Genome data: what do we learn?

William C Nierman*, Jonathan A Eisen†, Robert D Fleischmann‡ and Claire M Fraser§

Genome sequence information has continued to accumulate at a spectacular pace during the past year. Details of the sequence and gene content of human chromosome 22 were published. The sequencing and annotation of the first two *Arabidopsis thaliana* chromosomes was completed. The sequence of chromosome 3 from *Plasmodium falciparum*, the second sequenced malaria chromosome, was reported, as was that of chromosome 1 from *Leishmania major*. The complete genomic sequences of five microbes were reported. Approaches to using data from completely sequenced microbial genomes in phylogenetic studies are being explored, as is the application of microarrays to whole genome expression analysis.

Addresses
The Institute for Genomic Research, 9712 Medical Center Drive,
Rockville, MD 20850, USA
*e-mail: wnierman@tigr.org
†e-mail: jeisen@tigr.org
‡e-mail: rdfleisc@tigr.org
§e-mail: cmfraser@tigr.org

Current Opinion in Structural Biology 2000, **10**:343–348

0959-440X/00/$ – see front matter

Abbreviations

EST	expressed sequence tag
INH	isoniazid
IPTG	isopropyl-β-D-thiogalactopyranoside
ORF	open reading frame
PCR	polymerase chain reaction

Introduction

This review will focus on the progress reported in genome sequencing and analysis for the 12 months ending January 2000. Initial progress on the sequencing of the human genome is reported in papers on chromosome 16p and 16q, and chromosome 22. Two papers report the sequencing of two of the five chromosomes from the model plant *Arabidopsis thaliana*, chromosome 2 and chromosome 4. These are milestone papers in the sequencing of the first human and plant chromosomes, and provide the first systematic look at whole chromosome structure in these organisms. The complete sequences of five additional microbes were reported: from the evolutionarily deep-branching eubacterium *Thermotoga maritima*; from the most radiation-resistant organism known on the planet, *Deinococcus radiodurans;* from the hyperthermophile archaeon *Aeropyrum pernix*; and from the pathogens *Chlamydia pneumonia* and the second sequenced strain of *Helicobacter pylori*. A substantial fraction of the open reading frames (ORFs) identified in newly sequenced genomes or portions of genomes are for ORFs with no known function. This observation continues to dramatize the need for new approaches to large-scale functional analysis.

With the availability of 24 completely sequenced and annotated microbial genomes, and three chromosomes, methods for applying this dataset to phylogenetic studies are being explored. In addition, whole genome microarray technology is being applied to generate data for whole genome expression analysis.

Human genome sequencing

The sequencing and analysis of 11.8 Mb from the centromere-flanking region of human chromosome 16p and 16q was reported by Loftus *et al.* [1•]. This region was revealed to be gene poor, with only 84 predicted genes. Genes identified only by gene prediction software were not included in this count. The region may thus contain many more genes. One of the most interesting aspects of this region of the genome, as noted, was the presence of large tracts of highly homologous sequences (greater than 20 kb) distributed throughout the p-arm. Some of these ‘duplicons’ included portions of known genes and other apparently transcribed sequences. All but five of the predicted genes had at least one match to a human expressed sequence tag (EST). This analysis provided a first glimpse of the complexity of the genome and allowed an assessment of the usefulness of a number of current gene and exon prediction methods and an evaluation of the EST databases in gene structure determination.

The DNA sequence of the euchromatic portion of human chromosome 22 was reported by Dunham *et al.* [2••]. This is the first chromosome for which the “operationally complete” sequence was established by the Human Genome Project. The sequence consists of 12 contiguous segments spanning 33.4 Mb on chromosome 22q. The sequence regions contain a 23 Mb central contig, the largest continuous segment of DNA sequence known to date. No gap in the sequenced region is determined to be larger than 150 kb, as judged by fiber FISH analysis. Chromosome 22 is an acrocentric chromosome, with the short arm containing tandemly repeated rRNA genes and a series of other tandem repeat sequence arrays.

