—Reviews—

Fruits of Human Genome Project and Private Venture, and Their Impact on Life Science

Akiko IKEKAWA* and Sumiko IKEKAWA
1–13–2, Mihamaku, Makuhari-Nishi, Chiba, 261–0026, Japan

(Received June 29, 2001)

A small knowledge base was created by organizing the Human Genome Project (HGP) and its related issues in ''*Science*'' magazines between 1996 and 2000. This base revealed the stunning achievement of HGP and a private venture and its impact on today's biology and life science. In the mid–1990, they encouraged the developement of advanced high throughput automated DNA sequencers and the technologies that can analyse all genes at once in a systematic fashion. Using these technologies, they completed the genome sequence of human and various other organisms. These fruits opened the door to comparative genomics, functional genomics, the interdisprinary field between computer and biology, and proteomics. They have caused a shift in biological investigation from studying single genes or proteins to studying all genes or proteins at once, and causing revolutional changes in traditional biology, drug discovery and therapy. They have expanded the range of potential drug targets and have facilitated a shift in drug discovery programs toward rational target-based strategies. They have spawned pharmacogenomics that could give rise to a new generation of highly effective drugs that treat causes, not just symptoms. They should also cause a migration from the traditional medications that are safe and effective for every members of the population to personalised medicine and personalised therapy.

Key words——knowledge base; *Science* magazine; Human Genome Project; completion of human genome sequencing, genomics and proteomics; drug discovery and therapy

In order to get a rough grasp of today's life science and its movement, a small knowledge base was created by organizing Human Genome Project (HGP) and its related issues in ''*Science*'' magazines between 1996 and 2000. ''*Science*'' magazines contain News of the Week, News Focus, Policy Forum, Perspectives, Research Articles, Reports, and so on, all full of cutting-edge information together with the topics cited from various magazines such as *Nature*. This review shows a rough grasp of the efforts of HGP and their impact on today's life science revealed from this small base.

The tool used for base creation is HyperCard (version 2.3, Mackintosh). The base has a tree structure of two branchs, which further ramify into twigs. The HGP Effort branch has 9 twigs (HGP, Private Venture, Sequencing technology, Gene technology, Gene mapping, Genome sequencing, Genetic variation, News, and Policy forum). The Revolution branch has 12 twigs (Comparative genomics, Functional genomics, Computer and biology, Global gene expression, Proteomics, Molecular biology, Evolution, Plant biology, Drug and therapy, Biodiversity, Population genetics, and Human history). The News, Policy forum, Population genetics and Human history twigs are excluded in this review.

1. Fruits of Human Genome Project and Private Venture

1–1. History of the Human Genome Project[1,2)]

Fifteen years ago, astronomers in the University of California (UC) were already angling to build the world's biggest telescope. Sinsheimer, a biologist who was then UC chancellor, was looking for a project of similar magnitude in biology, unraveling the sequence of the human genome. In May1985, Sinsheimer hosted a meeting at his university to discuss the feasibility of his ambitious proposal. After hot debates and captivation of such reseachers as those at the U.K. Medical Research Council (MRC) and the Office of Health and Environmental Research at the Department of Energy (DOE) in February 1988, a National Research Council (NRC) panel endorsed the Human Genome Project (HGP) unanimously, calling for a rapid scale-up in ''new and distinctive'' funds to $200 million a year over the next 15 years. The panel recommended that the project should begin by constructing maps of the human chromosomes and full-scale sequencing be postponed until new technologies made it faster and cheaper. It was also the panel's

本総説は，平成13年度昭和大学薬学部を退官にあたり，平成7年から平成12年の間に行った研究を中心に記述されたものである．

recommendation to analyze the genomes of simple organisms, such as *Escherichia coli*, yeast, the roundworm, and eventually the mouse. In March 1988, then-the National Institutes of Health (NIH) director Wyngaarden announced that NIH would create a special office for genome research. In September 1988, he nabbed Watson to head it, and with that coup, NIH was firmly ensconced as the lead agency. It has remained so, even as the project gathered international collaborators and Britain's Wellcome Trust took on a prominent role.

Watson insisted that the goal at the first stage of the project was building maps of the human chromosomes, and claimed that they should not focus on finding the genes. Progress was rapid. By 1990, Sulston and colleagues had nearly completed the physical map of the worm—changing worm biology forever—and Olson and colleagues were proceeding apace on yeast. Faster and easier ways to clone and map DNA were coming on line, and sequencing trials were beginning.

That newfound harmony was shattered in June 1991. Venter and his colleague Adams had developed a new technique, called expressed sequence tags (ESTs), that enabled them to find genes at unprecedented speed. Never one of Watson's inner circle, Venter claimed he could find 80% to 90% of the genes within a few years, for a fraction of the cost. Venter left NIH in 1991 when he was offered $70 million from a venture capital company to try out his gene identification strategy at a new nonprofit, The Institute for Genomic Research (TIGR).

After Watson's sudden departure in April 1992, NIH picked gene hunter Collins of the University of Michigan to take the helm. It was a heyday for gene hunters. The early investments in the genome project paid off as increasingly sophisticated maps of the human and mouse genomes were compiled. With these maps in hand, the time it took to track down most disease genes dropped from a decade to perhaps 2 years. Every week, it seemed, another deadly disease gene was discovered. The consortium was growing as well, fueled by an infusion of funds from the Wellcome Trust, which in 1993 set up a major new sequencing lab, the Sanger Centre near Cambridge, with Sulston as its head.

But sequencing overall was lagging behind. At the existing rate and cost, Collins lamented when he took on the job, there was no chance they could finish the sequencing by 2005. Steady, incremental advances were enabling scientists to spew out longer ''sequence reads,'' and the cost was slowly dropping. Even so, reassembling the DNA fragments in correct order was tricky. In September1995, the Japanese government funds several sequencing groups for a total of $15.9 million over 5 years: Tokai University, University of Tokyo, and Keio University.

In 1995, Venter surprised the community by announcing that he and his colleagues had sequenced the first entire genome of a free-living organism, *Haemophilus influenzae*, at 1.8 megabases. What's more, they had done it in just a year using a bold new approach, whole-genome shotgun sequencing, that NIH had insisted wouldn't work and wouldn't fund. Sequencers in the publicly funded project had adopted a conservative, methodical approach—starting with relatively small chunks of DNA whose positions on the chromosome were known, breaking them into pieces, then randomly selecting and sequencing those pieces and finally reassembling them. Eventually, larger pieces called contigs would be hooked together. By contrast, Venter simply shredded the entire genome into small fragments and used a computer to reassemble the sequenced pieces by looking for overlapping ends.

On 9 May 1998, Venter announced that he had teamed up with Perkin-Elmer Corp., which was about to unveil an advanced, automated sequencing machine, to create a new company (Celera Genomics) that would single-handedly sequence the entire human genome in just 3 years—and for a mere $300 million using the whole genome shotgun method. What's more, said Venter, when he was done he would give the data away free to the community by posting it on his company's Web site. This time, he had 300 of Perkin-Elmer's sequencing machines. And to reassemble his sequenced fragments, Venter would use one of the world's fastest supercomputers.[3)]

Collins announced new goals for the public project in September 1998, just 4 months after Venter's surprise announcement. First, the consortium would complete the entire genome by 2003. And, in a dramatic departure from previous philosophy of 99.99% accuracy, the project would produce a ''rough draft,'' covering 90% of the genome, by the spring of 2001.[4)] In March1999, NIH again moved up the completion date for the rough draft, to spring 2000. Large-scale sequencing efforts are concentrated

in centers at Whitehead, Washington University, Baylor, Sanger, and DOE's Joint Genome Institute.[5]

In a crucial test of the shotgun strategy, Celera first tackled the 180-megabase genome of the fruit fly *Drosophila melanogaster*. Venter teamed up with a publicly funded team headed by Gerald Rubin of UC Berkeley, and by March 2000, they had pulled it off. This proved that the shotgun methods could work on a big, complex genome.[6]

The race was on, punctuated by dueling press releases. Venter announced in January 2000 that his crew had compiled DNA sequence covering 90% of the human genome, the public consortium asserted in March 2000 that it had completed 2 billion bases, and so on.[7] But, at a White House ceremony in June 2000, President Clinton lauded both scientists for their phenomenal achievement, and Collins and Venter lavished praise on one another.[8] In February 2001, The HGP consortium published its working draft in *Nature* (15 February), and Celera published its draft in *Science* (16 February).[2]

1-2. Genome Technology Developement In 1995, array technologies, that can analyse hundreds of genes simultaneously, was developed.[2,9] In 1997, automated high-throughput DNA sequencing mashine was developed.[2,10] These technologies accelerated complete genome sequencing of human and other organisms. These technologies also provided drug researchers entire sets of potential drug targets. As they discover more targets and create more potential drugs or leads to affect these targets, the number of samples they need to process overwhelms traditional approaches. So assay minituarization and high throughput screening systems consisting of automated instruments has become critical, in order to improve the efficiency and productivity, and to reduce the cost. The history of these technology developement is summarized below.

1-2-1. *Technology for DNA Analysis*

(2001 February 16 L. Roberts et al.)[2]

1984 May

Cantor and Schwartz of Columbia University develop pulsed field electrophoresis (*Cell*).

1985 December

Mullis and colleagues at Cetus Corp. develop PCR, a technique to replicate vast amounts of DNA.

1987 May

Burke, Olson, and Carle of Washington University in St. Louis develop YACs for cloning, increasing insert size 10-fold.

1990

Three groups develop capillary electrophoresis, one team led by Smith (*Nucleic Acids Research*, August), the second by Karger (*Analytical Chemistry*, January), and the third by Dovichi (*Journal of Chromatography*, September).

1995 May to August

Mathies and colleagues at UC Berkeley and Amersham develop improved sequencing dyes (*PNAS*, May); Reeve and Fuller at Amersham develop thermostable polymerase (*Nature*, August).

1995 October

Brown of Stanford and colleagues publish first paper using a printed glass microarray of complementary DNA (cDNA) probes.

1996 April

Affymetrix makes DNA chips commercially available.

1996 October 25 S. P. A. Fodor et al.[9]

DNA arrays containing up to 135,000 probes complementary to the 16.6-kilobase human mitochondrial genome by light-directed chemical synthesis. A two-color labeling scheme allows simultaneous comparison of a polymorphic target to a reference DNA or RNA.

(2001 February 16 L. Roberts et al.)[2]

1996 November

Hayashizaki's group at RIKEN completes the first set of full-length mouse cDNAs.

1997 July 18 S. P. A. Fodor et al.[9]

Affymetrix develops a DNA chip, consisting of a highly dense array of complementary probes, integrating light-directed combinatorial chemistry and laser confocal fluorescence scanning.

1998 February 20 L. G. Kostrikis et al.[11]

Kostrikis et al. develop an automated method for detecting mutations, called "spectral genotyping," in which alleles are identified by fluorescent colors generated in sealed amplification tubes, using molecular beacons.

1998 March 27 J. Alper[12]

A mass spectrometer vaporizes the DNA and accelerates the molecules through a vacuum chamber with the help of an electric field. MALDI-TOF, matrix-assisted laser desorption / ionization-time-of-flight mass spectrometry, can

analyze hundreds of DNA samples in a matter of a few minutes.

1999 October 15 F. S. Collins et al.[13]

The Mammalian Gene Collection (MGC) project is a new effort by the NIH to generate full-length complementary DNA (cDNA) resources. This project will provide publicly accessible resources to the full research community. The MGC project entails the production of libraries, sequencing, and database and repository development, as well as the support of library construction, sequencing, and analytic technologies dedicated to the goal of obtaining a full set of human and other mammalian full-length (open reading frame) sequences and clones of expressed genes.

2000 April 21 R. F. Service[14]

Mandecki develop microtransponders, essentially tiny silicon chip-based devices each just a few hundred micrometers on a side. Each transponder stores an ID number in memory and when prompted emits an identifying radio signal. They attach the probes to their silicon ID tags (the transponders). Next, they mix it into a solution containing the radio-tagged oligos. The researchers then simply flow their solution through a device. As the transponders and their genetic cargo stream through a narrow channel, the researchers shine a laser light on them. If a flash of light accompanies the radio signal, the gene is actively expressed.

2000 September 8 T. A. Taton, C. A. Mirkin, R.t L. Letsinger[15]

Scanometric DNA Array Detection with Nanoparticle Probes; A method for analyzing combinatorial DNA arrays using oligonucleotide-modified gold nanoparticle probes and a conventional flatbed scanner.

1–2–2. *Sequencing Technology*

(2001 February 16 L. Roberts et al.)[2]

1977

Maxam and Gilbert at Harvard University and Sanger at MRC independently develop methods for sequencing DNA (*PNAS*, February; *PNAS*, December).

1980 May

Botstein of the Massachusetts Institute of Technology (MIT), Davis of Stanford University, and Skolnick and White of the University of Utah propose a method to map the entire human genome based on RFLPs (*American Journal of Human Genetics*).

1982

Wada, at RIKEN in Japan, proposes automated sequencing and gets support to build robots with help from Hitachi.

1986 June

Hood and Smith of the California Institute of Technology (Caltech) and colleagues announce the first automated DNA sequencing machine (*Nature*).

1987 October

DuPont scientists develop a system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Applied Biosystems Inc. puts the first automated sequencing machine, based on Hood's technology, on the market.

1990 October

Lipman, Myers, and colleagues at the National Center for Biotechnology Information (NCBI) publish the BLAST algorithm for aligning sequences (*Journal of Molecular Biology*).

