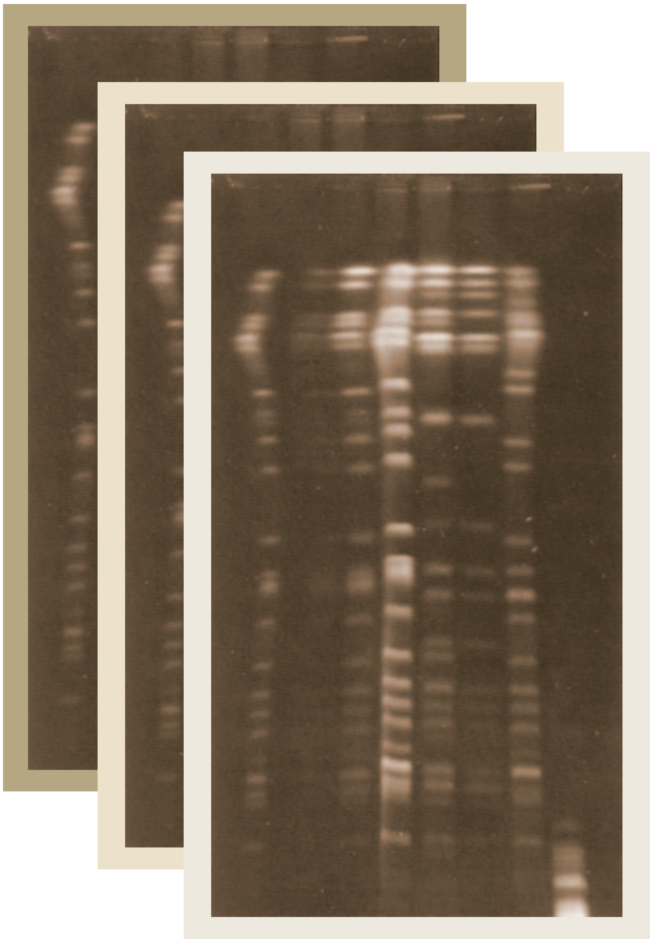
MILESTONE 1

Keep ‘em separate



In the early 1950s, the biochemical community started thoroughly investigating the chemical reactivity of DNA and RNA, with the hopes that these studies would help to elucidate the molecular structures and cellular functions of these macromolecules.

One topic of interest was the molecular mechanism of RNA and DNA hydrolysis; an important milestone in this field was published in 1952, when Markham and Smith reported that the hydrolysis of RNA proceeded via a cyclic phosphate intermediate, which was then further hydrolysed to produce a nucleoside 2′-monophosphate or 3′-monophosphate. A key development that led to this discovery involved the separation of complex mixtures of hydrolysed RNA using a simple device, termed an ‘electrophoresis apparatus’. The device was constructed from Whatman number 3 paper, several museum jars and various buffer solutions; the hydrolysed ribonucleic acids were deposited onto the paper, and a power supply was attached to the device. Applying the current led to the separation of the complex mixture into its components; remarkably, relatively minor differences in the structure of the molecules in the mixture led to observable differences in mobility across the paper. Using this approach, the authors were able to isolate ‘cyclic’ nucleotides, suggesting that the hydrolysis of RNA proceeded via the formation of a cyclic phosphate intermediate.

This general approach was not restricted to small molecules — ‘electrophoresis’ could also be performed on larger biomolecules if several adjustments were made. In 1955, Smithies demonstrated that gels made from starch solutions could be used to separate human serum proteins: when a hot starch solution was poured into a plastic tray, the cooled solution would form a solid (but brittle) ‘gel’. Protein samples could be loaded into the gel, which would then act as the stationary phase, much like the Whatman paper described above. Gel electrophoresis was further refined, and, 12 years later, Loening demonstrated that gels made from polymerized acrylamide and bisacrylamide (‘polyacrylamide gels’) had sufficient resolving power to separate high-molecular-weight pieces of RNA.

Despite the broad utility of the electrophoretic techniques, it was still relatively difficult to separate extremely large pieces of DNA — for example, whole chromosomes — from one another. In the mid-1980s, Schwartz and Cantor described a new approach, named ‘pulse-field gradient gel electrophoresis,’ which used short pulses from perpendicular electrical fields to separate large pieces of DNA. Pulse-field gel electrophoresis has since allowed biologists to undertake massive genotyping studies, as well as molecular epidemiological analyses of pathogens.

These four articles opened the door to numerous other DNA technologies described in this collection; many common molecular biology techniques would not be possible if there were not robust methods for rapidly purifying and analysing nucleic acids of various sizes. It is hard to imagine a modern molecular biology laboratory that does not use some kind of electrophoretic technique on a regular basis. Who could have predicted that spotting hydrolysed ribonucleic acids on Whatman paper would pave the way for the genomic era?

Joshua M. Finkelstein, Senior Editor, Nature

> “These four articles opened the door to numerous other DNA technologies described in this collection...”

ORIGINAL RESEARCH PAPERS Markham, R. & Smith, J. D. The structure of ribonucleic acids. I. Cyclic nucleotides produced by ribonuclease and by alkaline hydrolysis. *Biochem. J.* **52**, 552–557 (1952) | Smithies, O. Zone electrophoresis in starch gels: group variations in the serum proteins of normal human adults. *Biochem. J.* **61**, 629–641 (1955) | Loening, U. E. The fractionation of high-molecular-weight ribonucleic acid by polyacrylamide-gel electrophoresis. *Biochem. J.* **102**, 251–257 (1967) | Schwartz, D. C. & Cantor, C. R. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**, 67–75 (1984)

MILESTONE 2

The dawn of recombinant DNA

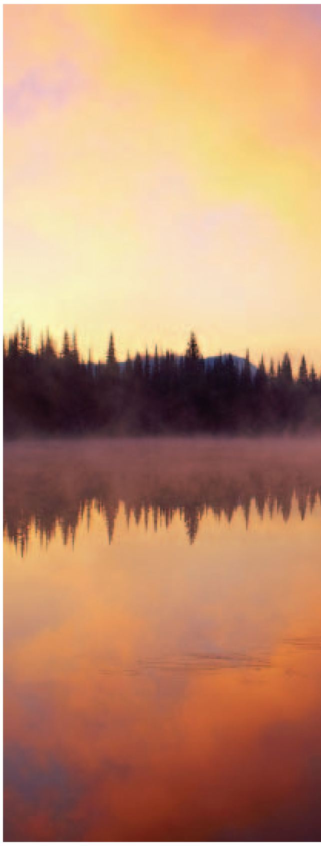
The ability to make recombinant DNA molecules is the cornerstone of modern molecular biology. Yet 40 years ago, it was hardly a conceivable accomplishment.

In the 1960s, biologists had realized that DNA recombination happens in the cell — for example, when breaks caused by ultraviolet irradiation are repaired — and the search for an enzyme that could join DNA molecules was on. The breakthrough came at the beginning of 1967, when Martin Gellert at the National Institutes of Health showed that *Escherichia coli* extracts could convert λ phage DNA 'hydrogen-bonded circles' into a covalently circular form. Within 6 months, Gellert and three other groups independently purified the enzymatic activity, which formed phosphodiester bonds between DNA ends held by hydrogen-bond pairing in a double-stranded configuration.

DNA ligase, which was the first ingredient for making recombinant DNA, was then at hand, but other ingredients, like restriction enzymes, (see Milestone 4) remained to be discovered. Another key concept was the use of plasmids as vectors for shuttling DNA into bacteria. Stanley Cohen, who was studying the role of plasmids in bacterial resistance to antibiotics at Stanford University, first worked out a 'transformation' method to make bacteria take up purified plasmid DNA.

Then, in 1973, Cohen and his Stanford colleague Annie Chang, in collaboration with Herbert Boyer and Robert Helling at the University of California in San Francisco, reported the first *in vitro* construction of a bacterial plasmid. Using the restriction enzyme *EcoRI*, they generated fragments from two plasmids (each conferring resistance to one antibiotic), joined them using DNA ligase and applied the mixture to transform *E. coli*. As they had hoped, a fraction of the transformed bacteria became resistant to both antibiotics while carrying a single hybrid plasmid. Not only had they demonstrated that bacterial plasmids constructed *in vitro* were functional in bacteria, but they had also described the first plasmid vector.

