

# Fowl sequence

Jeremy Schmutz and Jane Grimwood

Chickens have been an invaluable model organism for decades. Their usefulness in research, from genomics to breeding, will further increase with the sequencing of the genome of one chicken species.

The article on page 695 of this issue<sup>1</sup> describes the draft sequencing and initial analysis of the genome of *Gallus gallus* — more commonly known as the red jungle fowl, the predecessor of the domestic chicken, and a valuable experimental organism (Box 1). Describing the first avian genome to be sequenced, this paper and the two that accompany it<sup>2,3</sup> provide a valuable resource for a diverse set of scientists studying a diverse set of scientific problems. Those who will benefit include agricultural researchers attempting to breed the most productive strain by recognizing links between DNA sequences and attributes such as egg production; comparative genomicists desiring to accurately identify the functional elements of the human genome; and genome-sequence producers, who continue to debate the most effective way of sequencing a vertebrate's large genome.

The draft chicken genome sequence, as reported by the International Chicken Genome Sequencing Consortium<sup>1</sup>, has several features that distinguish it from the sequenced mammalian genomes — those of humans, mice, rats and dogs. Weighing in at about 1 billion DNA base pairs, the chicken genome is broken down into 1 pair of sex chromosomes (Z and W, with females being ZW and males ZZ) and 38 non-sex chromosomes (autosomes). The autosomes vary

greatly in size, being described as macrochromosomes (large) and microchromosomes (tiny). Microchromosomes, which range from 5 million to 20 million base pairs, are not common in mammals but are abundant in birds and some fish and reptile species. The consortium's analysis of these microchromosomes in chickens indicates that they are easily discernible from macrochromosomes at the sequence level, because of their relatively high levels of guanosine–cytosine (GC) base pairs (compared with adenine–thymine pairs) and relative lack of repetitive sequences.

Another notable difference between the chicken genome and the average mammalian genome is that the chicken sequence is about one-third the size. This is now explained in part by its markedly smaller amount of 'common repeats' (stretches of sequence that occur many times), including a reduction in the number of degraded copies of gene sequences, a simpler structure of large duplications, and fewer duplicated copies of genes overall.

So much for generalities; how does this draft sequence benefit the various 'special interests' groups mentioned above? First, it provides an initial framework for chicken breeders who want to understand how genetic variation influences traits that are important in the production of domestic

chickens, by allowing the traits to be mapped back to precise genomic locations and genes. These groups have traditionally used quantitative trait loci — an estimate of the occurrence rate of a desirable 'continuous' trait in a population — to link the genetics of a strain to that trait. Continuous traits show graded variation and are controlled by more than one gene; in humans, they include height.

With the draft sequence, however — together with the second paper in this issue<sup>2</sup> — it will be easier to link specific genetic variations with variations in physical traits. In that second paper, the International Chicken Polymorphism Map Consortium describes numerous single-base-pair differences — 2.8 million of them, in fact — between three lines of domestic chicken (broiler, layer and Silkie) and the red jungle fowl. The map they have developed should allow researchers to identify the genes, and the combinations of gene variations, that produce desirable traits in chicken breeding populations. It should also increase the odds of optimizing a particular trait in subsequent generations.

For some time now, researchers in comparative genomics who are studying the human genome have also been craving the genome sequence of a species in the chicken's rough evolu-

## Box 1 Chickens in the lab

Chickens have become one of the predominant model species for biological research, for numerous reasons.

- They are ideal for classical genetics. Chickens are easy to maintain, reproduce rapidly, have large brood sizes and have distinguishable characteristics, making it possible to track traits from parent to offspring.
- Many natural mutant variants exist. Over the past 60 years, chickens displaying extremes in heritable characteristics have been maintained in breeding populations, allowing researchers to investigate the underlying genetic mutations.