1991 December

Uberbacher of Oak Ridge National Laboratory in Tennessee develops GRAIL, the first of many gene-finding programs (*PNAS*).

1997 June 20 T. L. Hawkins et al.[10]

Hawkins et al. design an automated sequencing system to meet the laboratory throughput needs of HGP. With an articulated arm at the center of a set of modules for liquid handling, thermocycling, shaking, and storage, all coordinated by scheduling software, They are able to create a generic automation platform. They develop a procedure called solid-phase reversible immobilization (SPRI). Under certain conditions, DNA can be tightly bound to the surface of carboxyl-coated magnetic particles, extensively washed, and subsequently released back into solution. Then. they design the Sequatron, an automated, adaptable system for high-throughput genomics. The term Sequatron actually describes a generic platform with an articulated robotic arm, centralized control, and scheduling software.

1997 September 5 X. Michalet et al.[16]

DNA in amounts representative of hundreds of eukaryotic genomes is extended on silanized sur-

faces by dynamic molecular combing. Fluorescent hybridization of DNA probes on combed DNA allows direct mapping of their respective positions along the fibers.

(2001 February 16 L. Roberts et al.)[2]

1997 September

Molecular Dynamics introduces the MegaBACE, a capillary sequencing machine.

1998 March

Green and Ewing of Washington University and colleagues publish a program called phred for automatically interpreting sequencer data (*Genetic Research*). Both phred and its sister program phrap (used for assembling sequences) had been in wide use since 1995.

1998 May

PE Biosystems Inc. introduces the PE Prism 3700 capillary sequencing machine.

1998 May 8 E. Pennisi[17]

A new computer program that analyzes machine readouts is helping sequencing labs to scale up their efforts, at the final stage in which fragments of raw DNA data are arranged into a completed sequence.

1998 May 15 R. F. Service[18]

The gene sequencing machines rely on capillary electrophoresis, but add a high level of automation.

1998 July 17 M. Ronaghi, M. Uhlen, and P. Nyren[19]

A new DNA sequencing method is developed. Four nucleotides are added stepwise to the template hybridized to a primer. The PPi released in the DNA polymerase-catalyzed reaction is detected by the ATP sulfurylase and luciferase in a coupled reaction. As this procedure is repeated, longer stretches of the template sequence are deduced. An automated instrument has recently been developed.

1999 March 19 J. C. Mullikin and A. A. McMurray[20]

The latest automated high-throughput DNA sequencing machines;

Capillary tube apparatus

ABI Prism 3700 DNA Analyzer from Perkin-Elmer.

Molecular Dynamics MegaBACE 1000 from Amersham Pharmacia Biotech launched in 1998

Traditional Slab-shaped gel apparatus

ABI 377XL–96 slab gel sequencer

1–2–3. *Miniatuarization*

1998 October16 R. F. Service[21]

Coming Soon: The Pocket DNA Sequencer

Researchers and companies are working to shrink to pocket size all types of chemistry equipment, using microchips as ''microfluidics''.

1998 October 16 M. A. Burns et al.[22]

A device is developed that uses microfabricated fluidic channels, heaters, temperature sensors, and fluorescence detectors to analyze nanoliter-size DNA samples.

1999 January 15 B. H. Weigl and P. Yager[23] (TechView)

Microfluidic diffusion-based separation and detection

1999 April 16 P. Belgrader et al.[24]

A portable, real-time PCR instrument, consisting of an array of microfabricated silicon reaction chambers with integrated optical detectors, analyzes samples of 5 to 500 bacteria cells in as little as 7 minutes.

1999 October 15 J. Rogers[25]

The requirement for inexpensive, very accurate high throughput sequencing is even more crucial.

New instrumentation for sequence generation

Miniaturization of sequencing platforms

1999 December 10 C. A. Mirkin (Review)

P. Kim, C. M. Lieber

Nanoscale electromechanical systems, nanotweezers, based on carbon nanotubes have been developed for manipulation and interrogation of nanostructures.[26]

1–3. Efforts in Gene Mapping and Genome Sequencing The efforts of gene mapping and genome sequencing are summarized below.

1–3–1. *Gene Mapping*

(2001 February 16 L. Roberts et al.)[2]

1987 April

An advisory panel suggests that DOE should spend $1 billion on mapping and sequencing the human genome over the next 7 years and that DOE should lead the U.S. effort. DOE's Human Genome Initiative begins.

1987 October

Donis-Keller and colleagues at Collaborative Research Inc. publish the ''first'' genetic map with 403 markers, sparking a fight over credit and priority (*Cell*).

1989 September
Olson, Hood, Botstein, and Cantor outline a new mapping strategy, using sequence tagged sites (STSs).
1990 April
NIH and DOE publish a 5-year plan. Goals include a complete genetic map, a physical map with markers every 100 kb, and sequencing of an aggregate of 20 Mb of DNA in model organisms by 2005.
1992 October
U.S. and French teams complete genetic maps of mouse and human: mouse, average marker spacing 4.3 cM, Lander and colleagues at Whitehead (*Genetics, June*); human, average marker spacing 5 cM, Weissenbach and colleagues at CEPH (*Nature*, October).
1994 September
Murray of the University of Iowa, Cohen of Genethon, and colleagues publish a complete genetic linkage map of the human genome, with an average marker spacing of 0.7 cM.
1995 December
Researchers at Whitehead and Genethon (led by Lander and Hudson at Whitehead) publish a physical map of the human genome containing 15,000 markers.

1999 September 3 J. C. Venter et al.[27]
A whole-genome restriction map of *Deinococcus radiodurans*, a radiation-resistant bacterium able to survive up to 15,000 grays of ionizing radiation, is constructed with whole-genome shotgun optical mapping.

1999 November 12 E. Pennisi[28]
Two new maps of *Plasmodium falciparum* are produced this week. The first map is produced by Wellems, a malaria expert at the National Institute of Allergy and Infectious Diseases (NIAID) and his colleagues. The second map is produced by Schwartz of the University of Wisconsin, Madison, and his colleagues, by optical mapping. (*Nature Genetics*, November)

1999 November 12 Xin-zhuan Su et al.[29]
A genome-wide, high-resolution linkage map of the human Malaria parasite Plasmodium falciparum is produced.

2000 March 24 R. A. Hoskins et al.[30]
Hoskins et al. construct a bacterial artificial chromosome (BAC)-based physical map of chromosomes 2 and 3 of *Drosophila melanogaster*, which constitute 81% of the genome.

1-3-2. *Genome Sequencing*
Early in 1998, genome sequencing seemed daunting, 97% of the human genome remaining to be deciphered. Scince late in 1998, the sequencing rate increased dramatically due to the developement of advanced automated high-throughput DNA sequencer.[31,32]

1-3-2-1. *Human Genome Sequencing*

1998 May 8 E. Pennisi[31]
97% of the genome remains to be deciphered. Genome researchers aim to automate every step of DNA sequencing. NHGRI has only 40 million high-quality bases to show for the $52 million spent thus far.

1998 October 2 R. Waterston and J. E. Sulston[33]
In only 2 years, the world total of finished human genomic sequence has gone from 15 to 180 Mb (0.5 to 6%).

1999 September 24 D. Normile and E. Pennisi[34]
Within the next week or two, sequencing human genome chromosome 22 will be completed by the chromosome 22 consortium composed of British, Japanese, and U.S. researchers.
(2001 February 16 Leslie Roberts et al.)[2]
In December, they completed the first sequence of a human chromosome, number 22. (*Nature*, 2 December 1999)

2000 March 31 E. Marshall, E. Pennishi, and L. Roberts[9]
Both teams (HGP team and a privately funded team at Celera Genomics) are racing to complete a draft of the human genome in the next few months.

2000 April 21 E. Pennisi[35]
On 13 April, DOE has finished the working drafts of human chromosomes 5, 16, and 19.

2000 May 12 E. Pennisi[36]
On May 8, HGP anounces the completion of the rough draft of the human genome. (85% of the promised 90% available in Genbank) HGP consortium led by German and Japanese researchers publishes the complete sequence of chromosome 21 (*Nature*).

2000 June 30 E. Marshall[8]
On 26 June. at a White House ceremony, HGP and Celera jointly announce working drafts of the human genome sequence, declare their feud

at an end, and promise simultaneous publication. (Celera; 99% completed)

2001 February 16 E. Pennisi[37]

The HGP consortium publishes its working draft in *Nature* (15 February), and Celera publishes its draft in *Science* (16 February).

1–3–2–2. *Genome Sequencing of Other Organisms*

(2001 February 16 L. Roberts et al.)[2]

1984 July

MRC scientists decipher the complete DNA sequence of the Epstein-Barr virus, 170 kb (*Nature*).

1990 August

NIH begins large-scale sequencing trials on four model organisms:

Mycoplasma capricolum, Escherichia coli, Caenorhabditis elegans, and *Saccharomyces cerevisiae*. Each research group agrees to sequence 3 Mb at 75 cents a base within 3 years.

1991 October

The Japanese rice genome sequencing effort begins.

1995 July

Venter and Fraser of TIGR and Smith of Johns Hopkins publish the first sequence of a free-living organism, *Haemophilus influenzae*, 1.8 Mb (*Science*).

1996 October 25 A. Goffeau et al.[38]

An international consortium publicly releases the complete genome sequence of the yeast *S. cerevisiae*.

(2001 February 16 L. Roberts et al.)[2]

1997 September

Blattner, Plunkett, and University of Wisconsin colleagues complete the DNA sequence of *E. coli*, 5 Mb (*Science*).

1998 February

Representatives of Japan, the U.S., the E.U., China, and South Korea meet in Tsukuba, Japan, to establish guidelines for an international collaboration to sequence the rice genome.

1998 July 17 J. C. Venter et al.[39]

The genome of *Treponema pallidum* is sequenced by the whole genome random shotgun method.

1998 October 23 D. W. Meinke et al.[40]

Arabidopsis thaliana: The entire genome is scheduled to be sequenced by the end of the year 2000.

1998 November 6 J. C. Venter et al.[41]

Chromosome 2 of the Human Malaria Parasite *Plasmodium falciparum* is sequenced with the shotgun sequencing approach.

1998 December 11 The *C. elegans* Sequencing Consortium[42]

Sulston of the Sanger Centre and Waterston of Washington University and colleagues complete the genomic sequence of *C. elegans*.

1999 February 26 E. Pennisi[43]

Barrell, Parkhill et al. sequence the genome of a food-borne pathogen, the bacterium *Campylobacter jejuni*.

1999 July 16 C. Somerville, S. Somerville[44]

Nucleotide sequencing of the *Arabidopsis* genome is nearing completion. Sequencing of the rice genome has begun.

(2001 February 16 L. Roberts et al.)[2]

1999 September

NIH launches a project to sequence the mouse genome, devoting $130 million over 3 years.

1999 October 8 E. Pennisi[45]

HGP; producing a preliminary sequence of the mouse genome sequence by 2003, followed by a high-quality version by 2005.

1999 November 19 J. C. Venter et al.[46]

The genome of the radiation-resistant bacterium *Deinococcus radiodurans* R1 is sequenced by the whole-genome shotgun method.

2000 February 18 E. Pennisi[47]

The genome community; A good part of the mouse will be sequenced using the whole-genome shotgun method.

2000 March 3 E. Pennisi[48]

TIGR has just finished sequencing the *Caulobacter crescentus* genome.

2000 March 10 J. C. Venter et al.[49]

The genome of *Neisseria meningitidis* Serogroup B Strain MC58 is sequenced by the whole genome random shotgun method.

2000 March 24 J. C. Venter et al.[6]

Celera and academic collaborators sequence the 180–Mb genome of the fruit fly *Drosophila melanogaster*, the largest genome yet sequenced and a validation of Venter's controversial whole-genome shotgun method.

2000 April 14 E. Pennisi[50]

A rough draft of the entire rice genome is

prodeced by Monsanto and collaborators. By this winter, about 10% of the rice genome has been mapped by an international consolitium.

2000 May 5 M. Hagmann[51]

A 5-year sequencing effort of growing *Mycobacterium leprae* in a armadillo will begin by a team at the Institut Pasteur in Paris, in collaboration with the Sanger Centre in Cambridge, U.K..

Xylella fastidiosa, the first bacterial plant pathogen, is sequenced on 12 April by a consortium of some 30 labs in San Pauro state.

2000 September 1 V. J. DiRita[52]

Two circular chromosomes of *Vibrio cholerae* are sequenced by Heidelberg et al.. (*Nature*)

(2001 February 16 L. Roberts et al.)[2]

2000 October

DOE and MRC launch a collaborative project to sequence the genome of the puffer fish, *Fugu rubripes*, by March 2001.

2000 December

An international consortium completes the sequencing of the first plant, *Arabidopsis thaliana*, 125 Mb.

1-4. Genetic Variation Genetic variation provide powerful tools for a variety of medical genetic studies. Single-nucleotide polymorphisms (SNPs) are the most frequent type of variation in the human genome.[53] SNPs in genes or control regions may influence susceptibility to common diseases. Others probably have no function but could provide valuable markers for gene hunters. SNPs will be used as analytical tools, making it easier to trace inherited disease risks and abnormal responses to drugs.[54] Few SNPs are likely to be directly involved in disease, but a database of several hundred thousands will make it easier to track smaller segments of the genome and identify patterns of inheritance that affect health. The SNP map will make it possible to diagnose illnesses earlier and avoid giving drugs to patients likely to experience side effects, including drugs already in use.

