
Genome size: 201 million bases
Chromosomes: 6
Of interest because: A giant among the drosophilids, this large fly's showy wings are useful for studies of development and mating behaviour. *D. grimshawi* is also used to study fruitfly evolution and population biology in its native home of Hawaii. The Hawaiian islands are host to about one-third of all *Drosophila* species.

D. mojavensis
Genome size: 193 million bases
Chromosomes: 4
Of interest because: *D. mojavensis* survives the harsh environment of the Sonoran desert in the southwestern United States by drinking the juice of toxic cacti. Despite the dry habitat, males of the species lose 2–3% of their body weight every time they ejaculate. They mate several times a day.

S. J. CASTREZANA; N. GOMPEL

D. pseudoobscura
Genome size: 156 million bases
Chromosomes: 4
Of interest because: The second fruitfly genome to be sequenced, *D. pseudoobscura* was a favourite of the geneticist Theodosius Dobzhansky. He studied evolution in natural populations of the fly in the 1930s, and looked at some of the chromosomal rearrangements now evident by comparing genome sequences.