Meanwhile, Paul Berg had devised a similar experiment to transfer foreign DNA into mammalian cells, using the tumour virus SV40 as a vector. In 1972, he made a hybrid molecule *in vitro* by inserting λ phage sequences into SV40. These reports immediately raised concerns, as *E. coli*, which is a natural habitant of the human gut, could now carry hybrid DNA molecules containing SV40 oncogenes or other potentially harmful sequences. These fears led the community to a self-imposed


CORBIS

MILESTONE 3

Fully cooked FISH

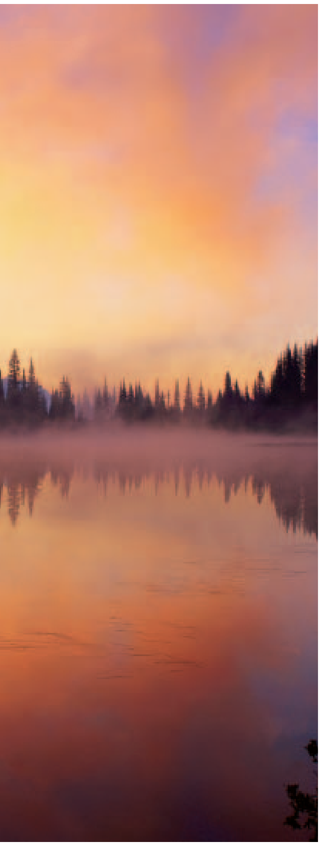
Fluorescence *in situ* hybridization (FISH) has become a methodological mainstay in many fields. It permits the detection, quantification and localization of genes and RNA at resolutions ranging from single nucleotides to whole chromosomes, or even whole cells, using a fluorescently labelled complementary DNA or RNA probe. Remarkably, while immunofluorescence detection of protein has been around since 1949, and immunofluorescent stains for nucleic acids soon followed, FISH was not 'fully cooked' until the early 1980s.

The principal development that led to FISH as we know it today was the determination of the sample fixation and permeabilization conditions necessary to fix cells in their native structure, while allowing the introduction and specific annealing of complementary labelled nucleic-acid molecules. A system that would permit easy testing and validation of the methods was also needed.

In the 1960s, Joe Gall and Mary Lou Pardue at Yale University were studying the extrachromosomal amplification of ribosomal RNA genes in *Xenopus laevis*. This system conveniently provides a known sequence at a high copy number and in an easily identifiable region of the cell. They were able to prepare tritiated complementary RNA of adequate purity and activity to determine the conditions permitting clear visualization of the extrachromosomally amplified ribosomal DNA. They predicted that labelling the RNA at a higher specificity would permit the detection of non-duplicated genes in chromosomes.

Pardue later went to work with members of the Max Birnstiel laboratory. Using *Drosophila melanogaster* polytene chromosomes, they could visualize the location of histone genes on individual chromosomes. However, the use of radioactivity, high background levels and long exposure times were problematic, and multiplexing was impossible. It was not long before fluorescently labelled nucleic acids provided the final crucial tool that gave us the sensitive multiplexing versions of FISH we have today.

Two different methods for fluorescently labelling nucleic acids are commonly used for FISH. The 'direct' method, first described by P. van Duijn and colleagues, relies on direct labelling of the nucleic acid with fluorophores and was the first



moratorium on recombinant DNA experiments. However, the foundation had been laid and progress soon resumed.

Veronique Kiermer, Chief Editor,
Nature Methods

ORIGINAL RESEARCH PAPERS Gellert, M. Formation of covalent circles of λ DNA by *E. coli* extracts. *Proc. Natl Acad. Sci. USA* **57**, 148–155 (1967) | Cohen, S. N., Chang, A. C. Y. & Hsu, L. Nonchromosomal antibiotic resistance in bacteria: genetic transformation of *E. coli* by R-factor DNA. *Proc. Natl Acad. Sci. USA* **69**, 2110–2114 (1972) | Cohen, S. N., Chang, A. C. Y., Boyer, H. W. & Helling, R. B. Construction of biologically functional bacterial plasmids *in vitro*. *Proc. Natl Acad. Sci. USA* **70**, 3240–3244 (1973) | Jackson, D. A., Symons, R. H. & Berg, P. Biochemical method for inserting new genetic information into DNA of simian virus 40. *Proc. Natl Acad. Sci. USA* **69**, 2904–2909 (1972)
FURTHER READING Gefter, M. L., Becker, A. & Hurwitz J. The enzymatic repair of DNA, I. Formation of circular λ DNA. *Proc. Natl Acad. Sci. USA* **58**, 240–247 (1967) | Olivera, B. M. & Lehman, I. R. Linkage of polynucleotides through phosphodiester bonds by an enzyme from *E. coli*. *Proc. Natl Acad. Sci. USA* **57**, 1426–1433 (1967) | Weiss, B. & Richardson, C. C. Enzymatic breakage and joining of deoxyribonucelic acid. *Proc. Natl Acad. Sci. USA* **57**, 1021–1028 (1967) | Zimmerman, S. B., Little, J. W., Oshinsky, C. K. & Gellert, M. Enzymatic joining of DNA strands: a novel reaction of diphosphopyridine nucleotide. *Proc. Natl Acad. Sci. USA* **57**, 1841–1848 (1967)

implementation of FISH. The 'indirect' method, developed later by the David Ward group and implemented in commonly used FISH kits, employs immunogenic or enzymatic detection of tagged nucleic-acid probes following hybridization.

As improvements to the 'recipe' continue to be made, FISH is finding its way onto the plate of increasing numbers of researchers and promises to do so for the foreseeable future.

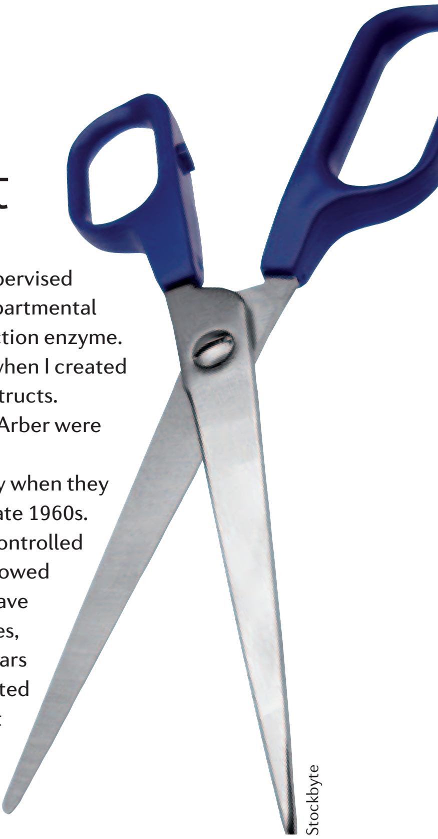
Daniel Evanko, Senior Editor,
Nature Methods

> “Remarkably, while immunofluorescence detection of protein has been around since 1949… FISH was not 'fully cooked' until the early 1980s.”

ORIGINAL RESEARCH PAPERS Gall, J. G. & Pardue, M. L. Formation and detection of RNA–DNA hybrid molecules in cytological preparations. *Proc. Natl Acad. Sci. USA* **63**, 378–383 (1969) | Pardue, M. L., Kedes, L. H., Weinberg, E. S. & Birnstiel, M. L. Localization of sequences coding for histone messenger RNA in the chromosomes of *Drosophila melanogaster*. *Chromosoma* **25**, 135–151 (1977) | Bauman, J. G., Wiegant, J., Borst, P. & van Duijn, P. A new method for fluorescence microscopical localization of specific DNA sequences by *in situ* hybridization of fluorochrome labelled RNA. *Exp. Cell. Res.* **128**, 485–490 (1980) | Langer, P. R., Waldrop, A. A. & Ward, D. C. Enzymatic synthesis of biotin-labeled polynucleotides: novel nucleic acid affinity probes. *Proc. Natl Acad. Sci. USA* **78**, 6633–6637 (1981)

MILESTONE 4

Making the cut



Stockbyte

I remember clearly the first time I made a supervised trip, as an undergraduate student, to the departmental freezer to obtain a precious aliquot of restriction enzyme. Its real value, however, only dawned on me when I created the first of countless recombinant DNA constructs.

It is unlikely that Stuart Linn and Werner Arber were aware of the far-reaching consequences of their discovery when they stumbled across restriction enzymes in the late 1960s. While studying a phenomenon called host-controlled restriction of bacteriophage growth, they showed that restriction enzymes of the host cells cleave unmethylated phage DNA in numerous places, thereby limiting their growth. A couple of years later, Hamilton Smith and Kent Wilcox reported the isolation and characterization of the first restriction enzyme — endonuclease R (later renamed HindII) — from extracts of *Haemophilus influenzae* strain Rd. Importantly, the enzyme degraded foreign DNA, such as that of phage T7, but did not affect native *H. influenzae* DNA.

Smith and Wilcox demonstrated that endonuclease R produces double-stranded 3′-hydroxyl, 5′-phosphoryl cleavage products. They proposed that the enzyme recognizes a specific sequence on the foreign DNA, and estimated from the number of breaks that the site would have to be five or six bases in length. Smith, together with Thomas Kelly, determined the recognition sequence using end-labelling techniques. This was an exceptional technical feat, as there was no method at the time for the analysis of terminal sequences beyond the dinucleotide level. They postulated that the internal symmetry of the recognition sequence, which was cleaved in the middle, was not surprising given that the enzyme carries out a symmetrical reaction on opposite strands.