1-4-1. *Projects creating SNPs Databases*

When genomics companies like Genset, Incyte, and Celera began building private SNP databases in the late 1990s, the National Human Genome Research Institute (NHGRI), worried that academic researchers would be shut out of the field, launched a new project to create a database of 60,000 to 160,000 SNPs in 1998.[54,55] In April 1999, a consortium of pharmaceutical giants and Britain's Wellcome Trust launched a novel venture: They would put together a database of 300,000 SNPs and give them away. These fiercely competitive companies are teaming up to bankroll work by a network of academic labs. The nonprofit SNP Consortium, or TSC (its official name) isn't altruistic, though. The companies backing the enterprise expect that SNPs will enable them to develop and sell drugs more effectively. And by creating a public database, they will avoid having to buy multiple, private data collections from the half-dozen or so biotech firms that have been collecting SNPs since 1997, hoping to stake a proprietary claim on the data. They are to find 300,000 SNPs in 2 years. Once found, the SNPs will be tracked to positions on the genome. The goal is to have 150,000 mapped in this way by mid-2001.[55]

The SNPs have been widely touted as the key to personalized medicine, with drugs tailored to an individual's genotype and simple tests to determine one's risk of specific diseases. But a closed meeting held in March 2000, sponsored by the SNP Consortium and NHGRI, concluded that those promises may be harder to achieve than expected, and that more SNPs may be required to track down a particular disease gene than previously estimated. In general, the more markers on a map, the easier it is to find genes. But with the cost of identifying each SNP hovering at about $100, there's a big incentive to use the fewest possible. Kruglyak published a paper in *Nature Genetics* in summer in 1999 asserting that 500,000 or even 1 million SNPs would be needed to track down susceptibility genes. Another confounding factor is that the usefulness of any one SNP varies enormously from population to population. Just one-third of the SNPs found so far seem to be widely applicable in all populations. Investigators who want to study a particular ethnic group will have to find more SNPs. Factoring in all these complexities, Collins left the meeting thinking that 600,000 to 1 million SNPs would be ideal. But, Collins added, considerable progress could still be made with a smaller set of markers.[54]

1-4-2. *Basic Research*

Wang et al. carried out large-scale identification, mapping, and genotyping of SNPs in the human genome, by a combination of gel-based sequencing and high-density variation-detection DNA chips. A total of 3241 candidate SNPs were identified. A genetic map was constructed showing the location of 2227 of these SNPs. Prototype genotyping chips were deve-

loped that allow simultaneous genotyping of 500 SNPs. The results provide a characterization of human diversity at the nucleotide level and demonstrate the feasibility of large-scale identification of human SNPs.[53]

A total of 3714 biallelic markers, spaced about every 3.5 kilobases, were identified by analyzing the patterns obtained when total genomic DNA from two different strains of yeast was hybridized to high-density oligonucleotide arrays. Because the extent of hybridization of a target sequence to an oligonucleotide probe depends on the number and position of mismatches between the two sequences, Winzeler et al. hypothesized that a substantial fraction of the allelic variation between any two strains of yeast could be detected simply by hybridizing genomic DNA from the two strains to the arrays and analyzing the hybridization differences.The markers were used to simultaneously map a multidrug-resistance locus and four other loci with high resolution.[56]

Progressive damage to mitochondrial DNA (mtDNA) during life is thought to contribute to aging processes. Here, Michikawa et al. revealed high copy point mutations at specific positions in the control region for replication of human fibroblast mtDNA from normal old, but not young, individuals. Furthermore, in longitudinal studies, one or more mutations appeared in an individual only at an advanced age. Some mutations appeared in more than one individual. Most strikingly, a T414G transversion was found, in a generally high proportion (up to 50 percent) of mtDNA molecules, in 8 of 14 individuals above 65 years of age (57 percent) but was absent in 13 younger individuals.[57]

Despite its high prevalence, very little is known regarding genetic predisposition to prostate cancer. A genome-wide scan performed in 66 high-risk prostate cancer families has provided evidence of linkage to the long arm of chromosome 1 (1q24–25).[58]

Schizophrenia is a complex disorder, and there is substantial evidence supporting a genetic etiology. Brzustowicz et al. carried out a genome-wide scan for schizophrenia susceptibility loci in 22 extended families with high rates of schizophrenia, and provided highly significant evidence of linkage to chromosome 1(1q21–q22). This linkage result should provide sufficient power to allow the positional cloning of the underlying susceptibility gene.[59]

1–4–3. *Application to Forensics*[60]

Forensic scientists are equipping police investigators with powerful tools for collecting and analyzing evidence. DNA profiling is a powerful technique for gauging the likelihood that a biological sample, such as blood or semen, came from a specific individual. The use of profiling to identify suspects was pioneered by geneticist Jeffreys of the University of Leicester, U.K. His approach, called multilocus profiling, used restriction enzymes to cleave the DNA at specific sites. The resulting fragments differ in size from person to person.

Modern profiling relies on short tandem repeat (STR) analysis, which debuted in the forensics world in 1994. This technique looks at specific areas of DNA molecules containing simple blocks of base pairs; the blocks are repeated end to end. The number of occurrences of each block, or repeat unit, varies by individual. Examining several DNA regions, or loci, and counting the number of repeated units in each area generates numbers that form a molecular label of the DNA's owner. This gives the DNA profile as a numerical tag for easy database comparison. The latest variation on the technique, introduced in Britain in mid–1999, targets 10 DNA loci—enough to guarantee that the odds of someone else sharing the same result are slimmer than one in a billion. In the United States, the FBI routinely examines 13 STR sites, and the chance of two unrelated individuals on average having the same DNA profile is about one in a million billion.

Geneticists can assess the likelihood that a person is a redhead simply by testing for mutations in the gene for the receptor for a hormone that spurs production of the pigment melanin. Ethnicity can be inferred from the frequencies of alternative forms, or alleles, of genes; allele patterns differ by racial origin.

Although it would require about 50 SNP sites to achieve the same level of confidence as with STR analysis, SNP analysis is easy in miniaturization, which means faster processing. There's probably 10,000 times as much mitochondrial DNA as there is nuclear DNA. Although mitochondrial DNA can't conclusively link an individual to a crime, it can point a finger at a family. The value of establishing family ties is not lost on missing-persons investigators. Many people who have disappeared, or died, or whatever, didn't bother to leave a nice, clean sample of their DNA. Tully helped develop a technique called minisequencing; looking at 12 of the bases that are most

likely to differ between individuals. This approach can slash the time it takes to get results from a batch of mitochondrial DNA samples from 3 months to 3 weeks.

In 1995, the United Kingdom became the first country to create a national DNA database. Based on the 10–loci STR approach, it now holds about 800,000 profiles of people suspected or convicted of an imprisonable offense. It is a major force in police intelligence, with about 600 ''hits'' every week from samples collected at crime scenes. Other countries with DNA databases—including Austria, Germany, the Netherlands, and New Zealand—are typically more restrictive about the circumstances under which DNA samples can be drawn from individuals. In the United States, each state has passed laws that allow a DNA sample to be collected from individuals convicted of rape or other sex crimes, and many also permit sampling from people convicted of murder, burglary, or even certain misdemeanors. The U.S. national DNA database system, called CODIS, or combined DNA index system, started up in October 1998. It collates data from all the state databases. There are maybe 100,000 profiles in CODIS now, but a whopping 750,000 blood samples from convicted felons have yet to be profiled and those profiles incorporated into CODIS.

There are too many privacy issues involved. You have to have a balance between the privacy of the citizen and the needs of the state. As forensic science grows more sophisticated, that balancing act will be ever harder to maintain.

2. Emerging New Fields

Complete genome sequencing of human and other organisms opened the door to large-scale comparative studies. Another response to the challenge of massive amounts of sequence data was the development of ''functional genomics'', ''bioinformatics'', and ''proteomics'' technologies in the mid–1990s. Functional genomics refers to the development and application of global (genome-wide or system-wide) experimental approaches to assess gene function or activity by making use of the information and reagents provided by genome mapping and sequencing projects. It is characterized by high-throughput or large-scale experimental methodologies, and the fundamental strategy is to expand the scope of biological investigation from studying single genes or proteins to studying all genes or proteins at once in a systematic fashion. In contrast to the intimate details of function that traditional biological disciplines provide, functional genomics and proteomics produce much broader but shallower information about large numbers of genes and proteins.[32,61]

2–1. Comparative Genomics

2–1–1. *Comparisons between Distantly Related Genomes*

About 26,000 to 38,000 genes are expected to be found in the draft version of our own genome,[62] a number that is only two to three times larger than the 13,600 genes in the fruit fly genome. Furthermore, some 10% of human genes are clearly related to particular genes in the fly and the worm. So, obviously, we share much of our genetic scaffold even with very distant relatives. The similarity between humans and other animals will become even more evident when genome sequences from organisms such as the mouse become available. For these species, both the number of genes and the general structure of the genome are likely to be very similar to ours.[63] DNA sequence comparison and comparative mapping are essential for identification of gene orthologs in distantly related species (genes in different species that are descended from a single gene of a common ancestor).

Comparative analysis of predicted protein sequences encoded by the genomes of *Caenorhabditis elegans* and *Saccharomyces cerevisiae* suggests that most of the core biological functions are carried out by orthologous proteins that occur in comparable numbers. The specialized processes of signal transduction and regulatory control that are unique to the multicellular worm appear to use novel proteins, many of which re-use conserved domains. Major expansion of the number of some of these domains seen in the worm may have contributed to the advent of multicellularity. The proteins conserved in yeast and worm are likely to have orthologs throughout eukaryotes; in contrast, the proteins unique to the worm may well define metazoans.[64]

Neurotransmitter receptors, neurotransmitter synthesis and release pathways, and heterotrimeric GTP–binding protein (G protein)-coupled second messenger pathways are highly conserved between *C. elegans* and mammals, but gap junctions and chemosensory receptors have independent origins in vertebrates and nematodes. Im *C. elegans*, most ion channels are similar to vertebrate channels but there

are no predicted voltage-activated sodium channels.[65]

The *C. elegans* genome sequence was surveyed for transcription factor and signaling gene families that have been shown to regulate development in a variety of species. In most of the gene families already have been genetically analyzed in *C. elegans*, about half of the genes detect probable orthologs in other species, and about 10 to 25 percent of the genes are, at present, unique to *C. elegans*. *C. elegans* is also missing genes that are found in vertebrates and other invertebrates.[66]

More than 3 percent of the protein sequences inferred from the *C. elegans* genome contain sequence motifs characteristic of zinc-binding structural domains, and of these more than half are believed to be sequence-specific DNA-binding proteins. The distribution of these zinc-binding domains among the genomes of various organisms offers insights into the role of zinc-binding proteins in evolution. In addition, the complete genome sequence of *C. elegans* provides an opportunity to analyze, and perhaps predict, pathways of transcriptional regulation.[67]

Dense genetic maps of human, mouse, and rat genomes that are based on coding genes and on microsatellite and SNP markers have been complemented by precise gene homolog alignment with moderate-resolution maps of livestock, companion animals, and additional mammal species. Comparative genetic assessment expands the utility of these maps in gene discovery, in functional genomics, and in tracking the evolutionary forces that sculpted the genome organization of modern mammalian species. Over half of 70,000 to 100,000 human ESTs for RNA transcripts are already mapped. Nearly every human genes has a mouse homolog.[68]

Improved technologies and the potential for valuable applications have put the prospect of dense gene maps of domesticated livestock and companion animal species within our reach. Some immediate practical applications of these maps that we envision include:

(i) supplying animal models for human genetic diseases based on explicit gene homology as monitors for pathogenesis and therapy;

(ii) an opportunity to identify candidate polygenes that affect human and veterinary disease;

Heritable canine maladies and feline genetic deseases have been attributed to genes homologous to human desease gene mutations.[68]

Venter et al. undertook a comparative analysis of the genomes of *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*—and the proteins they are predicted to encode— in the context of cellular, developmental, and evolutionary processes. The nonredundant protein sets of flies and worms are similar in size and are only twice that of yeast, but different gene families are expanded in each genome, and the multidomain proteins and signaling pathways of the fly and worm are far more complex than those of yeast. The conservation of biological processes from flies to mammals extends the influence of *Drosophila* to human health. The power of *Drosophila* genetics has been leveraged to elucidate mammalian pathways involved in cancer biology, the cell cycle, and receptor tyrosine kinase signaling. The fly has orthologs to 177 of the 289 human disease genes examined and provides the foundation for rapid analysis of some of the basic processes involved in human disease.[69]

Long-range regulatory elements are difficult to discover experimentally; however, they tend to be conserved among mammals, suggesting that cross-species sequence comparisons should identify them. To search for regulatory sequences, Loots et al. examined about 1 megabase of orthologous human and mouse sequences for conserved noncoding elements with greater than or equal to 70% identity over at least 100 base pairs. Ninety noncoding sequences meeting these criteria were discovered, and the analysis of 15 of these elements found that about 70% were conserved across mammals. Characterization of the largest element in yeast artificial chromosome transgenic mice revealed it to be a coordinate regulator of three genes, interleukin−4, 13, and 5, spread over 120 kilobases.[70]

Genomic data are helping to understand both ancient evolution and the relationships among modern species. They revealed a possible whale-hippo link and a new genetic analysis that indicates that there are fewer subspecies of pumas than previously thought.[71]