As was identified for chromosome 16 (above), long-range duplications were identified on chromosome 22; two of these were about 60 kb in size, with greater than 90% sequence similarity. The number of genes annotated for chromosome 22q is 679, using a combination of EST and protein databases, and gene prediction software (see News flash).

Arabidopsis thaliana genomes

The sequencing and analysis of chromosomes 2 and 4 from the model plant *A. thaliana* has been completed [3••,4••] as part of the international *Arabidopsis* Genome Initiative to sequence the entire *Arabidopsis* genome. This is a landmark accomplishment, revealing the structure of 30% of the *Arabidopsis* genome and setting a high standard for complete sequence coverage of organisms with larger and more complex genomes. The 16 Mb contig on the lower arm of chromosome 2 constituted, for a time, the largest sequence assembly recorded in the GenBank database; with the analysis of the human chromosome 22 large contig, this became the second largest.

Analysis of the sequence data has provided information on chromosome structure and organization for this plant. Gene density is relatively high and uniform outside the regions adjacent to the genetically defined centromeres, and is low in this pericentromeric region. The inverse is true of transposon and intermediate repeat densities, which are concentrated in the pericentromeric region. Large chromosome duplications were detected; a smaller duplication of about 1 Mb was detected between the lower segments of chromosomes 1 and 2, and a 4.6 Mb duplication was found between chromosomes 2 and 4. Other smaller regions of chromosome duplication were reported.

More than 60% of the predicted proteins on chromosome 2 have significant similarity to other *A. thaliana* proteins. About 20% of these chromosome 2 genes are found in tandem duplications that range in size from two to nine genes. A lower number of genes with *A. thaliana* homologs were observed on chromosome 4. Some of the duplicated genes are found in the large chromosomal duplications.

The sequencing of these chromosomes afforded the opportunity to explore the transfer of genes from mitochondria and chloroplasts to the nucleus. The analysis of genes on both chromosomes was accomplished using two methods for identifying nuclear genes that ancestrally may have been organellar genes. The presence of putative N-terminal signal peptides for chloroplast and mitochondrial localization was presumed to be evidence of an ancestral organellar location. Sequence similarity to genes in *Synechocystis,* the only sequenced cyanobacterium [5], was the second method. Three percent of the chromosome 2 genes met one or both of these criteria. The chromosome 4 result was 18%. This difference in number between the two chromosomes undoubtedly is the result of more stringent criteria applied to identifying such genes on chromosome 2. In addition, the chromosome 2 sequence revealed a 270 kb stretch of sequence that is nearly identical to that of the *Arabidopsis* mitochondrial genome. Its 99% identity to the published mitochondrial sequence suggests that this mitochondria to nucleus transfer event was very recent.

The sequencing of chromosomes 2 and 4 revealed 4037 and 3825 genes, respectively. The fractions of these genes with no predicted function on the basis of similarity to known genes are high, 48% and 40%, respectively. As for other sequenced genomes, assigning functions to these genes will require a major post-sequencing effort on the part of the *Arabidopsis* community.

Copenhaver *et al.* [6•] used the “virtually complete” centromere sequences of chromosomes 2 and 4 to compare the relationship between genetically defined centromere activity (suppressed recombination) and DNA sequence. A comprehensive analysis of the sequence motifs, DNA modifications and structural features that may contribute to centromere function is provided. Functional genes were observed to be located within the *Arabidopsis* centromeres. This is a potentially important observation for determining completion criteria for the genome sequencing projects of other multicellular organisms, as centromere sequences are frequently excluded from the regions that are to be sequenced.

Pathogenic protozoan genomes

The sequencing of the genome of the malaria pathogen *Plasmodium falciparum* is being accomplished by an international consortium using a chromosome by chromosome shotgun approach. There have already been major rewards from this project, including the discovery of both a new class of variant antigen and potential drug targets. Chromosome 2 was the first sequenced chromosome [7••]. The sequencing of chromosome 3 [8••] affords the kinds of comparisons cited above for the two *Arabidopsi*s chromosomes. The comparison showed a conservation of the order of features between the two chromosomes in the subtelomeric regions; this conservation includes members of multigene families involved in pathogenesis and antigenic variation. Additionally, a putative centromere has been identified in both chromosomes that is about 2 kb in size, has an extremely high A+T (97.3%) content and contains arrays of tandem repeats.

