Physical Maps of Chromosomes

David W Ussery, *The Technical University of Denmark, Lyngby, Denmark*

Physical maps of chromosomes are a way of visualizing the entire chromosome in a single figure. Physical maps can come in a variety of formats, including actual photographs of chromosomes, genetic maps where the relative locations of genes are mapped, restriction maps showing the location of restriction enzyme cutting sites along the deoxyribonucleic acid (DNA), and finally maps based on the known DNA sequence along the entire chromosome.

Advanced article

Article Contents

Online posting date: 15th March 2009

Chromosomes contain the genetic information for an organism, and a 'physical map' of a chromosome is usually thought of as an illustration or drawing showing the relative locations of genes along the single continuous sequence of deoxyribonucleic acid (DNA). This is in a same sense as a road map or an atlas. Both are scaled representations of something that cannot be easily visualized. The maps are a way of showing an overview of something on a single piece of paper that is either too large, in the case of lengthy motorways connecting cities, or too small, as in the locations of genes along a chromosome, to easily grasp. The earliest physical maps of chromosomes were quite crude, consisting of a thick line with a few marks on it, designating the relative locations of genes; these are often called 'genetic maps', and originally mapped the relative locations of genes based on genetic experiments, looking for crossovers between neighbouring genes. After the introduction of restriction enzymes to cut DNA sequences at specific sites, it became possible to make a 'restriction map' of a chromosome, and in some cases to more accurately estimate the size, in base pairs (bp) of DNA. More recently, it has become possible to sequence the entire DNA of a chromosome, and hence draw physical maps based now on the DNA sequence.

An organism's genome consists of the total amount of DNA from all its chromosomes. Many organisms contain multiple chromosomes. Individual chromosomes range in size from about a thousand bp (Kbp) of DNA to more than a trillion bp (i.e. a Tbp, or 10^{12} bp), with the general trend that virus chromosomes are smaller (at the time of writing, the smallest virus genome is the Coconut foliar decay virus, 1291 nt long (GenBank accession M29963), and the largest viral genome is the Acanthamoeba polyphaga mimivirus, 1 181 404 nt long [AY653733]), bacteria a bit larger (in the range of 150 Kbp to about 15 Mbp), whereas eukaryotic chromosomes range from about 200 Kbp to around 250 Mbp for humans and many animals, with some organisms having a thousand times more DNA (e.g. some amoebas are thought to have chromosomes in the range of a few Tbp). In the case of a small bacterial virus, the text containing the bases for the entire DNA sequence can fit on a single page. However, for many human chromosomes which are more than 100 000 000 bp long, it hardly makes sense to write this out as a string of As, Gs, Cs and Ts.

ELS subject area: Microbiology

How to cite:
Ussery, David W (March 2009) Physical Maps of Chromosomes. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0001425

Significance of Physical Maps

Photographs of chromosomes date back nearly 100 years, and the idea of some sort of relationship between regions along chromosomes and an organism's appearance shortly followed. One of the first traits to be linked to chromosomes was the sex of an individual. In the case of fruit flies, the sex is determined by the X chromosome, and it was found that this chromosome could sometimes be present in additional copies (e.g. XXY), as shown in **Figure 1** (Bridges, 1916). Based on physical maps, the idea that chromosomes somehow contain the genetic material could be seen and tested; at first, the level was simply observing missing or rearranged chromosomes, and relating these changes in the chromosomes to observed differences in organisms (fruit flies in many of the early experiments). Based on this evidence, it was possible to postulate that changes resulting from 'crossing-over' of chromosomes, which eventually led to the construction of genetic maps, or the relative location of genes along a chromosome.