2−1−2. *Comparisons between Closely Related Species*

We already know that the overall DNA sequence similarity between humans and chimpanzees is about 99%. When the chimpanzee genome sequence becomes available, we are sure to find that its gene content and organization are very similar (if not iden-

tical) to our own. Yet the few differences between our genome and those of the great apes will be profoundly interesting because among them lie the genetic prerequisites that make us different from all other animals. In particular, these differences may reveal the genetic foundation for our rapid cultural evolution and geographic expansion, which started between 150,000 and 50,000 years ago and led to our current overbearing domination of Earth. A chimpangee genome project is now developing to resolve the differences between humans and closest nonhuman relatives., whereas other primate gene maps (baboon and macaque) have been initiated to apply genetic assesment to these animals for behavior, vaccine developement, and genetic deseases.[63]

Although data on nucleotide sequence variation in the human nuclear genome have begun to accumulate, little is known about genomic diversity in chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*). Kaessmann, Wiebe, and Paabo reported a 10,154–base pair sequence on the chimpanzee X chromosome, representing all major subspecies and bonobos. Comparison to humans shows the diversity of the chimpanzee sequences to be almost four times as high and the age of the most recent common ancestor three times as great as the corresponding values of humans. Phylogenetic analyses show the sequences from the different chimpanzee subspecies to be intermixed and the distance between some chimpanzee sequences to be greater than the distance between them and the bonobo sequences.[72]

2–1–3. *Comparisons between Species*

Behr et al. carried out comparative genomics of BCG Vaccines by whole-genome DNA microarray. They obtained evidence for the ongoing evolution of BCG strains since their original derivation. A precise understanding of the genetic differences between closely related *Mycobacteria* suggests rational approaches to the design of improved diagnostics and vaccines.[73]

The pathogenic bacterium *Neisseria meningitidis* is the cause of meningococcal meningitis. Although a vaccine is available for *Neisseria* of serogroups A and C, there is no vaccine available for serogroup B. Comparison of these genome sequences revealed highly conserved surface proteins that may be valuable as vaccine antigens. It is likely that these conserved regions encode essential components of the pathogenesis pathway.[74]

A computational process similar to that used for the annotation of genome sequences by simultaneous comparison can be used to identify candidate targets within the genome and prioritize them for antimicrobial screening. For broad-spectrum agents, for example, bioinformatic analysis of genome sequences can be used to identify proteins that are highly conserved in the appropriate range of pathogens associated with a particular clinical indication. This becomes even more important when the same bioinformatic tools are used to identify highly conserved bacterial genes that lack a close human counterpart. Comparative genomics therefore provides, through simple computational analysis, a list of potential targets with useful bacterial spectrum and possible selectivity over humans.[75]

In addition to highly conserved targets which offer the opportunity to develop broad spectrum antibiotics, comparison of microbial genome sequences has also shown that a significant proportion of each genome encodes proteins that are functionally unknown, some of which are specific to that organism. These provide the opportunity to develop antibiotics with a high degree of specificity for a single organism (or a small set of related bacterial species). Such narrow specificity potentially offers long-term benefits by reducing problems arising from cross resistance. Organism-specific genes may not only provide the potential targets for novel therapeutic agents but also the principal components of rapid, PCR-based diagnostic tools.[75]

Large-scale comparisons of human genomes from many individuals are now possible with the emergence of high-throughput techniques for DNA sequence determination. Hypervariable microsatellites, also called short tandem repeats (STRs) and SNPs are highly informative in pedigree, forensic, and population assesment. They are also valuable for understanding the effect of variation on a particular disease or trait and looking for disease genes.[63]

2–2. Functional Genomics The next step begun by the successful sequencings will be to determine what the many genes code for. Functional genomics is the study of genomes to determine the biological functions of all the genes and their products. This field consists of all the work done to bridge the knowledge gap from DNA (or genes) to proteins (or functions).

Innovative techniques for functional genomics

range from simply comparing new sequences to those already in the databases to using ''microarrays'' to analyze how gene expression patterns change as conditions vary and generating wholesale lots of mutant organisms that can then be screened for interesting trait changes. In addition, plant scientists are venturing into the vast frontier of ''metanomics,'' which tracks the effects of particular mutations or environmental changes on a plant's entire metabolic repertoire. Based on the progress so far, the advances in tools and techniques will allow unprecedented progress in figuring out what genes do.[76]

Small nucleolar RNAs (snoRNAs) are required for ribose 2'–O–methylation of eukaryotic ribosomal RNA. Many of the genes for this snoRNA family have remained unidentified in *Saccharomyces cerevisiae*, despite the availability of a complete genome sequence. Lowe and Eddy used probabilistic modeling methods to computationally screen the yeast genome and identify 22 methylation guide snoRNAs, snR50 to snR71. Gene disruptions and other experimental characterization confirmed their methylation guide function.[77]

The functions of many open reading frames (ORFs) identified in genome-sequencing projects are unknown. New, whole-genome approaches are required to systematically determine their function. Each gene in the genome was deleted in a directed fashion and by marking each yeast gene with two molecular barcodes (UPTAG and DOWNTAG) that allows large numbers of deletion strains to be pooled and analyzed in parallel in competitive growth assays. Of the deleted ORFs, 17 percent were essential for viability in rich medium. The phenotypes of more than 500 deletion strains were assayed in parallel. Of the deletion strains, 40 percent showed quantitative growth defects in either rich or minimal medium.[78]

Because zebrafish is a vertebrate, it is genetically closer to humans than flies or worms, and its small size, quick generation time, and inexpensive care make it possible to keep thousands of fish in a single lab. Add to that the transparency of its young, and zebrafish is an ideal lab animal. Researchers are using various clever techniques to identify zebrafish mutants with which to probe the genes involved in a wide variety of human maladies, from obesity to bone diseases.[79]

Goodnow, an immunologist at the Australian National University, is using the latest technology and sequencing data to advance research on recessive mutations that cause adult-onset diseases. The plan is to generate mouse mutants on a massive scale in order to help assign functions to the genes being identified by HGP. Because mouse genes are thought to do the same thing as their human counterparts, scientists hope to translate the knowledge into clinical studies.[80]

2–3. Computer and Biology

2–3–1. *Bioinformatics*

''Bioinformatics'' gained common currency in the early 1990s to describe the tools and techniques for storing, handling, and communicating the massive and ever-increasing amounts of biological data emerging principally from genomics research and clinical trials. Researchers will have to face a profound challenge: how to deal with the masses of data that will come pouring out. It's not just how to analyze and interpret the data, but how to share and compare them. Right now, for example, people are using different platforms, as well as different methods of analyzing the data those platforms produce. This lack of standardization makes it difficult to relate the findings of the different labs and assess their quality.[81]

NIH Urged to Fund Centers to Merge Computing and Biology. A new network of interdisciplinary research between biologists and computing expertise is needed to fully tap the flood of new data. NIH should also take a more active role in organizing and curating the growing flock of biomedical databases, which hold everything from gene sequences to drug trial results.[82] As the size of GenBank, the public archive that contains every published DNA sequence, and the number of other biological databases grows, so does the need for ways to update and coordinate the information they contain. GenBank and even databases whose entries are reviewed and updated can have mistakes or missing data.[83]

2–3–2. *Biological Information Science*

Information processing on computers and a new kind of biological information science are crucial, and their impact on biology, medicine, and health care will be enormous.

Comparing the genomes of different species offers insight into the function of conserved noncoding regions of DNA sequence. Dubchak has developed a new method for global alignment of sequences, their comparison, and the display of their

identity.[84]

To perform any computational analysis of the biological function of a large number of genes, one needs to expand the concept of gene function. Soon, the function of a gene can be described by its expression profile in a large number of controlled experiments. Comparative computation can then be performed on gene function in ways not previously possible. Efforts to construct archival databases of gene expression experiments to facilitate predictive computations are under way.[85]

The largest impact of genomic technologies on biological research will come from the emerging ability to simulate cells and organisms on the computer. The goal is to simulate the causal and temporal behavior of a cell as a network of genes and gene products and to simulate the behavior of the organism as a network of cells. Quantitative and predictive simulations have the potential of reducing or replacing experimental effort. Shapiro has sought to define the genetic network that coordinates the initiation of DNA replication under both temporal and spatial constraints. Arkin is building a computational toolkit for computer-aided simulation and analysis of developmental switches. His program permits the experimenter to simulate chemical kinetic systems and parts of cellular pathways.[85]

2–3–3. *Structural Genomics*

It's the verifiable 3D maps that provide the most reliable information. Structural genomics applies high-speed techniques to make a systematic survey of all protein structures, cataloging the common ways in which proteins fold. That information could eventually lead to computer programs capable of predicting the shape and function of any protein from the simple linear sequence of A's, G's, C's, and T's in genes. The approach is expected to extend the genomics revolution from a catalog of genes to a catalog of the 3D shapes of the proteins for which the genes code. For industry, structural genomics promises not only a wealth of new drug targets but also help in eliminating those not likely to be useful.[86]

In December 1999, IBM announced a $100 million research initiative to build a supercomputer 500 times more powerful than the current record holder in order to model how proteins fold into the three-dimensional shapes. NIH, the Ontario provincial government, and three more countries—Japan, Germany, and the United Kingdom—either have announced plans to fund for structual genomics or are considering plans for investments.[87] Officials at NIH say they hope the new program will enable them to determine the structure of as many as 10,000 proteins in the next 10 years. It's also expected that the coming bolus of protein structures will reveal a large fraction of the estimated 1000 to 5000 protein folds thought to exist.[88]

2–3–4. *Genomic Medicine and the Future of Health Care*[85]

Genomic technologies and computational advances are leading to an information revolution in biology and medicine. Simulations of molecular processes in cells and predictions of drug effects in humans will advance pharmaceutical research and speed up clinical trials. Computational prognostics and diagnostics that combine clinical data with genotyping and molecular profiling may soon cause fundamental changes in the practice of health care.

A combination of rich cellular data, genomic profiling, and computational prediction may provide early detection of undesirable drug properties. For example, the effect of known toxic compounds can be assessed by measuring the genomic expression profile in cell cultures and accumulating a set of characteristic profiles as a background information base. The advantage of such methods lies in the much lower cost of cell culture tests as compared to tests in animals and clinical trials. Information from functional genomics experiments will be crucial for the predictive elimination of unpromising drug candidates.

Today's clinical trials are expensive and time-consuming. To accelerate the assessment of clinical outcomes using genomic technologies, a detailed and accurate link between molecular profiles and clinical outcomes is required. Computational processing and reference to information and knowledge bases about organismic and disease processes would allow conclusions about the likely results of therapy to be reached much faster than with classical macroscopic indicators of clinical outcomes.

2–3–5. *Diversity Digitized*[89]

The genomic research by the HGP was done by data shearing on the internet; which influences on many fields such as biodiversity.

The bioinformatics revolution is finally enabling biodiversity researchers to communicate efficiently with one another, providing a springboard and a common language for progress. This effort has the

broader advantage of at last putting biodiversity information into the public domain in accessible forms on the Internet. ''Species 2000'' is an Internet-based global research program that aims to create an index of the world's (known) species. The Global Biodiversity Information Facility aims to ensure interoperability among the various databases now emerging from biodiversity studies. Some research collections date back centuries, providing a historical context for studies of the world's organisms. Paleontologic databases can help researchers assess biodiversity on the scale of millions of years.

2–4. Global Gene Expression Analysis Array technologies, that can analyze hundreds of genes simultaneously and show patterns of gene expression, was developed in 1995.[2,9] Transcript profiling using DNA microarrays provides a rapid and systematic method for the high-throughput analysis of gene expression at the level of the whole genome, providing a specific analysis of expression of each individual gene monitored by mRNA concentrations.

2–4–1. *Application to Basic Research*

Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation (anaerobic growth), with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage. Brown et al. used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.[90]

Roberts et al. used genome-wide transcript profiling to monitor signal transduction during yeast pheromone response. Genetic manipulations allowed analysis of changes in gene expression underlying pheromone signaling, cell cycle control, and polarized morphogenesis. Diagnostic subsets of coexpressed genes reflected signaling activity, cross talk, and overlap of multiple mitogen-activated protein kinase (MAPK) pathways.[91]

Normal human fibroblasts require growth factors for proliferation in culture; these growth factors are usually provided by fetal bovine serum. In the absence of growth factors, fibroblasts enter a nondividing state, G_0, characterized by low metabolic activity. Addition of growth factors induces proliferation of fibroblasts. The temporal program of gene expression during a model physiological response of human fibroblasts to serum, was explored with a complementary DNA microarray representing about 8600 different human genes.[92]

Ly et al. compared gene expression in cells from healthy people of various ages and also from children with Hutchinson-Gilford progeria, a rare hereditary disorder that resembles an accelerated form of aging. They report that some of the gene changes they saw in aging fibroblasts could cause skin to wrinkle. They also found evidence for what may be a more global explanation of aging: an impairment of the machinery needed for normal separation of the chromosomes during cell division that could lead to genetic instability and a variety of disturbances in gene function.[93,94]

The Wisconsin team carried out microarray analysis of aging in mice skeletal muscle. Use of high-density oligonucleotide arrays representing 6347 genes revealed that aging resulted in a differential gene expression pattern indicative of a marked stress response and lower expression of metabolic and biosynthetic genes. Transcriptional patterns of calorie-restricted animals suggest that caloric restriction retards the aging process by causing a metabolic shift toward increased protein turnover and decreased macromolecular damage.[93,95]

But except for some stress-response genes, there was little overlap between the alterations the two groups saw. The fibroblasts, which are dividing cells, and the skeletal muscle cells, which have lost that ability, probably undergo aging through two different mechanisms.[93]