At 1.1 Mb, chromosome 3 contains predicted coding regions for 215 protein-coding genes and two tRNA-coding genes. From the presence of N-terminal organelle-targeting sequences, three predicted proteins from this chromosome were identified as being potentially targeted for import into the mitochondria and two others are targeted for import into a second organelle, the apicoplast.

Leishmania are protozoan pathogens that cause a spectrum of diseases in humans, resulting in considerable suffering and death. An international effort to map and sequence the *Leishmania* genome has been initiated by the *Leishmania* Genome Network (www.ebi.ac.uk/parasites/leish.html). The organism is diploid, with 36 chromosome pairs and a haploid genome size of 34 Mb [9]. The sequence of the smallest chromosome [10•], chromosome 1 of the reference strain *Leishmania major* Friedlin, was determined by sequencing mapped cosmid clones. The sequence revealed an unusual chromosome organization. The chromosome is composed of a 257 kb information-rich region containing 79 protein-coding genes. This region is flanked by telomeric and subtelomeric repeat elements that vary in size among chromosome 1 homologs. In the information-rich portion, the first 29 genes are all encoded on one DNA strand, whereas the remaining 50 are on the other, possibly reflecting the presence of two polycistronic transcription units. Of the 79 genes, 52% showed similarity to proteins with no know function. The majority of these were unique to *Leishmania* (41%).

Microbial genome sequencing

The complete genome sequences of three additional microbial species of environmental significance were reported during the review period. *A. pernix* K1 was the first sequenced aerobic hyperthermophilic archaeal species [11], with an optimum growth temperature of 95°C. The genome is 1.7 Mb in size and contains 2694 potential protein-coding genes identified on the basis of very permissive ORF-finding criteria. 1538 (57.1%) of these genes did not show significant similarity to sequences in the databases. *T. maritima* MSB8 is a deep-branching eubacterium with a 1.8 Mb genome. The sequence revealed 1877 predicted coding regions, 863 (46%) with unknown functional assignments [12••]. This organism has 81 of its genes in 15 clusters of 4–20 kb; these genes are most similar to genes from sequenced archaeal species. Gene order is conserved in many of these clusters, suggesting that lateral gene transfer may have occurred between thermophilic eubacteria and archaea. *D. radiodurans* R1 is a member of the most radiation-resistant and stress-resistant species known. The complete genome sequence [13••] revealed a genome size of 3.3 Mb in two chromosomes and two plasmids. The analysis of the gene content showed that extensive systems for DNA repair, DNA damage export, and desiccation and starvation recovery are present in this organism. There is considerable redundancy of the genes within these systems. Of the 3187 predicted genes, functional assignments could not be made for 1694 (53%) of them. Of these, 1002 had no database matches. These constitute a large set of genes that may participate in unknown ways in the stress resistance of this organism.

The complete genome sequencing of two bacteria allowed extensive comparison with a closely related strain or species. The sequencing of *H. pylori* J99 [14] afforded a comparison of this strain with the previously sequenced *H. pylori* 26695 strain [15]. Strain-specific genetic diversity has been proposed to account for different human disease phenotypes. The two sequenced strains were found to be very similar, with only about 6% of the genes specific to each strain. Over half of these strain-specific genes have no significant similarity with genes in public databases. The sequencing of *C. pneumoniae* [16] afforded a comparison with previously sequenced *Chlamydia trachomatis* [17]. These are obligate intracellular eubacteria that differ in their tissue tropism and disease spectrum. Analysis of the *C. pneumoniae* genome revealed 214 protein-coding sequences not found in *C. trachomatis*; again most without homology to known sequences. The major functionally identifiable addition to the *C. pneumoniae* genome is an expansion (from 9 to 21) in the number of genes encoding a polymorphic membrane protein.