From an historical perspective, it is hard to overstate the importance of physical maps of chromosomes. Here for the first time, after thousands of years of speculation about the nature of hereditary, was a tangible representation for the genetic material, which could be followed with a microscope. Of course, at the time, it was not known exactly what it was that was inside of these dark-staining threads that was responsible for carrying the genetic information,

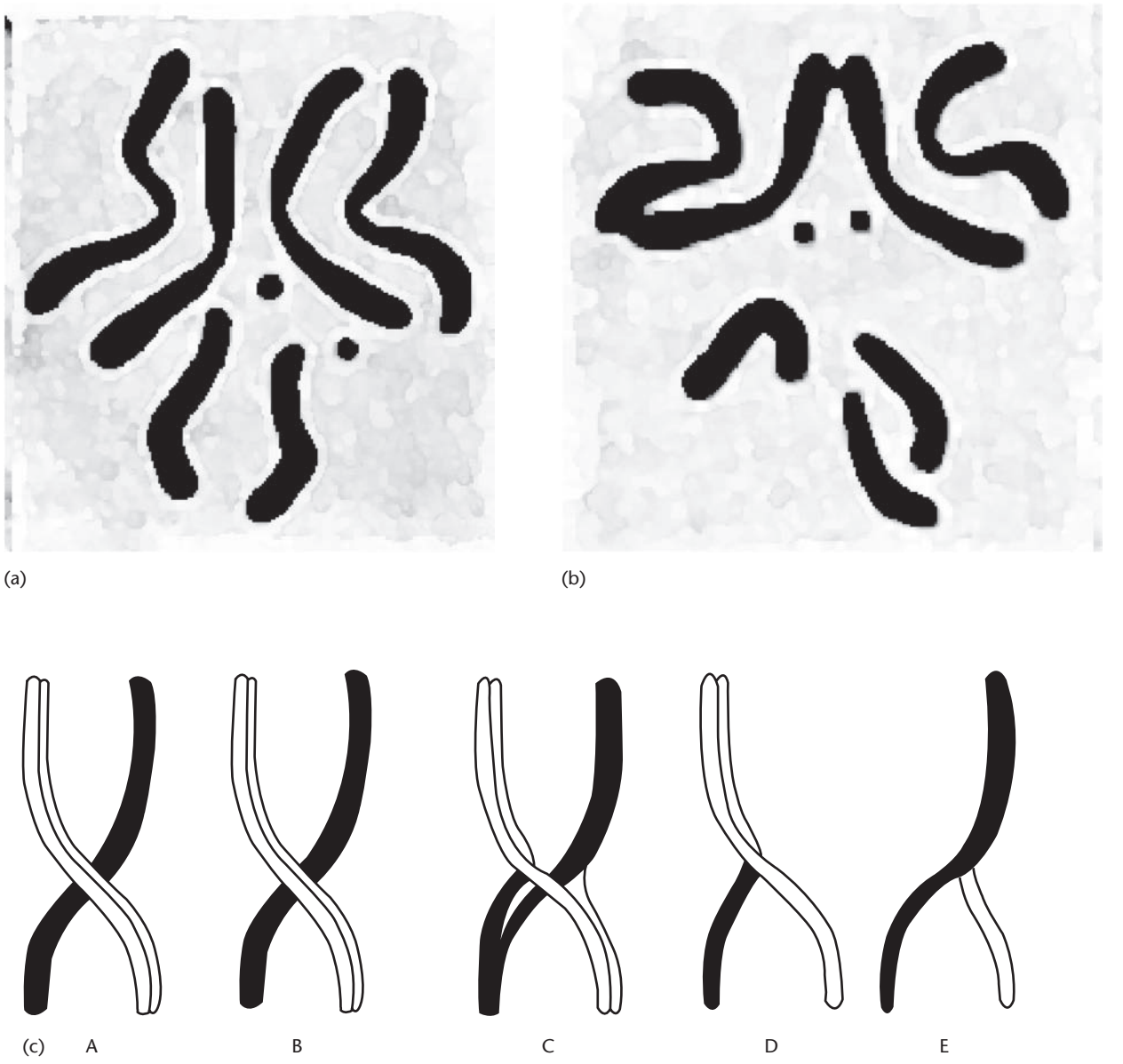

Figure 1 Picture of the four chromosomes from a wild type *Drosophila* female (top left), and XXY female (top right). Shown below is a mechanism for 'crossing-over', which explains the origins of the extra chromosome found in the XXY female. The drawings are based on Bridges (1916).

but nonetheless, at least its physical location was known, and genes could be mapped to relative positions on a line going from one end of the chromosome to the other. Now we know that chromosomes consist of a mixture of about half proteins and half DNA (equal volumes by weight), and that the genetic information is stored on one very long piece of DNA. For example, human chromosome 1 contains one continuous DNA molecule, about 250 000 000 bp long. Completely sequencing such a long chromosome is a real challenge, and at the time of writing none of the human chromosomes have been sequenced to one contiguous piece, although pretty detailed physical maps are available, showing the locations of about 25 000 predicted genes on the 23 chromosomes.