Metamorphosis is an integrated set of developmental processes controlled by a transcriptional hierarchy that coordinates the action of hundreds of genes. In order to identify and analyze the expression of these genes, White et al. constructed high-density DNA microarrays containing several thousand *Drosophila melanogaster* gene sequences. Many differentially expressed genes can be assigned to developmental pathways known to be active during metamorphosis, whereas others can be assigned to pathways not previously associated with metamor-

phosis. Additionally, many genes of unknown function were identified that may be involved in the control and execution of metamorphosis.[96]

Caulobacter crescentus is a bacterium that thrives in aquatic environments that lack sufficient nutrients for most other life-forms. Typically, *C. crescentus* has a whiplike appendage called a flagellum that it uses for swimming. But when it's time to reproduce, *C. crescentus* jettisons its flagellum, replacing it with a short stalk that anchors the tiny cell to a nearby surface. The DNA then replicates and the stalked cell divides asymmetrically, pinching off a new, mobile ''swarmer'' cell. These characteristic ''stalked'' and ''swarmer'' stages enable microbiologists to associate genetic changes with distinct stages of the cell cycle. On march 2000, TIGR has just finished assembling the entire genetic sequence of *C. crescentus*. A team at Stanford University headed by Shapiro, in cooperation with TIGR, had begun working with these data about 9 months before they were assembled. They used microarray to monitor the RNA transcribed from each gene at 15–minute intervals over *C. crescentus*'s life cycle. Other global expression studies will start soon on the anthrax genome. Even though the entire sequence may not be complete until June 2001, TIGR's Read plans to start microarray studies in July 2000 to identify target proteins.[48]

Single hematopoietic stem cells can give rise to at least eight distinct blood cell lineages and can maintain life long blood production in mice. Their hallmark property is the ability to strike a balance between self-renewal and a commitment to differentiation. The mechanisms that govern these stem cell fate decisions must be under tight yet flexible control. Phillips et al. performed a genome-wide gene expression analysis in order to define regulatory pathways in stem cells as well as their global genetic program. Subtracted complementary DNA libraries from highly purified murine fetal liver stem cells were analyzed with bioinformatic and array hybridization strategies.[97]

Bresnahan and Shenk used a human cytomegalovirus gene array to identify a previously unidentified class of viral transcripts. These transcripts, termed virion RNAs, were packaged within infectious virions and were delivered to the host cell on infection. This mechanism of herpesvirus gene expression allows for viral genes to be expressed within an infected cell immediately after virus entry and in the absence of transcription from the viral genome.[98]

The importance of analysing expression data by clustering can be seen further from studies on co-regulated genes. Coexpression may indicate genes whose protein products interact together in a heterologous complex, or which act in concert in the same pathway without direct physical interaction. Transcriptional co-regulation can thus indicate physical or functional interactions between proteins, and associating the regulatory patterns of known proteins with those of unknown function can be used to suggest functional linkages for further analysis.

2–4–2. *Application to Drug Discovery*

Toxicology may be on the verge of changing the way it collects raw data—adopting a process that could reduce animal use and improve test results. The new approach, called ''toxicogenomics,'' uses DNA arrays to profile gene expression in cells exposed to test compounds. The great promise of toxicogenomics is that it might be used to scan the entire human genome to see which genes are affected by specific chemicals. The immediate goal is to look at different classes of compounds and identify groups of genes that are tightly correlated with known classes of toxicants. These ''very informative'' genes could then be used to generate a next-generation array with a small number of genes. The condensed set could be used routinely to determine if a test chemical exhibits any of several common toxicities. The DNA tests are fast, efficient, and reduce live-animal expenses. If adapted for use in tissue cultures, the tests might even eliminate the need to sacrifice animals.[99]

From the perspective of drug discovery, the patterns generated from the parallel analysis of all genes in an organism using microarrays can give clues to the function of previously uncharacterized genes (target identification), as well as providing information about cellular responses to treatments with small inhibitor molecules (mechanism of action studies at all discovery phases). Several recent lines of evidence support the view that such gene expression profiles have a key role to play in antimicrobial drug discovery programs. This suggests that genetic inhibition of gene function by mutation or deletion can be used as the ''gold standard'' marker for specific gene inhibition, and that expression patterns generated after treatment with small molecule inhibitors can be related to the ''genetic patterns'' to define or confirm the

target and mechanism of action of the inhibitor. Microarrays can also be used to identify secondary (and potentially unwanted) mechanisms of action.[75]

2–4–3. *Application to Therapy: DNA Arrays Reveal Cancer in Its Many Forms*[82,100]

The use of microarrays to determine gene expression patterns is providing a wealth of new information that should aid in cancer diagnosis and ultimately in therapy. Researchers in several labs have used microarray technology to identify specific subtypes of a variety of cancers, including leukemias and lymphomas, the dangerous skin cancer melanoma, and breast cancer. In some cases, they can determine which cancers are likely to respond to current therapies and which aren't. In addition, the studies are giving researchers a fix on which genes are important for the development, maintenance, and spread of the various cancers, and are thus possible drug targets.

A team led by Lander and Golub began by comparing the profiles of acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), two blood cancers that are often hard to tell apart by standard pathological examination of the diseased cells. They showed that the expression patterns of those genes could identify which patients had AML and which had ALL without previous knowledge of these classes.

Since then, researchers have used gene expression patterns to reveal previously unknown cancer categories. Staudt's team working in collaboration with Brown and Botstein of Stanford University School of Medicine and their colleagues focused on patients with diffuse large B cell lymphoma, a common type of non-Hodgkin's lymphoma that affects more than 15,000 new patients annually in the United States and follows a highly variable clinical course. They found a great diversity in gene expression among the patients, despite their having the same diagnosis. Computer analysis of the expression patterns showed that the patients could be divided into two groups. One group expressed a set of genes characteristically turned on in B cells in the spleen and lymph nodes during an immune response. The other set didn't express those genes but did show activity of a set of genes that are turned on when blood B cells are stimulated to divide by an antigen. Their clinical pictures also varied: Those with the expression pattern of the spleen-lymph node B cells fared much better, with 75% alive 5 years after diagnosis, while 75% of the other group did not make it to that milestone. Lymphoma patients are usually treated first with chemotherapy, and if they relapse, they become candidates for a bone marrow transplant. But in the future, those patients who have a gene expression profile indicating a poor prognosis might move directly to bone marrow transplant, avoiding the first-line chemotherapy regimen, which can be debilitating.

Brown and Botstein and their colleagues have also found that breast cancers show distinguishable patterns of gene expression. Again, they could pick out two broad groups of the tumors, one of which was marked by expression of the gene for the receptor for the hormone estrogen while the other one wasn't. Cancer physicians have long known that breast cancer cells that lack estrogen receptors tend to be more aggressive. But those broad groups of tumors may contain a variety of subgroups. Whether there is a difference in outcomes for the subgroups is now addressed.

Researchers want to do more than just identify genes whose activity is turned up or down. They also want to find out which of those changes are important for cancer development and progression—the causes and not just the effects. In the 3 August 2000 issue of *Nature*, the NHGRI team used arrays to compare the gene expression patterns of highly metastatic melanoma cells with those of the much less metastatic cells from which they were derived. The comparison identified a suite of genes whose activity was apparently turned up as melanoma cells progressed to malignancy.

Researchers are using the arrays to determine how the activation of cancer-promoting oncogenes or the inactivation of tumor-suppressor genes perturbs the expression of other genes. Staudt and his colleagues used their Lymphochip to study the consequences of abnormal activation of an oncogene called BCL–6, a situation that commonly occurs in lymphomas. They found that BCL–6 activation leads to repression of a gene called blimp–1, which normally promotes the differentiation of B cells to become antibody-producing plasma cells, and also of a gene called p27kip1, which inhibits the cell division cycle. The net result is to lock cells in an undifferentiated, continuously dividing state. The MIT group, in collaboration with Eisenman's team, has also looked at the changes elicited in cells by activation of the MYC oncogene. They are discovering the malignant pathways of these tumors, and they can ask whether inter-

fering with these pathways can help.

Other applications of microarray technology to cancer that are also getting under way include studying how cancer cells respond to various chemotherapeutic agents and determining why some cells respond and some don't.

2–5. Proteomics Proteomics is the study of the full set of proteins encoded by a genome. HGP and similar gene-sequencing efforts are only the first step to understanding biology and disease. The next big step forward will be deciphering what the protein product of each gene does and how each gene is switched on and off, and how they interact with each other.[101] Neither DNA nor mRNA can identify how much protein is produced inside a cell or what it does. There is significant evidence that there is not necessarily a correlation between mRNA levels and protein levels. Chemical modifications such as phosphorylation play a key role in controlling protein activity; these modifications cannot be detected by screening nucleotides. The genome tells you what could theoretically happen inside the cell. Messenger RNA tells you what might happen, and the proteome tells you what is happening. If you want to know what's happening to a cell's proteins, you have to study the proteins themselves. In the next decade, we can hope to see large catalogs of protein interactions, predictive models of those interactions, and at least some ability to follow networks of those interactions in real time.[102]

Newer methods for analysis of protein interactions include descendants of conventional two-hybrid methods, other methods that rely on reconstitution of biochemical function in vivo, fluorescence resonance energy transfer methods, protein mass spectrometry, and evanescent wave methods. Taken together, these methods will help reveal not only the partners of particular proteins, but how tightly the interacting proteins touch one another, which surfaces they use to make contact, and where and when in living cells those contacts occur. With these methods, entire networks of interacting proteins can be analyzed.[103]

Genomic information opens new paths to biochemical discoveries. The finding in a genome of many pairs of protein sequences A' and B' that are both homologs to a single sequence A∧B in another genome suggests the possibility that A' and B' are binding partners and provides robust functional information about A' and B'. Systematic searches of this sort may lead to identifications of new pathways and protein complexes in organisms.[104]

Walhout and co-workers report that they have started to compile a global map of interactions between all of the proteins in the worm *Caenorhabditis elegans* involved in vulval development, using large-scale two-hybrid analysis. They needed to develop a global approach rather than analyzing the interactions of each individual protein. They developed a high-throughput method to analyze proteins in parallel that can be automated and should be suitable for the study of protein-binding interactions across the entire genome. The resulting map will help to elucidate gene function on a global scale. For those proteins that are conserved during evolution, interactions between two worm proteins may clear the way to finding homologous interactions in mammals.[105]

Chemical biologist MacBeath and chemist Schreiber report creating arrays of over 10,000 proteins on a piece of glass just half the size of a microscope slide. Three applications for protein microarrays were demonstrated: screening for protein-protein interactions, identifying the substrates of protein kinases, and identifying the protein targets of small molecules. Researchers also hope to array antibodies that bind to specific proteins. That would enable them to see which proteins are actually being produced in various tissues and, presumably, offer further clues to what causes various diseases.[106]

The technique, devised by Shokat and his colleagues, involves enlarging the active site of an enzyme so that it can bind an inhibitor that won't fit into the active sites of related—but unaltered—enzymes. Researchers can then insert the gene that encodes the modified enzyme into cells or living animals and turn off that enzyme by feeding them the inhibitor—without affecting other, very similar, enzymes.[107]

3D crystal structures can provide important insights into the role of proteins inside cells. Computer science aims to model how proteins fold into the three-dimensional shapes that allow them to orchestrate life within the cell. If successful, tuis could allow drug researchers to go right from the sequence of a disease-related gene to the predicted structure of its protein, in order to identify targets for therapeutic drugs. Today, virtually every aspect in the structure-determining process is being automated and scaled up. Researchers are now creating high-throughput

systems for cloning genes, expressing proteins, growing crystals, and collecting x–ray data. More powerful x–ray beams at synchrotrons around the world have improved the quality of diffraction data, and better computers and software have made interpreting results both easier and faster. Similar advances are propelling work in nuclear magnetic resonance (NMR) spectroscopy, a related approach to determining structures.[87]

The upstart Celera Genomics of Rockville, Maryland, is set to embark on an ambitious new effort to move beyond the human genome to conquer the next frontier: ''proteomics,'' an effort to identify all the proteins expressed in an organism and then track their ebb and flow. Working in tandem, the molecular scanner and a high-speed mass spectrometer could be a powerful combination. The move is the next logical step in understanding the role of all the genes they've decoded. It is also a critical step in developing novel drugs and tailoring medical care to the genetic makeup of individuals. Other companies are also pushing into proteomics as well. Virtually every major pharmaceutical company has a proteomics effort under way.[102]

3. Exciting Impact on Conventional Fields

Genome sequencing is also providing exciting changes in the comventional fields.

3–1. Mollecular Biology

3–1–1. *Creating Life*[108]

One important question posed by the availability of complete genomic sequences is how many genes are essential for cellular life. What is life in genomic terms?; What is a minimal set of essential genes?

Mycoplasma genitalium with 517 genes has the smallest gene complement of any independently replicating cell so far identified. Venter et al. used global transposon mutagenesis to identify nonessential genes in an effort to learn whether the naturally occurring gene complement is a true minimal genome under laboratory growth conditions. The analysis suggests that 265 to 350 of the 480 protein-coding genes of *M. genitalium* are essential under laboratory growth conditions, including about 100 genes of unknown function. This work represents an important step in the path toward the creation of minimal organisms, organisms with the smallest set of genes that allow for survival and reproduction. This research raised ethical, social, and religious issues, and posed challenge to our conception of the meaning of life.[109]

Researchers are attempting to model and eventually to create ''minimal organisms''. Scientists have proposed, and are working on, two different ways of creating a new organism with a minimal genome. The first, a ''top-down'' approach, entails removing or inactivating the entire set of genes of *M. genitalium* thought to be unnecessary. The second, and more technically challenging, ''bottom-up'' approach, entails synthesizing the proposed minimal genome and inserting it into an environment that allows metabolic activity and replication. The means for synthesizing relatively short pieces of DNA already exists, but assembling the entire genome of an organism and proving that the genome is capable of supporting a free-living life form within that environment has not been done. Recent work by Venter et al. represents a significant step in the ''top-down'' approach. These experiments define a ''minimal essential set'' of genes that, individually, are required for replication under permissive laboratory growth conditions.