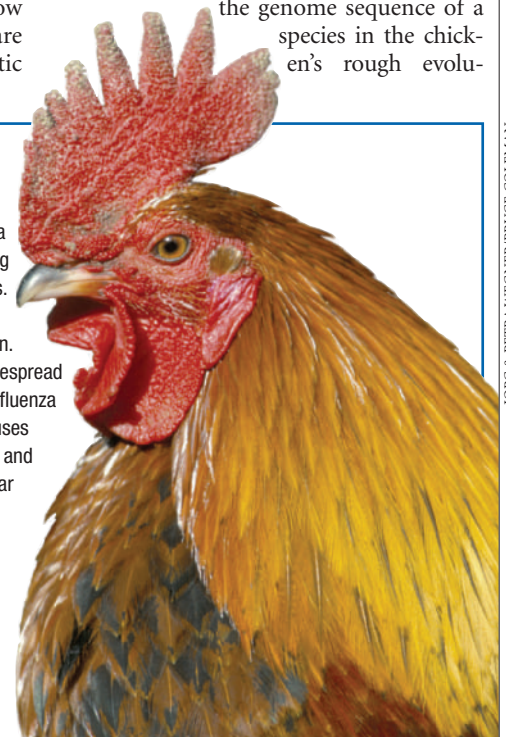
- They are ideal for studying vertebrate development. Chicken embryos develop in morphologically similar ways to mammals, yet, unlike mammalian embryos, are accessible to study (in the egg). Much of what we know about human limb formation has been uncovered through studies of chickens.

- They provide a model of human genetic diseases. Chicken lines have been created that show identical symptoms to those of patients with common debilitating diseases, including muscular dystrophy, epileptic seizures and decreased immunological responses. Studying these diseases

in chickens leads to a greater understanding of the genetic causes.

- They allow the study of viral infection. There have been widespread outbreaks of avian influenza and other animal viruses in humans. Chickens and mammals have similar immunological responses, allowing researchers to study the mechanisms of infection and the genetic basis of susceptibility.

J.S. & J.G.



JORG & PETRA WEGNER/BRUCE COLEMAN

tionary position. In general, researchers have a good grasp of how to identify those portions of a genome that are translated into proteins, by aligning sequences of messenger RNAs, the precursors of proteins, against genomic sequences of interest. One can also identify these 'coding' genomic sequences by comparing the DNA of organisms that are evolutionarily distant. For example, stretches of sequence that have been preserved in humans and fruitflies are likely to be very important for the functioning of the organisms. These sequence stretches are called conserved elements.

However, now that the human genome sequence is essentially finished<sup>4</sup>, researchers would like to do more than just identify the sequences that are translated into proteins. They also want to understand all of the regulatory structures present in a genome — structures that might, for instance, adjust the amount of protein manufactured from a particular gene. These structures are collectively known as functional elements, and the chicken, having diverged from humans more than 310 million years ago, is considered the best example so far of an 'outgroup' with which to identify them. Because enough differences between the human and chicken sequences have accumulated over this period, one can zero in on the precise base pairs that evolution has left alone for all these years — the base pairs most likely to be functional in the human genome. By comparison, the mouse, which split from humans only 75 million years ago, is too similar at the base-pair level, leading to difficulties in identifying functional elements<sup>5</sup>.

The consortium's initial analysis<sup>1</sup> describes 70 million base pairs of sequence that are highly conserved between chickens and humans. This includes base pairs within genes, but also base pairs that are between genes and therefore relate to potential functional elements (interestingly, many of these seem to be at a considerable distance from genes). Questions surrounding what these structures are and why evolution has constrained them over time will only be answered with targeted experiments, some of which are beginning to get under way<sup>6</sup>.

Finally, for those who concentrate on generating large-scale genomic sequences and resources, the chicken genome represents another in a series of grand experiments to balance two different approaches. Traditional clone-by-clone approaches (see, for example, refs 4, 7) — which involve cloning a genome into bacterial artificial chromosomes (BACs), mapping the clones, then sequencing them and assembling the sequences by using the map — are time-consuming but generally produce an accurate representation of all regions of the genome. Whole-genome shotgun<sup>8</sup> (WGS) is quicker, because it involves shattering the whole genome into pieces, sequencing the

fragments and assembling them by computer, but it often fails to represent all regions accurately.