Genetic Maps

Within a few years of the first photographs of chromosomes, as shown in **Figure 1**, very crude physical maps of *Drosophila* chromosomes predicted the relative physical location of a small number of genes along the chromosome. Sometimes, chromosomes can 'crossover', as shown in the bottom of **Figure 1**, which can result in hybrid

chromosomes. By doing experiments with flies having certain genetic traits, some genes were found to be linked to each other – that is, often they would crossover together. Genes which are further apart on the chromosome would crossover together less frequently, and from this 'linkage maps' could be constructed. The distance between genes is measured in map units, and one map unit is called a centimorgan, named after Thomas Hunt Morgan, who is one of the authors of the crossover illustration in **Figure 1**.

By the 1970s, extensively detailed physical maps of *Drosophila* chromosomes were available. However, apart from fruit flies, very few organisms had detailed genetic maps, which require extensive experiments, sometimes over several generations. Thus, doing experiments with an organism which lives for many years, progress was quite slow. When the first genetic maps were made in the early 1900s, at that time it was not realized that the 'genes' mapped were short regions of DNA along one very long piece of chromosomal DNA. The discovery of the structure of DNA by Watson and Crick in 1953 immediately pointed to a possible mechanism of how the genetic material might carry information, and soon the 'General Idea' or 'sequence hypothesis' was proposed – that this information was somehow contained in the sequence of the four DNA nucleotides along the chromosome. An extension of this was the 'Central Dogma', in which this information flows unidirectionally, from DNA sequence to ribonucleic acid (RNA) sequence to protein sequence. Thus the DNA sequence contains genetic information, in terms of coding for proteins. The 1970s also saw the introduction of widespread use of restriction enzymes, which would cut DNA sequences at specific sites, allowing for the construction of restriction maps of chromosomes.

Restriction Maps

With the discovery of restriction enzymes, it became possible to begin to examine the very long chromosomal sequences found in animals. Digestion of a piece of chromosomal DNA with different enzymes can produce a set of smaller fragments, which can easily be analysed by running them out on agarose gels. Using this methodology, it is possible to distinguish between chromosomes of organisms from different individuals, and to get a unique 'DNA fingerprint'. This can be used in forensics, for example, to determine whether a person's DNA matches that recovered at a crime scene, for example.

Restriction mapping also now allowed one to piece together a chromosome, based on a collection of DNA sequences, and not depending on genetic linkage experiments. This allowed resolution of fragments independent of the presence of genes. All that was needed was a piece of DNA. Restriction enzymes also paved the way for sequencing chromosomes, because instead of sequencing the entire chromosome from one end to the other, it was now possible to chop the DNA up into smaller pieces of known size, and determine their sequence, and then fit these together, like pieces in a puzzle, to come up with the complete sequence.

Optical mapping is an extension of restriction mapping. In this case, the chromosome is digested with a restriction enzyme and then instead of running the products on a gel, the fragments are prepared on a flat surface and photographed. Because the distance is known between the DNA bases stacked along the helix (0.34 nm per step), then the total number of bp can be estimated from the length of a stretched out piece of DNA. This method is becoming more popular, and for example, recently a set of genomes from 11 different *Escherichia coli* O157 isolates were anlaysed using optical mapping, after cutting with the *BamHI* restriction enzyme. The size varied between 5.3 and 5.6 Mbp, and there were several inversions found within the chromosomes (Kotewicz *et al*., 2007). Currently this method is considerably less expensive than sequencing the 11 genomes.