Schultz and his colleagues are attempting to find out what life would look like if DNA contained more than four nucleotide bases and proteins more than 20 amino acids. By reengineering DNA, RNA, and the proteins that interact with them, they hope to create synthetic organisms with a chemical makeup fundamentally different from all life that has existed on Earth for the last 3.8 billion years. The result will be proteins that incorporate amino acids other than the 20 commonly used by life to construct proteins. By adding these amino acids with completely new types of chemical behaviors, they hope to design bacteria to make proteins that work as novel catalysts and drugs, or that carry built-in tracers to help researchers decipher their structures. If they succeed, their biochemical reengineering could have a profound effect on everything from basic molecular biology to industrial chemistry.[110]

3–1–2. *Expectation and Anxiety about Creating Life*[108,110]

Creating a minimal genome would represent an important step forward in genetic engineering as it would permit the creation of organisms (new and existing) simply from knowing the sequence of their genomes. This research may provide insight into the origins of life, bacterial evolution, or the control of bacterial metabolism. In addition, definition of a

minimal genome could lead to a better understanding of the genomes of more complex modern organisms.

The first practical benefits might be in microbial engineering. Bacteria are now commonly engineered to produce useful products, ranging from industrial chemicals to insulin. A minimal organism might require less energy or produce fewer waste products that could contaminate the desired product. A minimal organism could be used as the basis for novel ''designer'' bacteria that are created to perform specific tasks, such as the breakdown of environmental toxins.

The building of new organisms raises intellectual property and commercialization issues that will affect the conduct of research and the ability of both industry and academia to continue developing the technology for public good. A new regulatory framework for intellectual property pertaining to genes and organisms is needed to ensure that public and commercial interests are protected.

The building of new organisms could, however, harm our health or the environment, by introducing ''alien'' species into the wild. The combination of large-scale sequencing of human pathogens, determination of function of disease-associated gene products, and development of technologies to assemble large pieces of DNA could lead to creation or release of organisms that could be used as biological weapons. The dangers of knowing the sequences of extremely deadly pathogens could pose threats to public health and safety that might outweigh the benefits. We need to give serious thought to monitoring and regulation at the level of national and international public policy.

There is a serious danger that the identification and synthesis of minimal genomes will be presented by scientists, depicted in the press, or perceived by the public as proving that life is reducible to or nothing more than DNA. Reducing life to genes has profound implications for several critical societal debates, including what constitutes human life and when life begins.

3–1–3. *New Insight in Chromosome Structure*

One of the rewards of having a *Drosophila melanogaster* whole-genome sequence will be the potential to understand the molecular bases for structural features of chromosomes that have been a long-standing puzzle. Benos et al. analysed 2.6 megabases of sequence from the tip of the X chromosome of *Drosophila* and identified 273 genes. Sequence analysis revealed that this region comprises 154 kilobases of DNA flanked by 1.2–kilobases of inverted repeats, each composed of a 350–base pair satellite related element. Thus, some aspects of chromosome structure appear to be revealed directly within the DNA sequence itself.[111)]

Copenhaver et al. used high-precision genetic mapping to define the regions that contain centromere functions on each natural chromosome in *Arabidopsis thaliana*. This investigation provides a platform for dissecting the role of individual sequences in centromeres in higher eukaryotes.[112)]

Genomic mobile elements called retrotransposons make up about 40% of the mammalian genome. During retrotransposition, these small pieces of DNA are duplicated by a ''copy and paste'' mechanism—they are transcribed into RNA, reverse-transcribed into DNA, and the complementary DNA is then inserted back into the genome at a new site. Although retrotransposons have been viewed as selfish DNAs that provide no benefit to their host cell, Kazazian states that these mobile pieces of DNA are busy reshaping our genome, making it more diverse and enabling us to survive and thrive through the vagaries of evolution. The LINE–1 or L1 element is one of the major retrotransposons without long terminal repeats (LTRs). Throughout evolution it is likely that L1 retrotransposition has donated functional domains to other proteins, expanded the diversity of the genome, and increased the genome's size through mobilization of non–L1 sequences.[113)]

Processed pseudogenes are nonfunctional intronless copies of genes. They are derived from mRNAs that have been reverse-transcribed and reinserted into the genome, a process similar to the duplication of L1 sequences. On human chromosome 22, processed pseudogenes account for 0.5% of genomic DNA. In addition, some of the many processed pseudogenes contribute new activities to the cell, such as providing new exons for preexisting genes. Processed pseudogenes have been generated by active L1 elements in tissue culture, which suggests that L1 proteins are the driving force behind processed pseudogene formation.[113)]

Alu elements are short, high-repeat DNA sequences that do not encode proteins. Alu elements may be mobilized by the trans action of L1 proteins as well. Roughly 1 million retrotransposed Alu elements

make up about10 to 12% of the human genome. The 300-base pair Alu elements are transcribed by RNA polymerase III into RNA that ends in a poly A tail. L1-encoded proteins are likely candidates for Alu mobilization because Alu elements are flanked by target site duplications that bear a close resemblance to the target site duplications of L1 elements, and DNA sequences at the sites of Alu insertions are similar to those found at L1 insertion sites. In addition to genome expansion through retrotransposition, Alu elements have shaped the genome through mispairing and unequal crossing-over, leading to deletions and duplications.[113]

3-1-4. *Exciting New Insights in Molecular Biology*

Little is known about the molecular mechanisms of taste perception in animals, particularly the initial events of taste signaling. Clyne, Warr, and Carlson identified large and diverse family of seven transmembrane domain proteins from the *Drosophila* genome database with a computer algorithm that identifies proteins on the basis of structure. Tissue specificity of expression of these genes, along with their structural similarity, supports the possibility that the family encodes a large and divergent family of taste receptors.[114]

Molecular biologist White and his team at TIGR announced that they have completed sequencing the *Deinococcus* genome. The decoded 3-million-base genome indicates that *Deinococcus* owes its extreme radiation resistance to the same repertoire of mechanisms for repairing DNA found in other organisms. It just has more of them than most other life-forms. One element of these defenses is an enzyme called MutT. Radiation damages cells in part by generating reactive forms of oxygen that oxidize key cellular compounds, including some of the nucleotide building blocks of DNA. These oxidized nucleotides can cause faulty DNA replication, but MutT protects against such mutations by helping rid the cell of the oxidized nucleotides. Most organisms have a single MutT gene, but with 20 MutT-like genes, *Deinococcus* is capable of removing a whole lot of oxidative products. Other results suggest that genetic engineers may be able to equip this hardy organism with genes that could enable it to degrade toxins and clean up metals at radioactive waste sites.[115]

Biologists who study the fungus *Candida albicans* have always assumed that this organism reproduces asexually because they have not found evidence of mating, meiosis, or a haploid stage of the life cycle. However, sequencing of the *C. albicans* genome has revealed the existence of a possible mating type locus. This finding has now been extended to demonstrate actual mating in the fungus. Two reports by Hull et al. and Magee et al. reveal that *C. albicans* does have a sex life. Apparently, this organism can be forced to mate, suggesting that mating may occur naturally, albeit rarely.[116]

Through genomic analyses of naturally occuring marine bacterioplankton, DeLong's teams have identified large populations of bacteria that convert sunlight hitting the sea surface into energy. These bacteria have light-harnessing abilities previously known to exist in a fungus and in nonbacterial microbes called archaea that hang out in the most hostile salty environments, such as salt ponds, where sustenance is scarce. They are equipped with a protein, known as bacteriorhodopsin, that enables them to thrive by harnessing light to generate ATP. The bacteria DeLong discovered use a type of bacterial chlorophyll that, until now, no one had found in bacteria in the open ocean. The bacterial rhodopsin was encoded in the genome of an uncultivated [gamma]-proteobacterium and shared highest amino acid sequence similarity with archaeal rhodopsins. They dubbed this new protein proteorhodopsin and tested its function by putting it into *Escherichia coli*. As they hoped, the modified bacteria not only made the protein, but they reacted to light by moving protons out of the cell and into the surrounding medium. That's characteristic of bacteriorhodopsin, whose proton-pumping activity helps set up a gradient whereby energy is generated by protons flowing back into the cell. Their data also indicate that a previously unsuspected mode of bacterially mediated light-driven energy generation may commonly occur in oceanic surface waters worldwide.[117]

3-2. Evolution

3-2-1. *Chimerism in Prokaryotic and Eukaryotic Genomes*[118]

Analyses of rRNA from many different organisms provided the basis for the clonal theory of the evolution of eukaryotic genomes from prokaryotes. This theory holds that genes have been passed directly from generation to generation, with modifications in the genes resulting in the appearance of new organisms. However, the availability of complete genome

sequences for many bacteria (prokaryotes) and for the yeast (a eukaryote) has called into question long-held views about the evolutionary tree of life, proposing chimerism in prokaryotic and eukaryotic genomes, which arises through transfer of groups of functionally similar genes between organisms.

As genomes contain large numbers of genes from different functional classes, it is now possible to analyze the evolutionary history of groups of genes that do similar jobs. Until recently, phylogenetic conclusions were based on the analysis of one or a few genes; now they are based on the analysis of hundreds. Thus, it is possible to ask questions about genome evolution that could never have been answered by analysis of only one gene.

In a recent analysis of the complete genome sequences of *Escherichia coli* (a proteobacterium), *Synechocystis* (a cyanobacterium), *Methanococcus* (an archaeabacterium), and *Saccharomyces* (a eukaryote), genes were found to fall into two functional superclasses: informational genes (those involved in transcription, translation, and related processes) and operational genes (those involved in housekeeping). Eukaryotes appear to have obtained their informational genes from an organism that is more closely related to *Methanococcus* than to either the proteobacterium or the cyanobacterium, whereas their operational genes seem to have come principally from an *Escherichia* relative. These new results begin to explain the mystifying, mixed origins of eukaryotic genomes.

Koonin and his co-workers observed that *Methanococcus* informational genes resembled their orthologs in yeast but not their orthologs in eubacteria (true bacteria). In contrast, *Methanococcus* operational genes were more closely related to those of their eubacterial relatives. They concluded that the *Methanococcus* genome is a chimera composed of genes from *Saccharomyces*, and genes from eubacteria.

Now, there is growing evidence that in prokaryotes, too, horizontal gene transfer and chimerism prevail. An investigation of prokaryote evolution found that operational genes, which represent approximately two-thirds of the prokaryotic genome, have been transferred laterally many times, whereas informational genes do not show characteristics in keeping with horizontal transfer. These results suggest that horizontal gene transfer is an important evolutionary mechanism in prokaryotes as well as in eukaryotes.

3–2–2. *Viral Origin*[119]

Most scientists agree that viruses are life-forms. They are not cells, but they have their own DNA or RNA genomes and can reproduce with the unwilling help of a cellular host. Yet although viruses are able to hijack organisms of all sorts, they have long been consigned to a taxonomic ghetto that had little to do with the origins of the three domains (Bacteria, Archaea, Eukarya). Recent work using the structure of viral genes and proteins to infer relationships between organisms has sparked some provocative ideas. Among them is the notion that viruses arose very early—perhaps before the three domains diverged—and the hypothesis that viruses, rather than being an aberrant branch on the tree of life, have played a major role in the evolution of multicellular organisms.

Earlier in 2000, Hendrix and Pittsburgh colleagues reported the complete DNA sequences of two bacteriophages that infect *Escherichia coli*. When the sequences were compared to each other as well as to those of two other outwardly similar bacteriophages, it became clear that all four have some DNA stretches in common. But these stretches are interspersed with longer sequences that vary greatly from one bacteriophage to the next. Such genetic mosaicism has often been taken as evidence that viral strains swap genes. Microbiologist Woese suggested that early life was a hotbed of ''lateral gene transfer'' between cells, and that this gene swapping was the key driver of evolution. Viruses might have been important early vehicles for such gene exchange.