Before long, Kathleen Danna and Daniel Nathans pioneered the application of restriction enzymes. They used endonuclease R to characterize the small oncogenic DNA virus SV40: the resulting 11 fragments were resolved by polyacrylamide gel electrophoresis, and their molecular weights were determined. Their prediction that restriction-enzyme analysis would be useful to map a genome region and to localize specific genes by testing for biological activity turned out to be visionary.

The 'recombination' potential of restriction enzymes was first demonstrated by Janet Mertz and Ronald Davis. They showed that the R1 restriction endonuclease produces 'staggered' breaks, generating 'cohesive' ends that are identical and complementary. Their findings suggested that any R1-generated ends can be joined by incubation with DNA ligase to generate hybrid DNA molecules. Thus, the era of recombinant DNA technology was born.

Arianne Heinrichs, Chief Editor,
Nature Reviews Molecular Cell Biology

ORIGINAL RESEARCH PAPERS Smith, H. O. & Wilcox, K. W. A restriction enzyme from *Hemophilus influenzae*. I. Purification and general properties. *J. Mol. Biol.* **51**, 379–391 (1970) | Kelly, T. J. Jr & Smith, H.O. A restriction enzyme from *Hemophilus influenzae*. II. Base sequence of the recognition site. *J. Mol. Biol.* **51**, 393–409 (1970) | Danna, K. & Nathans, D. Specific cleavage of simian virus 40 DNA by restriction endonuclease of *Hemophilus influenzae*. *Proc. Natl Acad. Sci. USA* **68**, 2913–2917 (1971)
FURTHER READING Linn, S. & Arber, W. Host specificity of DNA produced by *Escherichia coli*, X. *In vitro* restriction of phage fd replicative form. *Proc. Natl. Acad. Sci. USA* **59**, 1300–1306 (1968) | Mertz, J. E. & Davis, R. W. Cleavage of DNA by R1 restriction endonuclease generates cohesive ends. *Proc. Natl Acad. Sci. USA* **69**, 3370–3374 (1972)

MILESTONE 5

A reverse proves to be an advance

The central dogma of molecular biology states that DNA makes RNA makes protein, yet one of the most important DNA technologies stemmed from the discovery that the first of these steps can be reversed.

In 1970, puzzled by the ability of RNA tumour viruses to stably transform cells — without incorporation of a DNA copy of viral genes into the host genome — Baltimore, and Temin and Mizutani, looked for DNA polymerase activity in purified preparations of such viruses. The kinetics of the incorporation of radiolabelled thymine indicated that DNA was being synthesized. The reaction was sensitive to ribonuclease treatment, showing that it was RNA-dependent, whereas the product was deoxyribonuclease-sensitive but ribonuclease-insensitive, confirming that it was DNA.

Two years later, it was demonstrated that the reverse transcriptase could be used *in vitro* to synthesize cDNA from mammalian mRNAs. Verma *et al.* and Kacian *et al.* both added preparations of globin mRNAs to reverse transcriptase from avian myeloblastosis virus. They correctly hypothesized that the reaction would only work efficiently if they also added oligo(dT), which would hybridize to the poly(A) tail of the mRNAs and act as a primer. By hybridizing their DNA product to the original mRNA template, they confirmed that they had successfully synthesized cDNA.

Over the next two decades, reverse transcriptase was widely used for the cloning of expressed genes. However, one of the biggest technical hurdles, especially for the creation of comprehensive libraries, was that most cDNAs in a given reaction were not full length owing to premature termination. In addition, low-abundance transcripts were far less likely to be cloned than high-abundance ones. In the late 1990s, Carninci, Hayashizaki and colleagues developed several techniques to overcome these problems.

By biotin capping of the mRNAs, they ensured that only full-length cDNAs were selected. After first-strand cDNA synthesis, RNAse I was used to destroy any part of any mRNA that was not bound to cDNA. This caused the removal of the 5′ biotin cap from the mRNAs of all non-complete cDNAs. Magnetic beads were used to select only the full-length cDNAs for second-strand synthesis and cloning.

Their discovery that trehalose makes reverse transcriptase more thermostable meant that the reaction could be carried out at a higher temperature, at which the formation of fewer RNA secondary structures increased the number of full-length cDNAs produced.

Finally, in order to complete expression libraries, they selectively cloned rare new cDNAs by screening out abundant ones and those already cloned in existing libraries. To achieve this, they added biotinylated RNA from the original sample and existing libraries after first-strand

MILESTONE 6

Southern migration

It was a simple but clever idea that, in 1975, led Edwin Southern to invent a method that carries his name and that has revolutionized the study of DNA.

In the early 1960s, Julius Marmur and Paul Doty published a study that accurately described the conditions for the optimal renaturation of DNA complementary strands, upon denaturation by high temperatures. They proposed that high temperature was required to block the formation of weak bonds between non-complementary strands and to guarantee the proper pairing of complementary molecules.

Given these findings and the intrinsic features of DNA molecules, it became evident that a specific DNA fragment could be identified from a biological sample by letting it hybridize, following denaturation, to a radiolabelled complementary molecule. An RNA strand, for example, could be used as a probe to gain insight into gene structure and function. The hybridization could even occur when the DNA was trapped onto a nitrocellulose membrane.

At that time, DNA fragments could be obtained by digestion with the recently discovered restriction enzymes (see Milestone 4) and by separation through electrophoresis in agarose gel (see Milestone 1). However, DNA recovery from the gel inevitably led to loss of material and a notable decrease in the resolving power of electrophoresis.

Southern had a genial intuition that he could transfer the DNA fragments directly from the gel onto a nitrocellulose membrane laid on top of it. The agarose gel being permeable, he could force liquid to pass through the gel by piling up, on top of the nitrocellulose, a stack of dry filter paper. Drawn by the flowing liquid, the DNA fragments could be soaked out of the gel. Following hybridization with radiolabelled RNA, autoradiography of the membrane revealed the specific fragments containing the sequence of interest, which appeared as sharp bands in the same position as they had been on the gel. It was finally possible to detect a specific DNA sequence from a smear, without having to purify it away from the rest of the genome.

Southern used his method to study bacterial and eukaryotic ribosomal genes, but its practice soon became widespread. It facilitated the study synthesis. All of the cDNA that hybridized to the RNA was removed with magnetic beads, leaving rare cDNAs behind.

From its origin as an esoteric property of certain viruses, reverse transcription has become hugely important in molecular biology. Its influence extends from cloning to the development of microarrays to the annotation of genomes.

Patrick Goymer, Associate Editor,
Nature Reviews Genetics

ORIGINAL RESEARCH PAPERS Baltimore, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* **226**, 1209–1211 (1970) | Temin, H. M. & Mizutani, S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* **226**, 1211–1213 (1970) | Verma, I. M. *et al. In vitro* synthesis of DNA complementary to rabbit reticulocyte 10S RNA. *Nature New Biol.* **235**, 163–167 (1972) | Kacian, D. L. *et al. In vitro* synthesis of DNA components of human genes for globins. *Nature New Biol.* **235**, 167–169 (1972) | Carninci, P. *et al.* High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37**, 327–336 (1996) | Carninci, P. *et al.* Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. *Proc. Natl Acad. Sci. USA* **95**, 520–524 (1998) | Carninci, P. *et al.* Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10**, 1617–1630 (2000)

of gene structures and, only a few years later, the discovery of genetic defects (such as the loss of a restriction enzyme site underling sickle cell anaemia). Nowadays, the applications of the Southern blot span from basic to biomedical research, and from genetic engineering to forensics, yet the protocol remains surprisingly similar to that described by Southern over 30 years ago.

The method also inspired others to adopt a similar strategy for the study of different molecules. To emphasize the similarity with the Southern blot, the transfer of RNA and protein from a gel to a solid support were named northern and western blot, respectively.

Francesca Pentimalli, Assistant Editor,
Nature Reviews Cancer and Nature
Reviews Genetics

ORIGINAL RESEARCH PAPERS
Marmur, J. & Doty, P. Thermal renaturation of deoxyribonucleic acids. *J. Mol. Biol.* **3**, 585–594 (1961) | Southern, E. M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**, 503–517 (1975)

MILESTONE 7

Deciphering the code

A fundamental cornerstone of the era of "genetic engineering" was the development of technology that allowed researchers to determine a DNA sequence in its linear order. Given the advent of electrophoresis (see Milestone 1), the questions became how could one generate fragments at every position and how could the terminal base of the fragments be distinguished?