Complete Genome Sequence Maps

Soon after the discovery of restriction enzymes, people begin to realize that it would be possible to chop of the DNA sequence of a chromosome into smaller pieces, and sequence these. In 1978, the complete sequence of the bacteriophage phiX174 was published (Sanger *et al*., 1978). This was a landmark in that it was the first chromosome to be completely sequenced. A 'DNA structural atlas' for this sequence is shown in **Figure 2**; the atlas is based on the chromosomal DNA sequence, and graphically displays DNA structural properties of the sequence along the chromosome (Pedersen *et al*., 2000). Comparison of this atlas with the restriction maps shown in **Figure 3** reveals some of the additional information which can be obtained from having the whole DNA sequence of the chromosome. Not only is it clear where genes start and stop, but also there are variations in AT content (the fraction of As and Ts of the total content of Gs, As, Cs and Ts for the entire DNA sequence, shown as the red inner circle, where darker red means regions of higher amounts of As and Ts) throughout the genome. Furthermore, there is a structural signature for promoters – a region that is curved, followed by a region that is rigid and will melt easily; also, the presence of certain short palindromic repeats are associated with the end of the transcripts (Ussery *et al*., 2008).

The Sanger dideoxy method for sequencing eventually led to the development of new technologies, including the automation of DNA sequencing. Starting in the 1980s, there was a very large investment in increasing the speed of sequencing, such as the U.S. Human Genome Project, which invested $200 000 000 per year towards this goal. At the beginning of the project (in the mid-1980s), sequencing the 3 Gbp human genome at the same rate of the current technology would take more than a million years to finish! In large part because of this heavy investment in sequencing technology, the human genome was declared 'finished' to the level of a draft sequence, back in 2000. However,