Microbial evolution

The availability of 24 completed microbial genome sequences has provided new insights into microbial evolution and diversity. The molecular picture of evolution for the past 20 years has been dominated by the small-subunit rRNA phylogenetic tree derived by Woese and Fox, which proposes three nonoverlapping domains of living organisms: bacteria, archaea and eukaryotes [18]. Although archaea possess bacterial cell structures, it has been suggested that they are no more closely related to bacteria than to eukaryotes. This three domain proposal also posits that archaea and eukaryotes shared a common ancestor exclusive of bacteria or, in other words, the common ancestor of eukaryotes descended directly from within the archaeal lineage.

As a result of the completion of genome sequences from representatives of all three domains of life, it is now possible to examine evolutionary relationships among living organisms in a more comprehensive way. However, this task has turned out to be anything but straightforward. Incongruities can be seen everywhere in the phylogenetic tree, from its root to the major branchings, when single protein phylogenies are examined. It has become clear that gene evolution does not equal species evolution. This, in a large part, is a result of extensive lateral gene transfer, not only between bacteria, but also between bacteria and archaea. Additional reasons cited to account for these observations include gene displacement, gene duplication followed by specialization or extinction, and convergence at the molecular level.

By comparing the protein sequences encoded in the four archaeal species whose genomes have been completely sequenced, Makarova *et al.* [19•] have defined an evolutionary core of genes in archaeal genomes on the basis of their presence in all four species. These core genes (31%–35% of the genome content) code primarily for proteins involved in genome replication and expression. Specific metabolic functions are more sporadically present in the four genomes. An additional observation is that the core genes are more similar to eukaryotic counterparts, whereas genes present in only two or three of the species are most similar to bacterial homologs. The authors suggest that this may be due to lateral gene transfer of metabolic genes from bacteria in the evolution of archaea. The corollary to this observation is that the core genome replication and expression genes have not been laterally transferred. Perhaps the complexity of the multiprotein structures required for replication and expression — replication complexes, transcription complexes and ribosomes — makes efficient lateral transfer and fixation of the transferred gene or genes statistically extremely improbable.

An alternative to single gene phylogenies is to build 'average' phylogenetic trees for the whole genome on the basis of gene content. The first such 'gene content' phylogenetic analysis was reported by Snel *et al.* [20••], who showed that a distance tree based on the number of genes shared between genomes is remarkably similar to the rRNA tree for those same species. Subsequently, Fitz-Gibbon and House [21•] showed similar results using a parsimony analysis of the presence and absence of genes. In addition, Tekaia *et al.* [22] showed that a hierarchical classification method (which is a clustering-based method and thus is not a true phylogenetic method) also gives similar results. These studies show that there is a basic 'average' phylogenetic history for each species that is recoverable, even though individual genes may not follow this pattern.

Improved annotation from comparative genomics

A key part of genome sequencing projects is the identification and subsequent prediction of functions for genes in each genome. Although it is important to remember that these are only predictions, such predictions can be very useful for experimental studies of these species. Two important and creative new methods of function prediction were reported in 1999. The first involves the use of information on the presence and absence of particular genes in different species. Eisenberg and colleagues [23••] developed a method that clusters genes according to their distribution patterns across species, a method they refer to as phylogenetic profiling. They showed that genes that work in the same pathways frequently have such correlated distribution patterns that the functions of some unknown genes can be predicted on the basis of their having similar distribution patterns to genes with known functions. The correlated loss of genes in certain lineages is also a useful predictor of gene function [24•].

Another method for function prediction in which protein domain patterns are used to link proteins together into functional groups was reported by two groups simultaneously [25•,26•]. In this method (referred to as the Rosetta stone method by Marcotte *et al.* [25•]), the presence of a two-domain protein in one species is used to cluster proteins that contain either one (or the other) domain, using the assumption that the two separated domains probably work together. This can be done for proteins with any number of domains. The resulting linkages cluster genes that are known to be part of the same pathway and thus probably can be used to better predict functions for many genes.

Microbial genome expression analysis

Beyond trying to decipher molecular evolution, another formidable challenge in microbial genomics is how to make use of the new sequence information on a large scale in order to better understand biology. By using approaches that include gene chips, microarrays and proteome analysis, it should be possible to move from a static picture of a genome, as captured in a set of DNA and protein sequences, to the identification of gene networks and a better understanding of the dynamic nature of the regulation of gene expression in the microbial cell.