Three methods were revolutionary in achieving these goals. The first, cleaving single-stranded DNA (ssDNA) or denatured double-stranded DNA (dsDNA) labelled at one end, was published by Maxam and Gilbert in 1977. Fragments were generated in two steps: the base was removed, and then the weakened sugar bond was broken by the addition of alkali or amines. In this chemical method, the four bases were not specifically determined; rather, one of the four reactions cleaved at pyrimidines, one cleaved at C, and the other two had a preference for cleaving G > A or A > G. The four pools of fragments were separated in different lanes of a polyacrylamide gel and the sequence, with a read length of ~100 bases, was deduced from the ladder of bands.

Later that year, Sanger and colleagues published an enzymatic sequencing protocol. This method did not involve DNA breakage, but exploited the fact that dideoxy-nucleotides (ddNTPs), lacking a 3′-hydroxyl, cannot be extended by DNA polymerase. Consequently, one could set up four DNA synthesis reactions containing the same ssDNA template and primer, DNA polymerase, a mixture of the deoxynucleotide (dNTP) and ddNTP forms of one of the nucleotides, and the remaining three dNTPs (one of which was labelled). As both the ddNTP and dNTP were present, DNA polymerase would sometimes incorporate the correct dNTP, allowing further polymerization, and at other times would incorporate the ddNTP, causing chain termination. After DNA denaturation and polyacrylamide gel electrophoresis, a series of labelled bands corresponding to termination at a specific nucleotide could be read.

It was appreciated that in order to sequence genomes, it would be necessary to automate sequencing and gather information in real time. Into this void stepped the Hood laboratory with a third technique, a variation of the Sanger method that bypassed the rate-limiting step — detection of the labelled bands by autoradiography. Rather than using labelled dNTPs, the primer oligonucleotides in the four reactions were attached to fluorophores with different emission maxima. The four reactions yielded fragments tagged with different fluorophores, so that they could be combined and run in a single column gel. Simultaneously, the group developed a detection system that used a laser to read the fluorophore signals, and an analytical program that resolved the raw data into a series of peaks that corresponded to the sequence. When a single lane contained all four reactions, 200 bases were read with ease. Subsequently the fluorophores were attached to ddNTP terminators, removing the need for tagged primers and allowing all four reactions to be performed in one reaction, slab gels were replaced with capillaries and read lengths were extended to 600 or more bases.

These approaches facilitated the explosion of sequence-gathering studies that have evolved into our current bioinformatics-driven research.

Angela K. Eggleston,
Senior Editor, Nature

ORIGINAL RESEARCH PAPERS Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA* **74**, 560–564 (1977) | Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977) | Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986)

WEB SITE
DNA sequencing from Wikipedia:
http://en.wikipedia.org/wiki/DNA_sequencing

MILESTONE 8

Get out the map

Although the existence of variable loci in human DNA had been appreciated for some time, it was not until the late 1970s and early 1980s that a way to use polymorphisms for large-scale systematic mapping of human genes was proposed.

David Botstein, Raymond White, Mark Skolnick and Ronald Davis argued that a large number of DNA sequence polymorphisms must exist in the human population, and that some of these should be detectable as variants in the length of DNA fragments produced by restriction enzymes (restriction fragment-length polymorphisms or RFLPs). These RFLPs could be detected using Southern blotting experiments on human genomic DNA. Importantly, and unlike classical polymorphic antigenic and enzyme markers, these new loci could be identified in non-coding regions of the genome as well as within genes. Linkage relationships among RFLPs could be established using pedigrees, and genetic linkage to a locus of interest would allow a gene to be mapped and defined, even if the RFLPs were not in the gene. Botstein and colleagues estimated that at least 150 highly polymorphic regions at regular intervals in the human genome would make it feasible to construct a human genetic-linkage map and to localize disease genes.

The first practical demonstration of this came in 1983 with the mapping of the gene for Huntington disease. As proposed by Botstein and colleagues, a large number of recombinant DNA probes defining RFLPs in human DNA had by then been identified. In a collaborative effort involving researchers in the United States and Venezuela, James Gusella and colleagues screened the DNA of members of two Huntington families with 12 such probes, and identified one that defined a marker on chromosome 4 with close linkage to the disease locus.

A new field, positional cloning of genetic disease loci, was born. RFLPs were later involved in the mapping and positional cloning of the cystic fibrosis gene by John Riordan and colleagues. Early linkage studies with RFLPs also allowed the creation of the first genome-wide genetic maps, which, together with sequence-tagged sites introduced by Maynard Olson and colleagues, allowed the construction of sequence-based physical maps of genomes.

Natalie de Souza, Associate Editor,
Nature Methods

ORIGINAL RESEARCH PAPERS Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980) | Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983)
FURTHER READING Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–1073 (1989) | Olson, M., Hood, L., Cantor, C. & Botstein, D. A common language for physical mapping of the human genome. *Science* **245**, 1434–1435 (1989)

CORBIS

MILESTONE 9

Transformers, elements in disguise

In the post-genomic era, understanding biological processes increasingly relies on analysis of gene functions. Yet, until a few decades ago, studying eukaryotic gene function *in vivo* was impossible, as no efficient and reproducible procedures to transfer DNA into eukaryotic cells were available.

In the late 1970s, Gerald Fink and colleagues set the basis for studying eukaryotic genes by establishing a transformation protocol to introduce exogenous DNA into yeast cells permanently. A few years later, Mario Capecchi showed that microinjection of the herpes simplex virus gene encoding thymidine kinase into the nuclei of mammalian cells lacking this enzyme allowed the kinase activity to be recovered.

It was only in 1982, however, that DNA was successfully manipulated *in vivo* in a higher organism. Allan Spradling and Gerald Rubin characterized P-elements — mobile DNA elements — through analysis of *Drosophila melanogaster* strains that gave rise to progeny suffering from the hybrid dysgenesis genetic syndrome. They identified two groups of P-elements that differed in size and ability to move within the genome. Injecting 3-kb autonomous P-elements into *Drosophila* embryos lacking them revealed that the elements could insert into random genomic positions, inducing mutations in a fraction of the progeny. These findings initiated the use of P-elements in large-scale mutagenesis screens in *Drosophila*, given the advantage of gene cloning in the identified mutants.

In mice, site-directed mutagenesis was first used in 1987, when two research teams targeted the gene encoding hypoxanthine phosphoribosyl transferase (*Hprt*) by homologous recombination in embryonic stem (ES) cells. These cells were chosen for their unique potential to be manipulated *in vitro* and reintroduced into mouse blastocysts, producing chimeric animals that can transmit the new traits to subsequent generations. Kirk Thomas and Mario Capecchi engineered two classes of vectors that efficiently disrupted *Hprt* either by replacing endogenous sequences with the exogenous neomycin-resistance gene or by inserting the exogenous sequence into the *Hprt* locus. They then identified ES cells carrying mutated genes by selecting for acquired resistance to the drug G418 and the base analogue 6-TG. Oliver Smithies and co-workers also used this technique to correct the defects of three independent *Hprt*-mutated ES cell lines. Together, these groundbreaking studies paved the way for functional genomics and gene therapy.

Francesca Cesari,
Locum Associate Editor, Nature

ORIGINAL RESEARCH PAPERS Rubin, G. M. & Spradling, A. C. Genetic transformation of *Drosophila* with transposable element vectors. *Science* **218**, 348–353 (1982) | Spradling, A. C. & Rubin, G. M. Transposition of cloned P elements into *Drosophila* germ line chromosomes. *Science* **218**, 341–347 (1982) | Doetschman, T. *et al.* Targetted correction of a mutant *HPRT* gene in mouse embryonic stem cells. *Nature* **330**, 576–578 (1987) | Thomas, K. R. & Capecchi, M. R. Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell* **51**, 503–512 (1987)
FURTHER READING Hinnen, A. *et al.* Transformation of yeast. *Proc. Natl Acad. Sci. USA* **75**, 1929–1933 (1978) | Capecchi, M. R. High efficiency transformation by direct microinjection of DNA into cultured mammalian cells. *Cell* **22**, 479–488 (1980)

MILESTONE 10

Randomness versus order

Whereas randomness is avoided in most experimental techniques, it is fundamental to sequencing approaches. In the race to sequence the human genome, research groups had to choose between the random whole-genome shotgun sequencing approach or the more ordered map-based sequencing approach.

When Frederick Sanger and colleagues sequenced the 48-kb bacteriophage λ genome in 1982, the community was still undecided as to whether directed or random sequencing strategies were better. With directed strategies, DNA sequences were broken down into ordered and overlapping fragments to build a map of the genome, and these fragments were then cloned and sequenced. With the shotgun approach, DNA sequences were broken randomly, cloned, sequenced and then pieced together by analysing the overlap. Sanger *et al.* compared these strategies while sequencing bacteriophage λ and reported that the random approach was faster than any directed method.