3–2–3. *Conservation and Novelty in the Evolution of Genes*

The genome of the nematode *Caenorhabditis elegans*, now fully sequenced, affords remarkable insights into the origin and nature of multicellar life. Moreover, it raises challenging, often unforeseen, questions about the molecular processes and evolutionary consequences of genome change. New proteins and modules have been invented throughout evolution. Gene ''birth dates'' in *C. elegans* range from the origins of cellular life through adaptation to a soil habitat. Possibly half are ''metazoan'' genes, having arisen sometime between the yeast-metazoan and nematode-chordate separations. These include basement membrane and cell adhesion molecules implicated in tissue organization. By contrast, epithelial surfaces facing the environment have specialized

components invented within the nematode lineage. Moreover, interstitial matrices were likely elaborated within the vertebrate lineage. A strategy for concerted evolution of new gene families, as well as conservation of adaptive genes, may underlie the differences between heterochromatin and euchromatin.[120]

3–3. Plant Biology There are many opportunities to use the wealth of sequence information to accelerate progress toward a comprehensive understanding of the genetic mechanisms that control plant growth and development and responses to the environment. Gene sequencing are providing new approaches for gene discovery and development of plants carrying desired traits.[44]

The nutritional health and well-being of humans are entirely dependent on plant foods either directly or indirectly when plants are consumed by animals. Plant foods provide almost all essential vitamins and minerals and a number of other health-promoting phytochemicals. Because micronutrient concentrations are often low in staple crops, research is under way to understand and manipulate synthesis of micronutrients in order to improve crop nutritional quality. Genome sequencing projects are providing novel approaches for identifying plant biosynthetic genes of nutritional importance. The term ''nutritional genomics'' is used to describe work at the interface of plant biochemistry, genomics, and human nutrition.[121]

The composition of oils, proteins, and carbohydrates in seeds of corn, soybean, and other crops has been modified to produce grains with enhanced value. Genomics-based strategies for gene discovery, coupled with high-throughput transformation processes and miniaturized, automated analytical and functionality assays, have accelerated the identification of product candidates carrying the desired traits.[122]

The *Hs1pro–1* locus confers resistance to the beet cyst nematode, a major pest in the cultivation of sugar beet. Cai et al. cloned the *Hs1pro–1* gene with the use of genome-specific satellite markers and chromosomal break-point analysis. Expression of the corresponding complementary DNA in a susceptible sugar beet conferred resistance to infection with the beet cyst nematode.[123]

Generating wholesale lots of mutant plants can be used for screening of interesting trait changes. Researchers are using transposable elements to generate lots of mutant plants that can then be screened for interesting trait changes, such as drought tolerance or sweeter kernels. These technologies can help identify genes that plants turn on or off in response to stresses such as drought or salty soils—information the biotech industry welcomes eagerly.[124]

Tobacco mosaic virus (TMV) is used to shuttle genes into plant cells to trace their function. In one type of application, the researchers have created ''libraries'' by separately cloning thousands of genes from a plant, such as *Arabidopsis*, into TMV. They then infect tobacco plants in the greenhouse with the altered TMVs and screen the resulting plants for changes, such as disease or drought resistance, conferred by the transplanted gene.[124]

Researchers are also stepping back to get a broader picture of how a plant alters its patterns of gene expression or biochemistry over time or in response to changing environmental pressures. If a series of genes wink on and off together, they are likely to be operating in the same pathway. And if an unknown gene either over- or underexpressed in a plant affects the members of such a pathway, the unknown gene can be assigned to that pathway as well.[124]

Although microarray techniques can place new genes into known pathways, they may not reveal exactly what they do in those pathways. Researchers are building protein expression pattern databases for *Arabidopsis*, rice, maize, and pine trees. Comparisons of protein patterns from different plant organs and tissues show how the patterns change as a result of seasonal variations or water restriction.[124]

The field ''metanomics'' tracks the effects of particular mutations or environmental changes on a plant's entire metabolic repertoire. Researchers can get a quick idea of what biochemical pathway an unknown gene might perturb, or the side effects of a particular gene alteration, by comparing the metabolic profile of a genetically altered plant with those already logged into the databases. Plant researchers want to create metabolic ''maps'' or generalized profiles of the metabolites produced by various plants and plant tissues. Profiles are likely to change when the plant is subjected to various environmental conditions or when an unknown gene is mutated.[124]

4. Revolution in Drug and Therapy

Genome sequencing is revolutionalising the way in drug discovery and therapy.

4–1. Vaccine Developement Bacille Calmette-Guerin (BCG) vaccines are live attenuated strains of *Mycobacterium bovis* administered to prevent tuberculosis. To better understand the differences between *M. tuberculosis*, *M. bovis*, and the various BCG daughter strains, Behr et al. studied their genomic compositions by performing comparative hybridization experiments on a DNA microarray. A precise understanding of the genetic differences between closely related *Mycobacteria* suggests rational approaches to the design of improved diagnostics and vaccines.[73]

Neisseria meningitidis is a major cause of bacterial septicemia and meningitis. Sequence variation of surface-exposed proteins and cross-reactivity of the serogroup B capsular polysaccharide with human tissues have hampered efforts to develop a successful vaccine. To overcome these obstacles, Venter et al. used the entire genome sequence of a virulent serogroup B strain (MC58) to identify vaccine candidates. They identified the proteins that are surface exposed, that are conserved in sequence across a range of strains, and that induce a bactericidal antibody response, a property known to correlate with vaccine efficacy in humans.[125]

4–2. Search for New Drugs Filaria worms infect some 120 million people worldwide, causing the tropical diseases African river blindness and elephantiasis. In 1995, Bandi and his team reported that they had detected ribosomal DNA sequences indicating that the worms are infected by *Wolbachia* bacteria. In 1998, Bandi, Blaxter, and their colleagues confirmed these data. Based on their analyses, they concluded that the bacteria had long ago parasitized the nematodes. The bacteria seem to contribute to the nematode's reproductive success. Fleischer and his colleagues reported in the January *Journal of Clinical Investigation* that the antibiotic tetracycline kills the bacteria living in the reproductive tissue of nematodes found in mice. The treatment resulted in smaller, infertile nematodes. So Slatko and his colleagues have decided to sequence a *Wolbachia* genome, hoping the sequence may lead to better anti-*Wolbachia* drugs, which in turn may be more effective in killing the worms by preventing their reproduction.[115]

Not only is resistance to antibiotics escalating, but also a new range of organisms have to be considered as potential pathogens. Over the past 40 years, the search for new antibiotics has been largely restricted to well-known compound classes active against a standard set of drug targets. Recent advances in genomics have provided an opportunity to expand the range of potential drug targets and have facilitated a fundamental shift from direct whole-cell antimicrobial screening programs toward rational target-based strategies. All genes in microbial organisms are now available as potential targets. Ideal antimicrobial targets should be essential to microbial cell survival, highly conserved in a range of pathogens, and absent or radically different in humans.[75]

4–3. Pharmacogenomics and Pharmacogenetics [126] There is great heterogeneity in the way individuals respond to medications, in terms of both host toxicity and treatment efficacy. The overall pharmacologic effects of medications are typically not monogenic traits; rather, they are determined by the interplay of several genes encoding proteins involved in multiple pathways of drug metabolism, disposition, and effects. HGP, coupled with functional genomics and high-throughput screening methods, is providing powerful new tools for elucidating polygenic components of human health and disease. This has spawned the field of ''pharmacogenomics'', which aims to capitalize on these insights to discover new therapeutic targets and interventions and to elucidate the constellation of genes that determine the efficacy and toxicity of specific medications. In this context, pharmacogenomics refers to the entire spectrum of genes that determine drug behavior and sensitivity, whereas pharmacogenetics is often used to define the more narrow spectrum of inherited differences in drug metabolism and disposition, although this distinction is arbitrary and the two terms are now commonly used interchangeably. Ultimately, knowledge of the genetic basis for drug disposition and response should make it possible to select many medications and their dosages on the basis of each patient's inherited ability to metabolize, eliminate, and respond to specific drugs.

Pharmacogenomic studies will provide new insights for the development of medications that target critical pathways in disease pathogenesis. A better understanding of the links between genes and disease could give rise to a new generation of highly effective drugs that treat causes, not just symptoms. Pharmacogenomic studies will also permit the developement of medications that can be used to prevent diseases in individuals who are genetically predisposed to them. Such studies should also permit the develop-

ment of therapeutic agents targeted for specific, but genetically identifiable, subgroups of the population. This represents a migration from the traditional strategy of trying to develop medications that are safe and effective for every member of the population (individualizing drug therapy).[127]

Refined diagnosis and choice of personalized therapy will come, which take into account a patient's genotype and history and details of her molecular health profile. Personalized therapy is supported by an expanded spectrum of drugs developed to target particular disease subtypes on a particular genetic background. Molecular profiling is used to monitor the progress of the disease, and therapy may be adjusted flexibly. Overall, the primary goal of personalized medicine should be to increase the quality of life first, and life-span second.

4–4. Gene Testing and Gene Therapy The revolution in genetics has led to the determination of the precise genetic basis of common and uncommon hereditary diseases. The first fruits of this revolution are diagnostic—the ability to determine who is and who is not at risk for disease before the onset of symptoms. Such information is becoming essential for proper management of patients and their families. In individuals who inherit mutant genes, simple preventative measures often can reduce morbidity and mortality and allow more thoughtful planning for the future. The benefits of genetic testing are equally important for those family members who are found not to carry the relevant mutation; these individuals are spared unnecessary medical procedures and tremendous anxiety, though genetic testing is not without its problems.[127]

Because gene therapy is highly experimental and many patients are desperately ill, serious adverse events and even deaths will occur. It is vital to understand the reasons for unexpected results or clinical failures to allow the development of corrected procedures and improved experimental methods. The recent tragic death of a young gene therapy patient has marred the field of gene therapy research. Yet, encouraging new results from a study in which a severe immunodeficiency disease is corrected in two infants by delivery of the defective gene have brought cautious optimism back to the gene therapy field.[128]

4–5. Impact on Drug Companies Early in 1997, drug companies, biotechs, and Wall Street investors were putting their money down on efforts to unlock the secrets of human DNA. Not only do pharmaceutical companies expect genomics to deliver them more targets, they also believe that this surge of genetic information will help them develop drugs more quickly. Combination of combinatorial chemistry, bioinformatics, and genomics would have drug-discovery revolution.[129]

Sweden's drug companies have joined the transnational drug industry, but their researchers are finding the global job market tough going. In 1995, Pharmacia merged with U.S.-based drug giant Upjohn, and in 1998 spring, Astra merged with Britain's Zeneca, both accompaning inevitable job cuts in Sweden.[130] And Pharmacia & Upjohn Inc.—itself the product of a 1995 merger—announced in December 1999 that it would merge with Monsanto Co. to create ''a first-tier pharma company with a first-tier growth rate.'' On 17 January 2000, Britain's giant Glaxo Wellcome PLC merged with the equally huge SmithKline Beecham PLC. Just 3 weeks after the Glaxo SmithKline announcement, Pfizer Inc. and Warner-Lambert Co. linked hands to create a behemoth with an even bigger R&D platform.[131]

One of the forces behind the recent mergers is that each company must have enough market share and product lines to maintain ''a significant presence'' in the United States, Europe, and Japan, the three major world markets. Other forces stem from the need for ever-larger R&D establishments, both to take advantage of new opportunities and to cover the continuing rise in the cost of developing new drugs. HGP is yielding clues to thousands of new research targets for drug development. But following up on those clues will be hugely expensive and will require a supersized scale of operation. The easiest way to achieve that magnitude quickly is to merge.[131]

In a major switch from just 20 years ago, a growing proportion of drugs is now prescribed for chronic conditions rather than as short-term antibiotics or pain-killers. This change has forced government regulators to look more closely at each drug's long-range effects.[131]

Companies are also striving to shorten the time from discovery to marketing—to get the most out of the finite length of a patent and, often, to beat a competitor to market. But this competition raises costs by forcing companies to juggle more balls. They need to run more things in parallel, rather than sequentially.[131]

The cost of clinical trials makes up a huge share of pharmaceutical companies' R&D budgets, nearly 40% by one independent estimate, and both per-patient costs and the costs of sophisticated tests seem likely to keep rising. Another 10% goes toward developing processes to synthesize compounds on a huge scale while controlling their production precisely. In addition, a substantial fraction of R&D money goes to screen compounds that might produce a new or improved drug.[131)]

5. Conclusion

A small knowledge base was created by organizing the HGP and its related issues in ''*Science*'' magazines between 1996 and 2000. This base revealed the stunning achievement of HGP and a private venture and its impact on today's biology and life science. In the mid-1990, they encouraged the developement of advanced high throughput automated DNA sequencers and the technologies that can analyse all genes at once in a systematic fashion. Using these technologies, they completed the genome sequence of human and various other organisms. These fruits opened the door to comparative genomics, functional genomics, and the interdisprinary field between computer and biology, and proteomics. They caused a shift in biological investigation from studying single genes or proteins to studying all genes or proteins at once. They are also causing or will cause revolutional changes in traditional biology, drug discovery and therapy.

ACKNOWLEDGEMENT

The authors are deeply grateful to Toshio INOUE, Professor Emeritus of Tokyo University, for his suggesting what they could do without research funds and instruments.