In many laboratories, whole genome expression analyses based on the ORFs identified in genome sequencing projects are underway. PCR products prepared from each ORF are spotted in a high-density array on a glass microscope slide. These microarrays are probed with fluorescently labeled whole organism cDNA prepared from total RNA. High-resolution image scanners and analysis software quantify the signal intensity from each spot on the slide to determine the mRNA level from the cell of the ORF represented by the spot. This microarray methodology is particularly powerful in quantifying differential levels of expression of each ORF for cells grown under different conditions.

Whole genome expression analyses of *Escherichia coli* and *Mycobacterium tuberculosis* using whole genome microarrays were reported by Richmond *et al.* [27] and by Wilson *et al.* [28••], respectively. The *E. coli* project measured changes in RNA levels before and after exposure to heat shock and following treatment with isopropyl-β-D-thiogalactopyranoside (IPTG). Treatment with IPTG resulted in induction of the *lacZYA* and *melAB* operons. Heat shock significantly altered the expression levels of 119 genes, including 35 ORFs that were previously uncharacterized. Analysis of signal intensities suggested that at least 25% of the *E. coli* genes were expressed at detectable levels during growth in a rich medium. This analysis was intended to be an initial validation of the whole genome *E. coli* microarray.

The *M. tuberculosis* analysis determined alterations in RNA levels after treatment with the drug isoniazid (INH). This drug was selected for this first study because it is given to more tuberculosis patients than any other and because it is the drug against which resistance emerges most frequently. INH was found to induce several genes physiologically relevant to the drug's mode of action. INH selectively inhibits the synthesis of mycolic acids, the major component of the waxy outer lipid envelope of mycobacteria. The genes are induced within a fraction of a generation time following the addition of INH and some are directly involved in the processes inhibited by INH. Because the affected enzymatic pathway contains proven drug targets, perhaps other proteins operating in the same pathway, as revealed by the microarray data, might also be appropriate targets for new drug development. Patterns of the induction and repression of gene expression may also prove valuable in designing screens for novel compounds that exert similar effects. These kinds of applications are two of the valuable potential uses of microarray technology in drug discovery and validation.

Conclusions and future directions

The early events of the first phase of the development of the science of genomics can perhaps be listed as initiation of the Human Genome Project, high-throughput human EST sequencing and the completion of the total sequence of *Haemophilus influenzae*. This first phase can now be declared at an end, with the routine accomplishment of whole genome sequencing. An avalanche of genome sequence data is being generated and a second phase of the development of the science of genomics is necessary for producing the tools for dealing with this data. It is not merely a matter of developing software tools either for managing sequence data or for finding genes in eukaryotic DNA sequences but, more importantly, tools for moving from the sequence data to an understanding of the organism's biology are needed.

Approaches for making phylogenetic sense of the gene content of organisms are starting to be explored, as is the use of microarrays for expression analysis on a whole genome scale. Future genome scale analyses will directly lead to powerful approaches for vaccine and drug development, for improved diagnostics, for understanding host–pathogen interactions and for dissecting the genetic components of human disease. Additionally, these analyses will reveal the inner workings of biological systems and provide insight into the evolutionary history of life on earth.

The next step is to address the issue of the fact that a substantial fraction of the gene content of even the most extensively studied organisms are for genes with no known function. The great immediate challenge remains finding efficient ways to identify the function of these genes and determining how all the genes work together to make an organism what it is. This is the great challenge that our success in genomics has brought to us. Our future accomplishments in this endeavor will probably reveal things about biology and biological systems that have, to date, been out of reach of even the human imagination.

News flash

By searching the EST-based TIGR (The Institute of Genomic Research) Human Gene Index against the human chromosome 22 sequence, Liang *et al.* [29] obtained 1326 high-scoring (98% identity, more than 150 base pairs) matches, including 1153 that did not hit genes annotated in the published sequence [2••]. Thus, the number of genes on chromosome 22 is probably more than twofold greater than the number initially reported. This also suggests that the number of genes in the human genome is significantly higher than the 45,000 lower estimate presented by the Chromosome 22 Consortium.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Loftus BJ, Kim UJ, Sneddon VP, Kalush F, Brandon R, Fuhrmann J,
• Mason T, Crosby ML, Barnstead M, Cronin L *et al.*: **Genome duplications and other features in 12 Mb of DNA sequence from human chromosome 16p and 16q.** *Genomics* 1999, **60**:295-308.