One problem with the random approach, however, was that of filling gaps when the sequence was nearly complete (or closure), as randomly selected clones were often redundant. For instance, Sanger *et al.* used — in their opinion mistakenly — direct sequencing strategies to finish the last 10% of the bacteriophage λ sequence. In 1991, Al Edwards and Thomas Caskey proposed a method to maximize efficiency by minimizing gap formation and redundancy: sequence both ends (but not the middle) of a long clone, rather than the entirety of a short clone.

Although the shotgun approach was now accepted for sequencing short stretches of DNA, map-based techniques were still considered necessary for large genomes. Like the directed strategies, map-based sequencing subdivided the genome into ordered 40-kb fragments, which were then sequenced using the shotgun approach. In 1995, however, Robert Fleischmann and colleagues used a whole-genome shotgun approach to sequence the 1,800-kb genome of *Haemophilus influenzae* — the first complete genome of a free-living organism. The authors had randomly generated large 40-kb fragments and had thereby bypassed the mapping stage. In doing so, they had proved that genome-assembly programmes that matched overlap were reliable and that whole-genome shotgun sequencing worked, in principle.

The *H. influenzae* genome, however, was a mere DNA fragment compared with the 1,500-fold longer ~3 billion base-pair human genome. In 1996, Craig Venter and colleagues proposed that the whole-genome shotgun approach could be used to sequence the human genome owing to two factors: its past successes in assembling genomes and the development of bacterial artificial chromosomes (BAC) libraries, which allowed large fragments of DNA to be cloned.

A showdown ensued, with the biotechnology firm Celera Genomics wielding whole-genome shotgun sequencing and the International Human Genome Sequencing Consortium wielding map-based sequencing. Yet, when the dust settled, it was a draw — both groups published their initial drafts of the human genome concurrently in 2001.

Asher Mullard, Assistant Editor,
Nature Reviews Microbiology and
Nature Reviews Molecular Cell Biology

ORIGINAL RESEARCH PAPERS Sanger, F. *et al.* Nucleotide sequence of bacteriophage λ DNA. *J. Mol. Biol.* **162**, 729–773 (1982) | Edwards, A. & Caskey, C. T. Closure strategies for random DNA sequencing. *Methods* **3**, 41–47 (1991) | Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995) | Venter, J. C. *et al.* A new strategy for genome sequencing. *Nature* **381**, 364–366 (1996)

> “In the race to sequence the human genome, research groups had to choose between the random whole-genome shotgun sequencing approach or the more ordered map-based sequencing approach.”

MILESTONE 11

Chain reaction

Daniel Jones

Today, it would be difficult to conceive of a biology laboratory without a polymerase chain reaction apparatus—‘the PCR machine’. Standard molecular practices that we take for granted, such as creating constructs for expressing tagged proteins, amplifying genes from minute samples for cloning and introducing mutations into genes, would be a completely different story, a much more complicated one, without the advent of PCR.

This technique, which has revolutionized biological laboratory research, was first developed about two decades ago. Yet in 1971, in a *Journal of Molecular Biology* report, Har Gobind Khorana and colleagues had already described a process called repair replication for synthesizing short DNA duplexes and single-stranded DNA by polymerases. This report outlined several features that are hallmarks of PCR, but fell short of an experimental test. It predicted, for example, that ▶

the DNA duplex would have to be denatured to single strands, that an excess of primer to template would be required to overcome secondary structures generated by single-stranded template and that, following completion of the reaction by DNA polymerase, the cycle would have to be reinitiated if the template duplex had renatured.

However, it took another almost 17 years before PCR, as we know it, was described. Kary Mullis and Fred Faloona, in a 1987 paper in *Methods in Enzymology*, experimentally defined the basic steps of PCR. The authors speculated about the potential applications of the method, many of which are now routine molecular biology procedures. Interestingly, 2 years before this landmark report was published, Norman Arnheim and colleagues demonstrated the power of PCR as a diagnostic tool by showing that such an approach could be used to rapidly amplify the β-globin sequence from clinical samples to determine whether it possessed the mutation for sickle cell anaemia.

It was not until 1988 that Henry Erlich and colleagues described the use of a thermostable DNA polymerase from *Thermus aquaticus*, which was a key innovation that allowed annealing and extension at high temperatures. This crucial improvement did away with the need to replenish the enzyme after each cycle, increased specificity, yield and sensitivity, allowed the process to be used to generate longer products and, finally, paved the way for its automation. PCR, as we know it, was born.

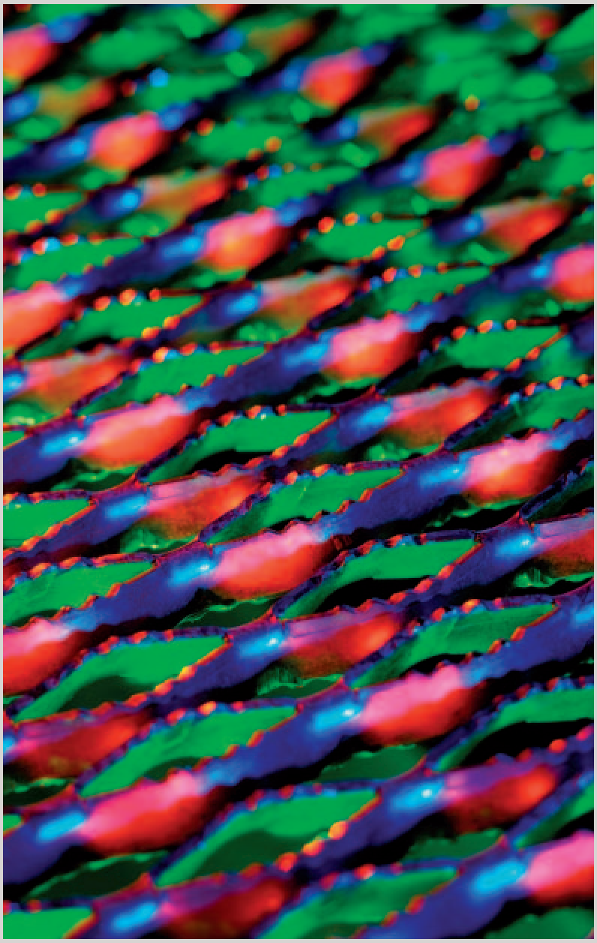
Sowmya Swaminathan, Senior Editor,
Nature Cell Biology

ORIGINAL RESEARCH PAPERS Kleppe, K., Ohtsuka, E., Kleppe, R., Molineux, I. & Khorana, H. G. Studies on polynucleotides. XCVI. Repair replications of short synthetic DNAs as catalyzed by DNA polymerases. *J. Mol. Biol.* **56**, 341–361 (1971) | Saiki, R. K. *et al.* Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354 (1985) | Mullis, K. B. & Faloona F. A. Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. *Meth. Enzymol.* **155**, 335–350 (1987) | Saiki, R. K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491 (1988)

> “Today, it would be difficult to conceive of a biology laboratory without a polymerase chain reaction apparatus—‘the PCR machine’.”

MILESTONE 12

A repeat success



Following the discovery of the DNA structure in the 1950s, attention gradually shifted to the next big thing: finding those parts of the human genome that differed between individuals. In 1985, Alec Jeffreys, Victoria Wilson and Swee Lay Thein described a highly variable segment of DNA that would help in this quest: the minisatellite. This was not only a sensitive tool for human and other genetic studies but it also had applications in person identification, which opened it up for use in paternity analysis, immigration disputes and forensic science.

Molecular markers were already being used in linkage studies, for example, and in antenatal diagnosis. Yet the variability of these at-best dimorphic markers — made of DNA segments with variable length that were created by endonucleases — would not stretch to give the resolution needed to distinguish individuals easily.

In 1984, a tandemly arranged 33 base-pair sequence had been detected in an intron of the human myoglobin gene; the basis of the 1985 paper was the realization that a probe derived from this kind of repeat could pick up many variable-length segments in the genome simply by Southern blotting. Because of their repeated nature such sequences, or minisatellites, are prone to expansion and contraction, and so each locus can vary among individuals. Each minisatellite has a ‘core’ motif of 6–100 base pairs: if more than one core sequence was used then a ‘profile’ or ‘fingerprint’ of an individual could be obtained that was, in effect, unique — as the paper reported, this allowed even close relatives to be unambiguously identified.

The authors were aware of the wide potential applications of ‘DNA fingerprinting’, although it is unlikely that they had predicted its runaway success in fields ranging from conservation biology to forensics. Minisatellites made their first appearance in court that same year, in an immigration dispute over the identity of a young boy returning to the UK from Ghana. Minisatellite profiling of the boy’s alleged mother and three siblings confirmed the relationship claim made by the defence, and he was granted permission to remain in the country.