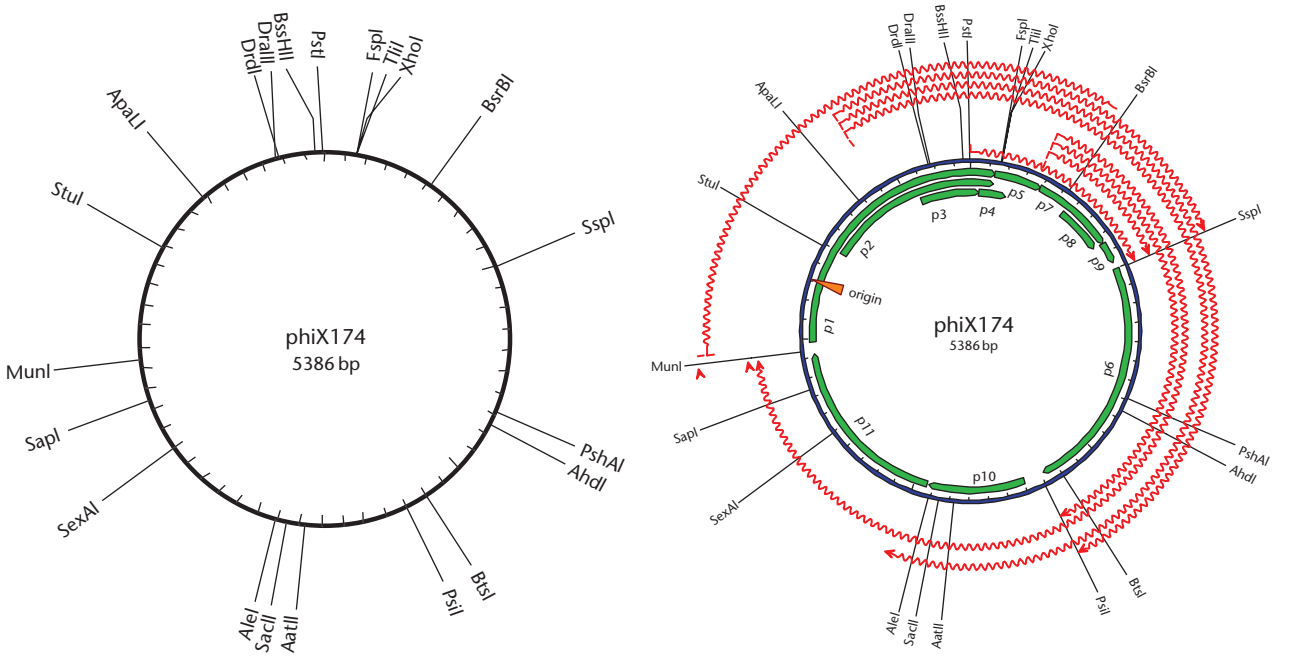

Figure 2 Genome atlas for phiX174. The outer three lanes correspond to DNA structural properties (curvature, stacking energy and nucleosomal 'position preference'). All the genes go in the same direction (clockwise), and hence are annotated as blue. Two different types of inverted repeats are indicated in the next two lanes, followed by GC skew and the inner-most circle is the relative AT content (red being more AT rich, and blue represents GC rich). Hansen Permissions.

because of the large numbers of repeats, the human genome is not likely to be completely sequenced, in terms of having one contiguous piece for each chromosome, for many years. In the case of *Drosophila*, an attempt is underway to systematically sequence all the heterochromatin regions containing many repeats (Hoskins *et al*., 2007), although at present such an effort to map more than 60% of the human genome containing repeats is not being done. That is, less than half the human genome has been literally sequenced, in terms of knowing the four bases (GATC) for every position along the chromosome. Having said that, the ENCODE project has attempted to thoroughly go through about 1% of the human genome, and look for genes and other experimental information in a careful and systematic way. The results have found that there is much more there than has been thought by many, and a lot more complicated in terms of what really defines a gene (Gerstein *et al*., 2007). Plans are underway to scale this project for the whole human genome, over the next several years. Currently, the draft of the human sequence has allowed for the mapping of about 25 000 genes along the 23 human chromosomes. If all the chromosomes were of the same size, this would be a bit more than a thousand genes per chromosome; however, the largest human chromosome is quite large, almost 10% of the total size of the genome, and is thus estimated to contain roughly 2500 genes.

Technology continues to progress, and the cost of sequencing is dropping dramatically. The first draft of the human genome cost more than US$3 billion, and took 15 years to finish. The second human genome (and the first individual, J. Craig Venter) to be sequenced cost $100 million, and took 9 months to finish. The third human genome (Jim Watson) to be sequenced took less than 2 months, and cost less than $1 million. Currently there is a prize of US$10 000 000 to sequence 100 human genomes in 10 days or less, for about US$10 000 per genome. This means that it would cost only US$1 to sequence a bacterial genome! At the time of writing, there are more than 750 sequenced bacterial genomes which have been published, with an additional 1200 sequences 'in progress'. The largest bacterial genomes that have been completely sequenced are around 10 Mbp, and code for more than 9000 genes, or more than a third as many genes as in the human genome.

Figure 4 is a genome atlas for *Bradyrhizobium japonicum*, strain USDA 110 (Kaneko *et al*., 2002), which is the fifth largest of the bacterial genomes sequenced at the time of writing. Its genome sequence has 8317 genes annotated, or more than three times the total number of genes on the largest of human chromosome 1, even though the bacterial chromosome is 25 times smaller. Notice that, although the chromosome shown in **Figure 4** has about a thousand times as many genes as in the chromosome shown in **Figure 2**, the atlases look generally the same, although of course there is a lot more data behind the latter figure. But still this allows one to get a general idea of regions of variation of the DNA sequence along the chromosome. One obvious area of interest in **Figure 4** is the 'symbiosis island', identified by Kaneko *et al*. in their original genome paper. Based on oligomer skew analysis of the whole chromosome, we estimate the replication origin to be around 682 Kbp, in agreement with the location of the parA and