The authors provide an early observation of the organization of the human genome, as revealed through the Human Genome Project sequencing. The use of gene finding tools and procedures is also reported.

2. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE,
•• Bruskiewich R, Beare DM, Clamp M, Smink LJ *et al.*: **The DNA sequence of human chromosome 22.** *Nature* 1999, **402**:489-495.

This landmark paper reports the "operationally complete" sequence of the first sequenced human chromosome. Human genome organization characteristics reported previously [1•] are confirmed in this larger scale analysis.

3. Lin X, Kaul S, Rounsley S, Shea TP, Benito M, Town CD, Fujii CY,
•• Mason T, Bowman CL, Barnstead M *et al.*: **Sequence and analysis of chromosome 2 of *Arabidopsis thaliana*.** *Nature* 1999, **402**:761-768.

Two papers [3••,4••] represent landmark reports of the first two plant chromosomes to be sequenced. The reports describe the chromosome organization, repeat content, gene content and provide a description of the status of the migration of organellar genes to the nuclear genome. The insertion of a major portion of the mitochondrial genome into chromosome 2 is reported.

4. Wambutt R, Murphy G, Volckaert G, Pohl T, Dusterhoft A,
•• Stiekema W, Entian KD, Terryn N, Harris B, Ansorge W *et al.*: **Sequence and analysis of chromosome 2 of *Arabidopsis thaliana*.** *Nature* 1999, **402**:769-777.

See annotation to [3••].

5. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S *et al.*: **Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions.** *DNA Res* 1996, **3**:109-136.

6. Copenhaver GP, Nickel K, Kuromori T, Benito M-I, Kaul S, Lin X,
• Bevan M, Murphy G, Harris B, Parnell LD *et al.*: **Genetic definition and sequence analysis of *Arabidopsis* centromeres.** *Science* 1999, **286**:2468-2474.

The authors report the findings of a focused analysis of the *Arabidopsis* centromeres that are derived from the genome sequencing project. The most striking finding is that functional genes are located within the genetically defined centromeres on chromosomes 2 and 4.

7. Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L,
•• Koonin EV, Shallom S, Mason T, Yu K, Fujii C *et al.*: **Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*.** *Science* 1998, **282**:1126-1132.

This paper, together with [8••], reports the sequencing of two *P. falciparum* chromosomes. These papers describe a new class of variant antigen and potential drug targets for the development of malaria vaccines and therapeutics. The analysis of the sequences of chromosomes 2 and 3 has allowed the identification of putative centromeres.

8. Bowman S, Lawson D, Basham D, Brown D, Chillingworth T,
•• Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T *et al.*: **The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*.** *Nature* 1999, **400**:532-538.

See annotation to [7••].

9. Ivens AC, Lewis SM, Bagherzadeh A, Zhang L, Chan HM, Smith DF: **A physical map of the *Leishmania major* Friedlin genome.** *Genome Res* 1998, **8**:135-145.

10. Myler PJ, Audleman L, de Vos T, Hixson G, Kiser P, Lemley C,
• Magness C, Rickel E, Sisk E, Sunkin S *et al.*: ***Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes.** *Proc Natl Acad Sci USA* 1999, **96**:2902-2906.

The authors report the unusual gene and chromosome organization of this organism. Chromosome 1 has a gene-rich core containing 79 genes; 29 are encoded in tandem on one DNA strand, whereas the remaining 50 are on the other.

11. Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, Takahashi M, Sekine M, Baba S, Ankai A *et al.*: **Complete genome sequence of an aerobic hyper-thermophilic renarchaeon, *Aeropyrum pernix* K1.** *DNA Res* 1999, **6**:83-101.

12. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH,
•• Hickey EK, Peterson JD, Nelson WC, Ketchum KA *et al.*: **Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*.** *Nature* 1999, **399**:323-329.