Today, DNA fingerprinting is alive and well; if anything, it has grown in recognition. However, the name is one of the few things that have been preserved, given the large changes to the field brought about by new markers and technologies — not least the invention of PCR.

Tanita Casci, Senior Editor,
Nature Reviews Genetics

ORIGINAL RESEARCH PAPERS Jeffreys, A. J., Wilson, V. & Thein, S. L. Hypervariable ‘minisatellite’ regions in human DNA. *Nature* **314**, 67–73 (1985) | Jeffreys, A. J., Brookfield, J. F. & Semeonoff, R. Positive identification of an immigration test-case using human DNA fingerprints. *Nature* **317**, 818–819 (1985)

MILESTONE 13

Size matters

The year 1987 marked a quantum leap for DNA cloning with a tenfold increase in vector capacity. A vector carrying a 50-kb insert, the largest available at the time, was useful for examining a single gene but far too small to also contain all regulatory regions. There was also growing interest in constructing comprehensive libraries covering the whole genomes of higher organisms — a supremely daunting task if undertaken in 50 kb fragments.

To increase capacity, Maynard Olson and colleagues at Washington University exchanged the classical cloning vehicle, a circular *Escherichia coli* plasmid, for what they called a yeast artificial chromosome (YAC): a linear DNA molecule that mimics a yeast chromosome, complete with centromere and telomeres. All necessary YAC elements were incorporated into circular plasmids that could be linearized *in vitro* during insertion of the exogenous DNA. The resulting linear fragment, carrying as much as several hundred kb of foreign DNA, faithfully replicated in yeast.

YACs were eagerly adopted by the scientific community, particularly to establish physical maps of whole genomes. Yet, with their increased use, some problems surfaced: many clones turned out to be chimaeras of noncontiguous DNA fragments, inserts were occasionally unstable and purification of YACs proved challenging due to contamination from endogenous yeast chromosomes. People started to miss the simplicity of a bacterial cloning system.

In 1992, the time was ripe to revisit *E. coli* with the aim of adapting it for large-fragment cloning. A group at the California Institute of Technology led by Melvin Simon modified an endogenous circular plasmid in *E. coli*, the fertility (F) factor present at one or two copies per cell, to create a cloning vector. In reference to its yeast cousin, they called it bacterial artificial chromosome (BAC). With a cloning capacity of ~300 kb, BACs are not as potent as YACs, but they have all the advantages of a bacterial vector: stability, and ease of manipulation and purification.

Today YACs are still in use; however, BACs have become the workhorses in genomic research for any application that requires large DNA inserts.

Nicole Rusk,
Associate Editor, Nature Methods

ORIGINAL RESEARCH PAPERS Burke, D. T. *et al.* Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236**, 806–811 (1987) | Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragment of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* **89**, 8794–8797 (1992)

MILESTONE 14

ChIPping away at protein–DNA interactions

Neil Smith

We now know that chromatin is not simply an inert packaging structure for DNA, but provides a dynamic environment with an important role in regulating processes such as gene transcription, DNA repair and replication. However, over much of the twentieth century the study of chromatin progressed more slowly than that of DNA, often because appropriate tools were not available.

The chromatin immunoprecipitation (ChIP) assay has become an indispensable tool, but the required techniques took many years to evolve. The assay allows transcription factors, chromatin proteins (such as histones) or post-translational modifications to these proteins to be mapped to specific regions of genomic DNA. A crucial aspect of the ChIP assay is to preserve physiologically relevant interactions between DNA and chromatin proteins, which was made possible by crosslinking methods.

Formaldehyde crosslinking became popular as it works well with histones and is easily reversible. In an early report in 1978, Jackson used formaldehyde crosslinking and electrophoresis to demonstrate histone–DNA and histone–histone interactions in isolated nuclei. Some years later, Varshavsky, Lis and their colleagues published influential papers in which an immunoprecipitation step was introduced with specific histone antibodies, which was performed after protein–DNA complexes had been formaldehyde or ultraviolet crosslinked and sheared. This approach allowed investigators to show, for example, that chromatin at the heat-shock protein 70 promoter of *Drosophila* was perturbed upon heat shock.

▶

A further key advance was the development of selective antibodies, such as those to modified histones, for immunoprecipitating specific protein–DNA complexes. In the late 1980s, ChIP with these antibodies provided a functional link between histone acetylation and transcription.

Later, PCR was used to amplify the DNA purified from the antibody complexes; however, with the advent of DNA microarrays, the ChIP assay was dramatically extended to locate transcription factor binding sites on a genome-wide scale, a technique that acquired the moniker 'ChIP-chip'. In two pioneering papers, Ren, Iyer and their colleagues combined ChIP of transcription factors in yeast with a PCR step to amplify and label the immunoprecipitated DNA before hybridization to microarrays of gene promoters. A genome-wide view of the sites bound by transcription factors under various cellular conditions could now be combined with microarray gene expression analysis to determine the direct function of these factors. Since then, ChIP-chip has also allowed mapping of chromatin proteins and histone modifications across the genome, and so has proved a powerful tool to investigate the influence of chromatin on gene transcription and other DNA processes.

Alex Eccleston, Senior Editor, Nature

ORIGINAL RESEARCH PAPERS Solomon, M. J., Larsen, P. L. & Varshavsky, A. Mapping protein–DNA interactions *in vivo* with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**, 937–947 (1988) | Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000) | Iyer, V. R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001)
FURTHER READING Jackson, V. Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent. *Cell* **15**, 945–954 (1978) | Gilmour, D. S. & Lis, J. T. Detecting protein–DNA interactions *in vivo*: distribution of RNA polymerase on specific bacterial genes. *Proc. Natl Acad. Sci. USA* **81**, 4275–4279 (1984) | Hebbes, T. R., Thorne, A. W. & Crane-Robinson, C. A direct link between core histone acetylation and transcriptionally active chromatin. *EMBO J.* **7**, 1395–1402 (1988) | Braunstein, M., Rose, A. B., Holmes, S. G., Allis, C. D. & Broach, J. R. Transcriptional silencing in yeast is associated with reduced nucleosome acetylation. *Genes Dev.* **7**, 592–604 (1993) | Strahl-Bolsinger, S., Hecht, A., Luo, K. & Grunstein, M. SIR2 and SIR4 interactions differ in core and extended telomeric heterochromatin in yeast. *Genes Dev.* **11**, 83–93 (1997) | Kuo, M. H. & Allis, C. D. *In vivo* cross-linking and immunoprecipitation for studying dynamic protein:DNA associations in a chromatin environment. *Methods* **19**, 425–433 (1999)

MILESTONE 15

BLAST-off for genomes

The generation of genome sequences from a wide range of organisms has opened the field of comparative genomic analysis, assisting the annotation of individual genomes, and bringing new insights into genome evolution. Central to the field of comparative genomics have been programs used to search and align protein or DNA sequences based on a measure of similarity.

Sequence alignment can involve either global alignment, in which the two sequences are aligned over their entire length, or local alignment, in which subregions of the two sequences are aligned. The latter has been more widely applied, as DNA sequences generally show isolated regions of similarity.

An exact solution to the global alignment problem was developed by Saul Needleman and Christian Wunsch in 1970, by applying dynamic programming to find the optimal alignment between two sequences. In 1981, Temple Smith and Michael Waterman extended this dynamic programming approach to solve the local alignment problem. These exact solutions placed sequence comparisons on a firm mathematical grounding, and formed the basis for the early alignment search algorithms.

However, the exact solutions proved slow in practice, especially for searching large databases, spurring the development of faster *heuristic* approaches. One early successful heuristic algorithm to enable efficient searching of large databases, FASTA, was presented by David Lipman and William Pearson in 1988. This simplified the problem by searching

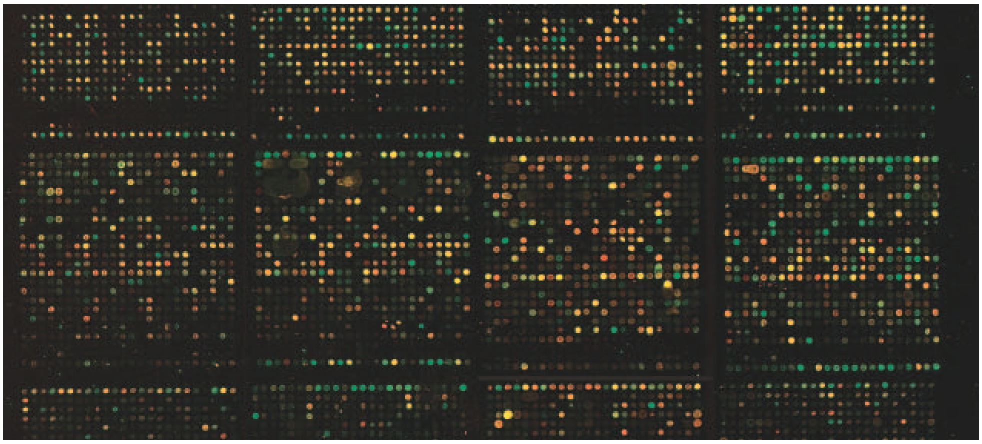
> “…BLAST indexes the query sequence and scans against a database.”

for short regions of exact match and then extending them. In 1990, Stephen Altschul and colleagues presented the basic local alignment search tool (BLAST), which instead searched for all short matches above a given scoring threshold, and showed that this improved speed. In addition, Altschul developed a statistical framework for sequence alignment that provided a conceptual basis for understanding similarity measures, and a method for assessing the statistical significance of a given alignment. BLAST indexes the query sequence and scans against a database, whereas Jim Kent in 2002 showed how the reverse could increase speed (at tolerably reduced sensitivity), with the BLAST-like alignment tool (BLAT).