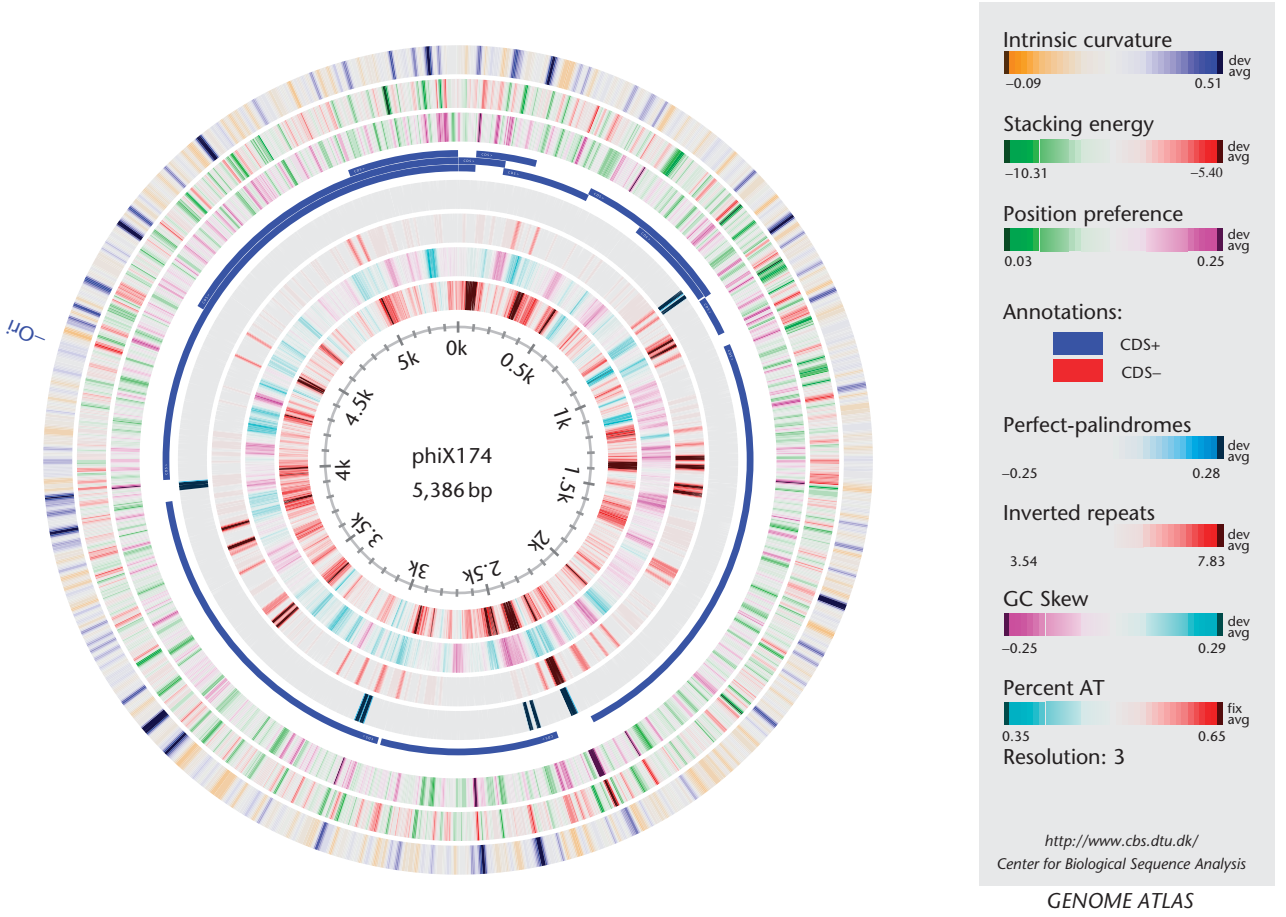

Figure 3 Restriction map (left), and gene map (right) for the bacteriophage phiX174. Note that the genes are all in the same orientation, as indicated by the green arrows. The red lines represent different mapped transcripts. These figures were kindly provided by Flemming Hansen.

parB genes, and the replication terminus to be around 5002 Mbp. Note that the dark green regions in the nucleosomal 'position preference' lane (third from the outside) are possible regions of more flexible chromatin structure, which might contain highly expressed genes (Pedersen *et al*., 2000). Thus a great deal can be learned about a genome, just from a simple overview of its sequence plotted in a structural atlas.