REFERENCES

1) Roberts L., *Science*, **291**, 1182–1188 (2001).
2) Roberts L., Pennisi E., Marshall E. *et al.*, *Science*, **291**, 1195–1200 (2001).
3) Marshall E., Pennisi E., *Science*, **280**, 994–995 (1998).
Venter J. C., Adams M. D. *et al.*, *Science*, **280**, 1540–1542 (1998).
Marshall E., *Science*, **284**, 1906–1909 (1999).
4) Collins F. S. *et al.* and the members of the DOE and NIH planning groups, *Science*, **282**, 682–689 (1998).
5) Marshall E., *Science*, **284**, 1439–1441 (1999).
Pennisi E., *Science*, **284**, 1822–1823 (1999).
6) Pennisi E., *Science*, **287**, 2182–2184 (2000).
Adams M. D., Venter J. C. *et al.*, *Science*, **287**, 2185–2195 (2000).
Myers E. W., Venter J. C. *et al.*, *Science*, **287**, 2196–2204 (2000).
Kornberg T. B., Krasnow M. A., *Science*, **287**, 2218–2220 (2000).
7) Marshall E., Pennisi E., Roberts L., *Science*, **287**, 2396–2398 (2000).
8) Marshall E., *Science*, **288**, 2294–2295 (2000).
Pennisi E., *Science*, **288**, 2304–2307 (2000).
9) Fodor S. P. A. *et al.*, *Science*, **274**, 610–614 (1996).
Fodor S. P. A., *Science*, **277**, 393–395 (1997).
10) Hawkins T. L., Lander E. S. *et al.*, *Science*, **276**, 1887–1889 (1997).
11) Kostrikis L. G., Tyagi S., Mhlanga M. M., Ho D. D., Kramer F. R., *Science*, **279**, 1228–1229 (1998).
12) Alper J., *Science*, **279**, 2044–2045 (1998).
13) Strausberg R. L., Feingold E. A., Klausner R. D., Collins F. S., *Science*, **286**, 455–457 (1999).
14) Service R. F., *Science*, **288**, 425–427 (2000).
15) Taton T. A., Mirkin C. A., Letsinger R. L., *Science*, **289**, 1757–1760 (2000).
16) Michalet X. *et al.*, *Science*, **277**, 1518–1523 (1997).
17) Pennisi E., *Science*, **280**, 816 (1998).
18) Service R. F., *Science*, **280**, 995 (1998).
19) Ronaghi M., Uhlen M., Nyren P., *Science*, **281**, 363–365 (1998).
20) Mullikin J. C., McMurray A. A., *Science*, **283**, 1867–1868 (1999).
21) Service R. F., *Science*, **282**, 399–401 (1998).
22) Burns M. A. *et al.*, *Science*, **282**, 484–487 (1998).
23) Weigl B. H., Yager P., *Science*, **283**, 346–347 (1999).
24) Belgrader P. *et al.*, *Science*, **284**, 449–450 (1999).
25) Rogers J., *Science*, **286**, 429 (1999).
26) Mirkin C. A., *Science*, **286**, 2095–2096 (1999); Kim P., Lieber C. M., *ibid.*, **286**, 2148–2150 (1999).
27) Lin J., Venter J. C. *et al.*, *Science*, **285**, 1558–1562 (1999).

28) Pennisi E., *Science*, **286**, 1263–1264 (1999).
29) Xin-zhuan Su *et al*., *Science*, **286**, 1351–1353 (1999).
30) Hoskins R. A. *et al*., *Science*, **287**, 2271–2274 (2000).
31) Pennisi E., *Science*, **280**, 814–817 (1998).
32) The Science, News and Editorial Staffs, *Science*, **286**, 2239–2243 (1999).
33) Waterston R., Sulston J. E., *Science*, **282**, 53–54 (1998).
34) Normile D., Pennisi E., *Science*, **285**, 2038–2039 (1999).
35) Pennisi E., *Science*, **288**, 417–419 (2000).
36) Pennisi E., *Science*, **288**, 939 (2000).
37) Pennisi E., *Science*, **291**, 1177–1180 (2001).
38) Goffeau A., Tettelin H., Oliver S. G. *et al*., *Science*, **274**, 546–567 (1996).
39) Fraser C. M., Venter J. C. *et al*., *Science*, **281**, 375–388 (1998).
40) Meinke D. W., Cherry J. M., Dean C., Rounsley S. D., Koornneef M., *Science*, **282**, 662–682 (1998).
41) Gardner M. J., Venter J. C. *et al*., *Science*, **282**, 1126–1132 (1998).
42) Pennisi E., *Science*, **282**, 1972–1974 (1998). The C.elegans Sequencing Consortium, *Science*, **282**, 2012–2018 (1998).
43) Pennisi E., *Science*, **283**, 1243 (1999).
44) Somerville C., Somerville S., *Science*, **285**, 380–383 (1999).
45) Pennisi E., *Science*, **286**, 210 (1999).
46) White O., Venter J. C. *et al*., *Science*, **286**, 1571–1577 (1999).
47) Pennisi E., *Science*, **287**, 1179–1181 (2000).
48) Pennisi E., *Science*, **287**, 1572–1573 (2000).
49) Tettelin H., Venter J. C. *et al*., *Science*, **287**, 1809–1815 (2000).
50) Pennisi E., *Science*, **288**, 239–241 (2000); Bloom F., *Science*, **288**, 973 (2000).
51) Hagmann M., *Science*, **288**, 800–801 (2000).
52) DiRita V. J., *Science*, **289**, 1488–1489 (2000).
53) Wang D. G., Lander E. S. *et al*., *Science*, **280**, 1077–1082 (1998).
54) Roberts L., *Science*, **287**, 1898–1899 (2000).
55) Marshall E., *Science*, **284**, 406–407 (1999).
56) Winzeler E. A. *et al*., *Science*, **281**, 1194–1197 (1998).
57) Michikawa Y., Mazzucchelli F., Bresolin N., Scarlato G., Attardi G., *Science*, **286**, 774–779 (1999).
58) Smith J. R., Collins F. S. *et al*., *Science*, **274**, 1371–1374 (1996).
59) Brzustowicz L. M., Hodgkinson K. A., Chow E. W. C., Honer W. G., Bassett A. S., *Science*, **288**, 678–682 (2000).
60) Watson A., *Science*, **289**, 850–854 (2000).
61) Boguski M. S., *Science*, **286**, 453–455 (1999).
62) Cohen J., *Science*, **275**, 769 (1997); Pennisi E., *ibid*., **288**, 1146–1147 (2000).
63) Paabo S., *Science*, **291**, 1219–1220 (2001).
64) Chervitz S. A. *et al*., *Science*, **282**, 2022–2028 (1998).
65) Bargmann C. I., *Science*, **282**, 2028–2033 (1998).
66) Ruvkun G., Hobert O., *Science*, **282**, 2033–2041 (1998).
67) Clarke N. D., Berg J. M., *Science*, **282**, 2018–2022 (1998).
68) O'Brien S. J. *et al*., *Science*, **286**, 458–481 (1999).
69) Rubin G. M., Venter J. C. *et al*., *Science*, **287**, 2204–2215 (2000).
70) Loots G. G. *et al*., *Science*, **288**, 136–140 (2000).
71) Pennisi E., *Science*, **284**, 2081 (1999).
72) Kaessmann H., Wiebe V., Paabo S., *Science*, **286**, 1159–1162 (1999).
73) Behr M. A. *et al*., *Science*, **284**, 1520–1523 (1999).
74) Nassif X., *Science*, **287**, 1767–1768 (2000).
75) Rosamond J., Allsop A., *Science*, **287**, 1973–1976 (2000).
76) Gura T., *Science*, **287**, 412–414 (2000).
77) Lowe T. M., Eddy S. R., *Science*, **283**, 1168–1171 (1999).
78) Winzeler E. A. *et al*., *Science*, **285**, 901–906 (1999).
79) Vogel G., *Science*, **288**, 1160–1161 (2000).
80) Finkel E., *Science*, **288**, 1572–1573 (2000).
81) Marx J., *Science*, **289**, 1670–1672 (2000).
82) Malakoff D., *Science*, **284**, 1742 (1999).
83) Pennisi E., *Science*, **286**, 447–450 (1999).
84) Spengler S. J., *Science*, **287**, 1221–1223 (2000).
85) Sander C., *Science*, **287**, 1977–1978 (2000).
86) Service R. F., *Science*, **287**, 1954–1956 (2000).
87) Service R. F., *Science*, **286**, 2250 (1999).
88) Service R. F., *Science*, **289**, 2254–2255 (2000).

89) Sugden A., Pennisi E., *Science*, **289**, 2305 (2000).
90) DeRisi J. L., Iyer V. R., Brown P. O., *Science*, **278**, 680–686 (1997).
91) Roberts C. J. *et al.*, *Science*, **287**, 873–880 (2000).
92) Iyer V. R., Brown P. O. *et al.*, *Science*, **283**, 83–87 (1999).
93) Marx J., *Science*, **287**, 2390 (2000).
Pennisi E., *Science*, **286**, 664 (1999).
94) Ly D. H., Lockhart D. J., Lerner R. A., Schultz P. G., *Science*, **287**, 2486–2492 (2000).
95) Lee C.-K., Klopp R. G., Weindruch R., Prolla T. A., *Science*, **285**, 1390–1393 (1999).
96) White K. P., Rifkin S. A., Hurban P., Hogness D. S., *Science*, **286**, 2179–2184 (1999).
97) Phillips R. L. *et al.*, *Science*, **288**, 1635–1640 (2000).
98) Bresnahan W. A., Shenk T., *Science*, **288**, 2373–2376 (2000).
99) Lovett R. A., *Science*, **289**, 536–537 (2000).
100) Golub T. R., Lander E. S. *et al.*, *Science*, **286**, 531–537 (1999).
101) Brenner S., *Science*, **287**, 2173–2174 (2000).
102) Service R. F., *Science*, **287**, 2136–2138 (2000).
103) Mendelsohn A. R., Brent R., *Science*, **284**, 1948–1950 (1999).
104) Marcotte E. M. *et al.*, *Science*, **285**, 751–753 (1999).
105) Walhout A. J. M., *Science*, **287**, 116–122 (2000).
Kim S. K., *Science*, **287**, 52–53 (2000).
106) Service R. F., *Science*, **289**, 1673 (2000).
MacBeath G., Schreiber S. L., *Science*, **289**, 1760–1762 (2000).
107) Strauss E., *Science*, **289**, 2029–2031 (2000).
108) Cho M. K., Magnus D., Caplan A. L., McGee D., the Ethics of Genomics Group, *Science*, **286**, 2087–2090 (1999).
109) Hutchison III C. A., Venter J. C. *et al.*, *Science*, **286**, 2165–2169 (1999).
110) Service R. F., *Science*, **289**, 232–235 (2000).
111) Benos P. V., Glover D. M. *et al.*, *Science*, **287**, 2220–2222 (2000).
112) Copenhaver G. P., Preuss D. *et al.*, *Science*, **286**, 2468–2474 (1999).
113) Kazazian Jr. H. H., *Science*, **289**, 1152–1153 (2000).
114) Clyne P. J., Warr C. G., Carlson J. R., *Science*, **287**, 1830–1834 (2000).
115) Pennisi E., *Science*, **283**, 1105–1106 (1999).
116) Gow N. A. R., Brown A. J. P., Odds F. C., *Science*, **289**, 256–257 (2000).
Hull C. M., Raisner R. M., Johnson A. D., *Science*, **289**, 307–310 (2000).
Magee B. B., Magee P. T., *Science*, **289**, 310–313 (2000).
117) Pennisi E., *Science*, **289**, 1869 (2000); Beja O. *et al.*, *Science*, **289**, 1902–1906 (2000).
118) Lake J. A., Jain R, Rivera M. C., *Science*, **283**, 2027–2028 (1999).
119) Balter M., *Science*, **289**, 1866–1867 (2000).
120) Hutter H. *et al.*, *Science*, **287**, 989–994 (2000).
121) DellaPenna D., *Science*, **285**, 375–379 (1999).
122) Mazur B., Krebbers E., Tingey S., *Science*, **285**, 372–375 (1999).
123) Cai D. *et al.*, *Science*, **275**, 832–834 (1997).
124) Gura T., *Science*, **287**, 412–414 (2000).
125) Pizza M., Venter J. C. *et al.*, *Science*, **287**, 1816–1820 (2000).
126) Cohen J., *Science*, **275**, 776 (1997).
Kleyn P. W., Vesell E. S., *Science*, **281**, 1820–1821 (1998).
Evans W. E., Relling M. V., *Science*, **286**, 487–491 (1999).
127) Marshall E., *Science*, **275**, 782 (1997).
Yan H., Kinzler K. W., Vogelstein B, *Science*, **289**, 1890–1892 (2000).
128) Ye X, Wilson J. M. *et al.*, *Science*, **283**, 88–91 (1999).
Friedmann T., *Science*, **287**, 2163–2165 (2000).
Anderson W. F., *Science*, **288**, 627–629 (2000).
129) Cohen J., *Science*, **275**, 767–772 (1997).
130) Rose J., Nilsson A., *Science*, **286**, 2063 (1999).
131) Agnew B., *Science*, **287**, 1952–1953 (2000).

要約

1996年から2000年の間にサイエンス誌に掲載されたヒューマンゲノムプロジェクト（HGP）とこれに関連する記事を整理して小さな知識ベースを構築した。この知識ベースから公的研究資金が導入されたHGPと私的な研究資金を導入したベンチャー企業との血のにじむ努力によって成し遂げられた驚くべき偉業と，その成果が今日のバイオテクノロジー及び生命科学に引き起こしている革命的変化を明解に把握することが出来た。1990年代半ばには，これらのグループの努力は高速自動DNAシークエンサーや一度にすべての遺伝子をシステマティックに分析出来るテクノロジーの開発を促した。これらのテクノロジーを用いることによって遺伝子配列決定の速度が加速され，両グループはヒトやその他の生物の完全な遺伝子配列を決定することが出来た。これらの成果によって比較ゲノミックス，ファンクショナルゲノミックス，コンピューターとバイオテクノロジーの境界領域，プロテオミックス等の新しい分野が生み出された。生物学的研究は少数の遺伝子や蛋白質に着目する研究から一度にすべての遺伝子や蛋白質に着目する方向へシフトし，従来のバイオロジー、医薬品開発，治療の分野にも革命的変化が引き起こされた。可能性のある医薬品のターゲットの範囲が拡大し，医薬品開発プログラムの，合理的なターゲットに基礎を置く方策へのシフトが促された。ただ症状を抑えるのではなく原因を扱う新世代の高度に効果的医薬品を生み出すことの出来るファーマコゲノミクスという分野が誕生した。今後はだれにでも安全で効果のある従来の薬物治療からパーソナライズドメディシン，パーソナライズドセラピーへのシフトも認められるであろう。