The sequencing of full-length genomes increased interest in developing tools for genome-wide multiple alignments. Paving the way, in the 1980s, Clustal drew on information in phylogenetic trees to go from pairwise to multiple sequence alignments, although mainly for protein sequences. MUMmer, which in 1999 was one of the first to move to whole-genome multiple alignments, proved useful for aligning bacterial genomes. This was soon followed by BLASTZ and MultiZ from Webb Miller and colleagues; these were useful for cross-species comparisons, allowing alignment and searching of mammalian chromosome-length sequences.

Orli Bahcall, Associate Editor,
Nature Genetics

ORIGINAL RESEARCH PAPERS Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981) | Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. *Gene* **73**, 237–244 (1988) | Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85**, 2444–2448 (1988) | Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990) | Altschul, S. F. & Gish, W. Local alignment statistics. *Methods Enzymol.* **266**, 460–480 (1996) | Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27**, 2369–2376 (1999) | Kent, W. J. BLAT: the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002) | Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003)



MILESTONE 16

The microarray revolution

Imagine attempting to measure the expression level of every gene in the genome one at a time. Now imagine assaying thousands of genes all at once. It was the advent of microarray technology that made possible this leap from low to high throughput, allowing researchers to ask questions on a scale that was previously unattainable.

The landmark study that showed that DNA could be made into microarrays was published in 1991 by Stephen Fodor and colleagues at the Affymax Research Institute. They showed that a diverse set of oligonucleotides could be chemically synthesized on a glass slide through photolithography, a process using precisely aimed beams of light to direct chemical reactions to specific spots. This allows for miniaturization because the density of spots is limited only by the diffraction of light.

A paper published in 1995 by Patrick Brown and colleagues at Stanford University brought attention to the exciting potential of microarray technology. They used a microarray of 45 *Arabidopsis* complementary DNAs to which they hybridized fluorescently-labelled total cellular messenger RNA (mRNA). The intensity of fluorescence at a spot reflected the amount of mRNA hybridized, which in turn reflected the level of that particular mRNA in the initial sample. This paper showed for the first time that the expression of many genes in a small sample could be quantitatively monitored in parallel.

Microarray-based expression profiling has useful applications in medical research, as it reveals molecular portraits of gene expression in disease states. Yet the utility of microarrays is not limited to measuring gene expression. Once the technology became established, researchers began to use microarrays to measure other important biological phenomena. For example, microarrays are being used to genotype single-nucleotide polymorphisms by hybridizing the DNA of individuals to arrays of oligonucleotides representing different polymorphic alleles. An application called array-comparative genomic hybridization is being used to detect genomic structural variation, such as segments of the genome that have varying numbers of copies in different individuals; this is accomplished by hybridizing the total genomic DNA to an array of oligonucleotides representing DNA fragments distributed evenly throughout the genome. Epigenetic marks associated with certain areas of the genome, such as chromatin modifications, are being profiled using microarrays in an application called ChIP-chip (see Milestone 14).

Microarray technology has grown from a pioneering method applied by innovators at the cutting edge to a ubiquitous technique that has allowed researchers to investigate 'big-picture' questions in biology. This miniaturized technology has brought about a major revolution.

Emily Niemitz, Associate Editor,
Nature Genetics

ORIGINAL RESEARCH PAPERS Fodor, S. P. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773 (1991) | Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995)
FURTHER READING Southern, E. M. *et al.* Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics* **13**, 1008–1017 (1992) | Pease, A. C. *et al.* Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl Acad. Sci. USA* **91**, 5022–5026 (1994) | Hoheisel, J. D. Microarray technology: beyond transcript profiling and genotype analysis. *Nature Rev. Genet.* **7**; 200–210 (2006) | Allison, D. B., Cui, X., Page, G. P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nature Rev. Genet.* **7**, 55–65 (2006) | Fan, J. B., Chee, M. S., Gunderson, K. L. Highly parallel genomic assays. *Nature Rev. Genet.* **7**, 632–644 (2006)

MILESTONE 17

Silencing by stealth

Genetic manipulation can allow one to determine gene function in tractable organisms, but other systems seemed unusable for these studies, until the discovery of RNA interference (RNAi), a naturally occurring process in eukaryotes.

Studies in the mid-1980s found that introducing antisense RNA into a cell could shut down mRNA expression, but the process by which this happened was not understood. A breakthrough occurred in 1998, when Fire, Mello and colleagues examined the specific requirements for antisense RNA activity in a *Caenorhabditis elegans* system. They found that both mRNA and protein levels were reduced, and that double-stranded RNA (dsRNA) was much more effective than single-stranded RNA (ssRNA), suggesting that a model involving simple base-pairing with the mRNA was not sufficient. Unexpectedly, the effect also persisted into the progeny of the injected animals. The authors proposed that dsRNA could be used as a genetic tool to investigate the function of any coding region.

A year later, Hamilton and Baulcombe addressed the basis of a putative antiviral protection phenomenon, known as post-transcriptional gene silencing (PTGS), in plants and fungi. They set out to find the endogenous RNAs involved in this process and discovered that plant PTGS was tightly associated with the presence of ~25 nucleotide (nt) RNAs antisense to the transcripts investigated. This study confirmed that antisense RNA mediated an endogenous process.

The term RNAi became commonly used after publication of a paper by Tuschl and colleagues in 2001. By this time, it was known that longer dsRNAs were processed to 21–23 nt pieces. In this work, Tuschl and colleagues showed that 21–22 nt RNAs with short 3′ overhangs, which they called small-interfering RNAs (siRNAs), were effective in promoting mRNA degradation. They also found that the mRNA was cleaved 9–10 nt from the 5′ end of the siRNA, and that the end from which processing occurred dictated whether the active strand was sense or antisense.

From these studies, the mechanism of RNAi was sufficiently understood that researchers could begin to make specific siRNAs that would cleave mRNAs at a desired location. This technique has since revolutionized our views of the value of non-coding RNAs, the regulation of gene expression in development and the potential for specific gene silencing in a therapeutic setting.

Angela K. Eggleston,
Senior Editor, Nature

ORIGINAL RESEARCH PAPERS Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998) | Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**, 950–952 (1999) | Elbashir, S. M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188–200 (2001)

MILESTONE 18

The next generation arrives

Although Sanger sequencing served researchers admirably for almost three decades, in recent years there has been increasing pressure to produce ever-larger amounts of sequence as rapidly and cheaply as possible. These demands have catalysed the development of a new generation of sequencing technologies.

Sequencing can be made faster and cheaper if a single volume of reagents is used in parallel on thousands or millions of targets. In 1998, Ronaghi and colleagues showed that a method they had recently developed, known as pyrosequencing, could be carried out on a solid support and was therefore suitable for such multiplexing. In pyrosequencing, pyrophosphate — released upon nucleotide addition by DNA polymerase — is converted to ATP. This triggers a luciferase enzyme to produce light, which is used to detect an incorporation event, so that a sequence read can be built up over successive rounds using different deoxynucleotides. Importantly, because nucleotide addition is detected by the emission of photons, this method is well-suited to detection using simple optics and automated data collection.

Another important advance came in 2003, when Mitra and colleagues described an approach that allowed the multiplexing of both template amplification and sequencing. They adapted an existing method, known as polymerase-colony technology, which involves the amplification of millions of DNA molecules by PCR in an acrylamide gel. Because the products are prevented from diffusing away, a spherical colony of DNA — known as a polony — is formed for each target. This paper showed that sequencing can be carried out on polonies, allowing many reactions to take place in parallel.