Summary

There are several different types of physical maps, ranging from photographs of chromosomes to genetic maps to restriction maps to actual complete chromosome sequences, which can then be used to generate 'chromosome atlases'. Obviously, having the full sequence is the most optimal in terms of generating high-density information maps of the chromosome. Fortunately, with sequencing technology becoming less expensive, it is now realistic to start thinking about comparison of multiple genomes within a population. For example, soon it will be possible to sequence all the organisms in a culture collection, and compare the many different isolates. As a first step towards beginning to analyse this data, we have developed 'BLAST atlases', which allow the comparison of multiple genomes of the same taxonomic group. Links to some examples of these can be found in the supplemental information.

Supplemental Information

NCBIs list of sequenced microbial genomes: http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi

Genome atlases for sequenced microbial genomes: http://www.cbs.dtu.dk/services/GenomeAtlas/

'BLAST atlases', comparing closely related sequenced bacterial genomes: http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/zoomatlas/

Information on xprize: http://www.xprize.org/xprizes/genomics_x_prize.html

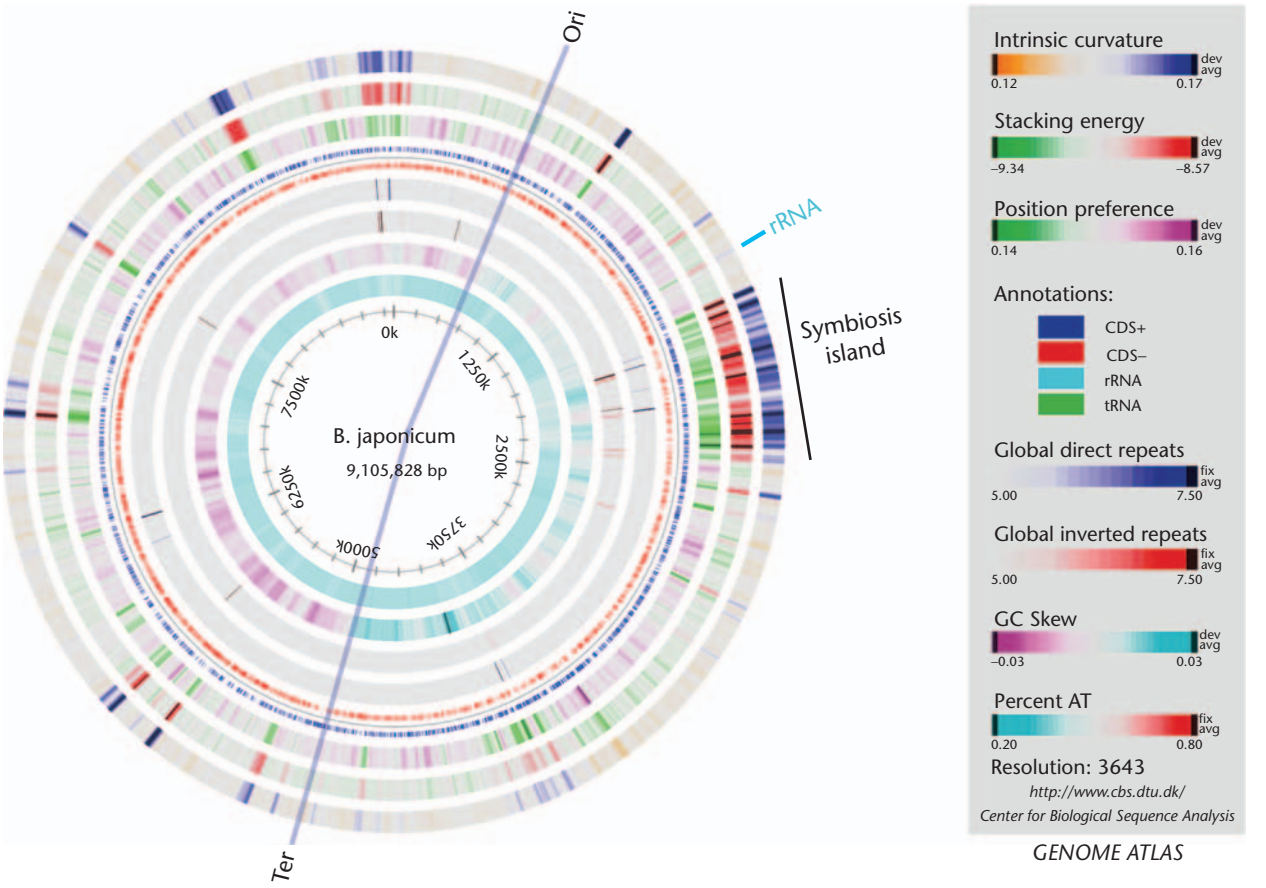

Figure 4 Genome atlas for the bacterial genome of *Bradyrhizobium japonicum*, strain USDA 110. The lanes are similar to those in the atlas in **Figure 2**, except the repeats are now global direct and inverted repeats, which represent regions of at least 100 bp with homology either on the same strand (direct repeats), or on the opposite strand (inverted repeats). As is characteristic for many bacterial genomes, the overall level of repeats is quite low, with less than 5% of the genome containing repeats with more than 80% homology.

References

Bridges CB (1916) Non-disjunction as proof of the chromosome theory of heredity (concluded). *Genetics* **1**: 107–163 PMID: 17245853 [PubMed].

Gerstein MB, Bruce C, Rozowsky JS *et al*. (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Research* **17**: 669–681 PMID: 17567988 [PubMed – in process].