The authors report the observation that 81 of the 863 open reading frames identified by genome sequence analysis of *T. maritima*, a deep branching eubacterial species, are organized in 15 clusters that are most similar to genes from sequenced archaeal species. Gene order is conserved in these clusters, providing the most direct evidence for lateral gene transfer from archaeal to eubacterial species.

13. White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD,
•• Dodson RJ, Haft DH, Gwinn ML, Nelson WC, Richardson DL *et al*: **Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1.** *Science* 1999, **286**:1571-1577.

The authors' analysis of the gene content of the most radiation-resistant organism reveals the presence of all known genes for DNA repair, DNA damage export, and desiccation and starvation recovery; extensively redundant systems are also present in this organism. In addition, functional assignments could not be made for approximately half of the 3187 predicted genes, many of which will undoubtedly be shown to participate, via unknown mechanisms, in the stress-resistant characteristics of this organism.

14. Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, de Jonge BL *et al.*: **Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*.** *Nature* 1999, **397**:176-180.

15. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA *et al.*: **The complete genome sequence of the gastric pathogen *Helicobacter pylori*.** *Nature* 1997, **388**:539-547.

16. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q *et al.*: **Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*.** *Science* 1998, **282**:754-759.

17. Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW, Stephens RS: **Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*.** *Nat Genet* 1999, **21**:385-389.

18. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proc Natl Acad Sci USA* 1977, **74**:5088-5090.

19. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL,
• Wolf YI, Koonin EV: **Comparative genomics of the archaea (euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell.** *Genome Res* 1999, **9**:608-628.

The authors identify a set of core genes that are not observed to be laterally transferred across taxa. These are genes that code for genome replication and expression proteins.

20. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene
•• content.** *Nat Genet* 1999, **21**:108-110.

The authors demonstrate the use of the gene content of whole genomes to build phylogenetic trees from shared gene content. Such trees are remarkably similar to previously described trees built from rRNA similarity.

21. Fitz-Gibbon ST, House CH: **Whole genome-based phylogenetic
• analysis of free-living microorganisms.** *Nucleic Acids Res* 1999, **27**:4218-4222.

The authors obtained results similar to those described in [20••] using a parsimony analysis of the presence or absence of genes in sequenced microbial genomes.

22. Tekaia F, Lazcano A, Dujon B: **The genomic tree as revealed from whole proteome comparisons.** *Genome Res* 1999, **9**:550-557.

23. Pellegrine M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO:
•• **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.

The authors showed that genes that work in the same pathways frequently have correlated distribution patterns across species. Thus, the function of some unknown genes can be predicted on the basis of their having similar distribution patterns as genes with known functions.

24. Eisen JA, Hanawalt PC: **A phylogenomic study of DNA repair
• genes, proteins, and processes.** *Mutat Res* 1999, **435**:171-213.

The authors used a combination of evolutionary and genomic analysis, which they refer to as phylogenomics, to study DNA repair proceses.

25. Marcotte EN, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D:
• **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402**:83-86.

These authors, together with those of [26•], used protein domain patterns to cluster proteins that are functionally linked. The method involves using a two-domain protein in one species to cluster proteins in a second species that contain either one (or the other) domain, using the assumption that the two separated domains work together.

26. Enright AJ, Lliopoulos I, Kyrpides NC, Ouzounis CA: **Protein
• interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402**:86-90.

See annotation to [25•]

27. Richmond CS, Glasner JD, Mau R, Jin H, Blattner FR: **Genome-wide expression profiling in *Escherichia coli* K-12.** *Nucleic Acids Res* 1999, **27**:3821-3835.

28. Wilson M, DeRisi J, Kristensen HH, Imboden P, Rane S,
•• Brown PO, Schoolnik GK: **Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization.** *Proc Natl Acad Sci USA* 1999, **96**:12833-12838.

The authors used microarrays to explore gene expression levels in *M. tuberculosis* before and after exposure to the tuberculosis drug isoniazid. Genes within the drug target pathway are observed to be induced by the drug treatment, as are others of unknown function.

29. Liang F, Hole I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J: **Assessing the gene context of the human genome.** *Nat Genet* 2000, in press.