The chicken sequence presented here is a halfway house: it is not a straight WGS assembly, but has been revised according to a physical map of 180,000 BAC clones, detailed by Wallis *et al.*<sup>3</sup> on page 761. This map was crucial in ordering and localizing the sequence pieces generated by WGS. Thus the assembly captures an impressive 98% of the sequence over most of the genome, with that number falling slightly in very GC-rich regions. The authors were also able to locate partial or complete sequences of at least 97% of coding genes that were previously known to exist.

However, the genome has received no directed 'finishing' work, and issues do still exist — there is a distinct lack of continuity in 10% of the gene-rich regions, and there are perhaps 1.4 million base pairs of sequence that are in the wrong position. Recent studies<sup>9</sup> suggest that, even with algorithmic improvements, WGS assemblies fail to resolve large-scale duplications in vertebrate genomes; even with a BAC map, recently duplicated

sequences in the chicken assembly are poorly resolved<sup>1</sup>. And the authors suggest that one reason why they were able to resolve most of the WGS sequence was the minimal repetitive content of the chicken genome, so the experience will not necessarily translate to all vertebrate genomes. As we move forward in this post-genomic era, we must learn from all past experience, so that we can maintain the high quality we have come to expect from genome-sequencing projects. ■

Jeremy Schmutz and Jane Grimwood are at the Stanford Human Genome Center, 975 California Avenue, Palo Alto, California 94304, USA. e-mails: jeremy@shgc.stanford.edu jane@shgc.stanford.edu

1. International Chicken Genome Sequencing Consortium *Nature* **432**, 695–716 (2004).
2. International Chicken Polymorphism Map Consortium *Nature* **432**, 717–722 (2004).
3. Wallis, J. W. *et al.* *Nature* **432**, 761–764 (2004).
4. International Human Genome Sequencing Consortium *Nature* **431**, 931–957 (2004).
5. Mouse Genome Sequencing Consortium *Nature* **420**, 520–562 (2002).
6. ENCODE Project Consortium *Science* **306**, 636–640 (2004).
7. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
8. Venter, J. C. *et al.* *Science* **291**, 1304–1351 (2001).
9. She, X. *et al.* *Nature* **431**, 927–930 (2004).

Microbiology

## Jekyll or hide?

George A. O'Toole

Many bacteria can adopt different lifestyles: in a free-living state, they are virulent and cause disease; in a surface-attached community, they are less virulent but may go unnoticed. How is this 'decision' made?

In the November issue of *Developmental Cell*, Goodman and colleagues<sup>1</sup> report the identification of a regulatory system in the bacterium *Pseudomonas aeruginosa* that determines whether it causes disease or lies low and simply persists. This bacterium is of interest to the medical community because of its ability to infect people whose immune system is damaged, who have sustained serious burns or an eye injury, or who suffer from cystic fibrosis. Goodman *et al.* found that inactivating a so-called two-component regulatory system in *P. aeruginosa* results in a strain with a markedly decreased ability to cause disease, but an increased ability to form surface-attached, persistent communities known as biofilms (Fig. 1).

Although there are many variations on bacterial two-component regulatory systems, their basic job is to constantly sample the external environment and transmit this information to the bacterial interior. This allows the organism to adapt to an ever-changing environment. Goodman *et al.*<sup>1</sup> discovered a new protein component of a new such regulatory system, a component that they call RetS.

They also found that a RetS-deficient *P. aeruginosa* strain was better than a wild-type strain at forming a biofilm on both an abiotic surface, namely glass, and a biotic surface, cultured hamster cells. The RetS-deficient bacteria were, however, less able to damage the hamster cells they colonized, and to cause disease in a mouse model of pneumonia. Outside the lab, the ability of *P. aeruginosa* to form biofilms is best known with respect to abiotic surfaces such as catheters, but it might also be able to produce biofilms on tissues within a host, in diseases such as otitis media (earache) and cystic fibrosis<sup>2</sup>. It seems, then, that Goodman *et al.* might have identified a control element that allows this bacterium to switch between a virulent, disease-causing state and a biofilm state in a mammalian host. The biofilm state, although less virulent, might allow the microbe to persist for longer.

To understand better how the protein might control this pathogenesis/persistence switch, the investigators used DNA microarrays to identify all the genes in the organism that are regulated by RetS. They found that, in the RetS-deficient strain, the expression of genes required to make a 'type III secretion