Hoskins RA, Carlson JW, Kennedy C *et al*. (2007) Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* **316**: 1625–1628.

Kaneko T, Nakamura Y, Sato S *et al*. (2002) Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Research* **9**: 189–197 PMID: 12597275 [PubMed – indexed for MEDLINE].

Kotewicz ML, Jackson SA, LeClerc JE and Cebula TA (2007) Optical maps distinguish individual strains of *Escherichia coli* O157: H7. *Microbiology* **153**: 1720–1733 PMID: 17526830 [PubMed – in process].

Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH and Ussery DW (2000) A DNA structural atlas for *E. coli*. *Journal of Molecular Biology* **299**: 907–930 PMID: 10843847 [PubMed – indexed for MEDLINE].

Sanger F, Coulson AR, Friedmann T *et al*. (1978) The nucleotide sequence of bacteriophage phiX174. *Journal of Molecular Biology* **125**: 225–246 PMID: 731693 [PubMed – indexed for MEDLINE].

Ussery DW, Borini S and Wassenaar TM (2008) *Computing for comparative genomics: Bioinformatics for microbiologists*. London: Springer-Verlag.

Further Reading

Gregory TR (ed.) (2005) *The Evolution of the Genome*. London, UK: Elsevier Academic Press.

Morgon TH (1919) *The Physical Basis of Heredity, Monographs on Experimental Biology*. Philadelphia: J.B. Lippincott.

Morton NE (2005) Linkage disequilibrium maps and association mapping. *Journal of Clinical Investigation* **115**: 1425–1430 PMID: 15931377 [PubMed – indexed for MEDLINE].

O'Brian S, Menninger JC and Nash G (eds) (2006) *Atlas of Mammalian Chromosomes*. Hoboken, NJ: Wiley.

Valouev A, Schwartz DC, Zhou S and Waterman MS (2006) An algorithm for assembly of ordered restriction maps from single

DNA molecules. *Proceedings of the National Academy of Sciences of the USA* **103**: 15770–15775 PMID: 17043225 [PubMed – indexed for MEDLINE].

Yim YS, Moak P, Sanchez-Villeda H *et al*. (2007) A BAC pooling strategy combined with PCR-based screenings in a large, highly repetitive genome enables integration of the maize genetic and physical maps. *BMC Genomics* **8**: 47 PMID: 17291341 [PubMed – indexed for MEDLINE].

Zhao S and Stodolsky M (2004) *Bacterial Artificial Chromosomes: Volume 1: Library Construction, Physical Mapping, and Sequencing (Methods in Molecular Biology)*. Totowa, NJ: Humana Press.