In 2005, two papers illustrated the benefits of advances in sequencing technology, describing the rapid sequencing of whole bacterial genomes. Shendure and colleagues used polony-based amplification combined with another new method — sequencing by ligation — which, similar to pyrosequencing, involves successive rounds of detection. Here, a primer is hybridized to a known sequence next to the DNA target. DNA ligase then joins oligonucleotides, which are fluorescently labelled at one position, to the primer. Because the ligase prefers to join molecules when the bases in double-stranded DNA match, the fluorescent signal from the ligated oligonucleotide can be used as a readout of the target sequence, again allowing automated data collection. In the second 2005 paper, Margulies and colleagues described a sequencing system that uses fibre-optic slides with more than 1 million wells. They showed that robust pyrosequencing could be carried out in the resulting picolitre volumes, and sequenced an impressive 25 million bases in a single run.

With high-throughput sequencing now becoming available to increasing numbers of researchers, next-generation approaches are set to bring major advances in genetics and genomics, from the rapid sequencing of new genomes, to the large-scale characterization of genetic variation in populations, to personalized genomes.

Louisa Flintoft, Senior Editor, Nature Reviews Genetics

ORIGINAL RESEARCH PAPERS Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363, 365 (1998) | Mitra, R. D., Shendure, J., Olejnik, J., Edyta-Krzymanska-Olejnik, & Church, G. M. Fluorescent *in situ* sequencing on polymerase colonies. *Anal. Biochem.* **320**, 55–65 (2003) | Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005) | Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005)

FURTHER READING Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. Advanced sequencing technologies: methods and goals. *Nature Rev. Genet.* 5, 335–344 (2004)

MILESTONE 19

The clickable genome

> These heavily used genome browsers are now so much a part of the fabric of genome-based biological research that their contribution to progress would be difficult to overestimate.

Consider a thought experiment in which the sequences of the human genome and the genomes of several model organisms were finished to a high standard, but the repositories that were needed to house and make sense of these data in an accessible manner did not exist. The sequence databases could be filled, but the usefulness of the data to biologists would be greatly diminished. In place of a powerful resource, we would have an 'alphabet soup'.

The earliest repositories for DNA sequences established in the early-to-mid 1980s were the European Molecular Biology Laboratory Data Library, GenBank at the National Center for Biotechnology Information (NCBI) and the DNA Data Bank of Japan. Although these databases have served the community admirably, as the pace of sequencing increased, it became clear that new graphical user interfaces would have to be developed in order to facilitate the viewing and manipulation of both the sequences and the subsequent annotations that would make them meaningful. An early and influential effort in this regard was ACeDB, a genomics database that was originally developed for the *Caenorhabditis elegans* Genome Project that could display genetic, cosmid and sequence maps in a flexible manner.

When the assembled draft sequence of the human genome was published in February 2001, it was made available through three public portals: Ensembl, the University of California, Santa Cruz (UCSC) Genome Browser and the NCBI Map Viewer. Mainly funded by the Wellcome Trust, Ensembl is a joint project of the European Bioinformatics Institute and the Wellcome Trust Sanger Institute. It presents a range of views of the human genome and the genomes of an ever-increasing number of other organisms. Importantly, these are 'clickable' genomes, so the user can home in on small regions of a genome of interest to see protein-coding genes, RNA-coding genes, single-nucleotide polymorphisms, nucleotide composition, pseudogenes, contigs, expressed sequence tags, comparative alignments to other genomes and links to a suite of other databases that constitute the ongoing effort to produce a deep functional annotation of sequenced genomes. The UCSC Genome Browser, which was produced in its initial form by the Santa Cruz group that carried out the first genome assembly for the public Human Genome Project, can also present a view of the genome at any scale, and offers annotations in a series of 'tracks' that can be added or eliminated depending on the interests of the user. These heavily used genome browsers are now so much a part of the fabric of genome-based biological research that their contribution to progress would be difficult to overestimate.

Alan Packer, Senior Editor,
Nature Genetics

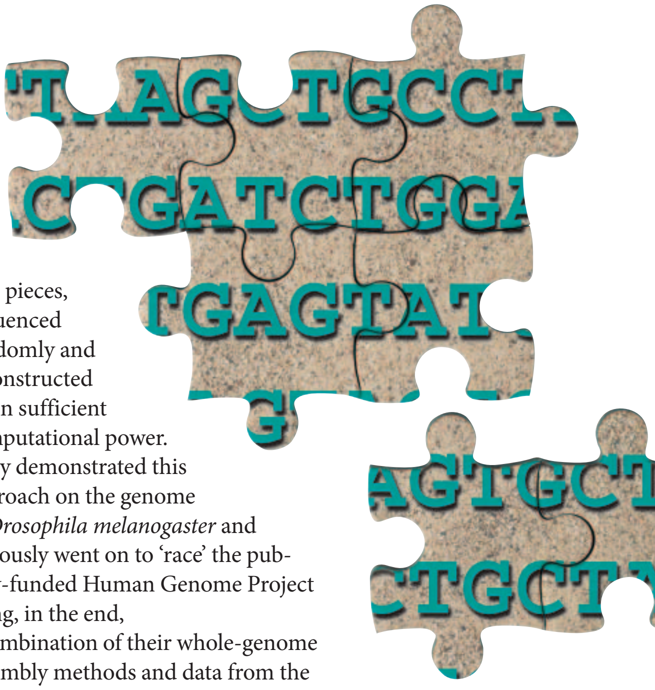
ORIGINAL RESEARCH PAPERS Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002) | Birney, E. *et al.* An overview of Ensembl. *Genome Res.* **14**, 925–928 (2004)

FURTHER READING Reed, J. L., Famili, I. & Thiele, I. & Palsson, B.O. Towards multidimensional genome annotation. *Nature Rev. Genet.* **7**,130–141 (2006) | Eichler, E. E., Clark, R. A. & She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nature Rev. Genet.* **5**, 345–354 (2004) | Chintapalli, V.R., Wang, J. & Dow, J. A. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nature Genet.* **39**, 715–720 (2007) | Boffelli, D., Nobrega, M. A. & Rubin, E. M. Comparative genomics at the vertebrate extremes. *Nature Rev. Genet.***5**, 456–465 (2004) | Ureta-Vidal, A., Ettwiller, L. & Birney, E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nature Rev. Genet.* **4**,251–262 (2003)

WEB SITES
DNA Data Bank of Japan: http://www.ddbj.nig.ac.jp
Ensembl: http://www.ensembl.org
NCBI Map Viewer: http://www.ncbi.nlm.nih.gov/mapview
UCSC Genome Browser: http://www.genome.ucsc.edu

MILESTONE 20

Putting it all together



You might remember this problem from your childhood: when you lose the top to your puzzle box, you are confronted with lots of pieces and no idea what they are supposed to look like when assembled. Genome sequencers faced the same dilemma when beginning large-scale DNA sequencing. They did the same thing that you might: they started at known landmarks and systematically built up the larger picture.

In order to assemble short stretches of DNA sequence from each read into a larger whole, particularly on a large scale, bioinformaticists developed algorithms that could take input directly from fluorescent sequencing machines. The earliest programs to achieve wide use were called Phred, Phrap and Consed, developed by Phil Green and colleagues. Phred initially went through the sequence reads and assigned a 'base call' to the chromatogram output from the machine. Phrap then assembled the list of bases from multiple reads into the most likely single path through the sequence. Users then viewed and edited the output with Consed, to generate higher-quality sequences as required. These programs were developed for, and used on, the public Human Genome Project.

Gene Myers and colleagues later developed an algorithm that used the end-pair information from sequencing subclones and could assemble larger sequences. They postulated that the whole genome could be cut into pieces, sequenced randomly and reconstructed given sufficient computational power. They demonstrated this approach on the genome of *Drosophila melanogaster* and famously went on to 'race' the publicly-funded Human Genome Project using, in the end, a combination of their whole-genome assembly methods and data from the public project. However, so-called shotgun whole-genome assemblies are now the method of choice for large genome projects, and the field has moved on to

MILESTONE 21

Fishing for genes

Gene-prediction tools have co-evolved with advances in genome-sequencing capabilities, to make sense of the ever-increasing amount of data.

Early programs sifted through sequences to identify open reading frames. Later, richer representations of the features that distinguish coding from noncoding sequences were used, and methods treated gene prediction as a pattern-recognition problem. These programs, such as Genie, Genscan, GLIMMER and FGENESH, used linear-discriminant analysis, Markov models, neural networks or a combination of methods to detect the variation. Relatively simple models can be used for microbial gene identification: GLIMMER was applied to 10 completed microbial genomes in 1999.

The presence of exon splice sites complicates eukaryotic gene prediction. Each eukaryotic gene is 'marked' by start and stop codons and has splice sites (for example, ATG … GT-AG … GT-AG … Stop). Markov model-based applications then complete the identification — these algorithms are designed to compare sliding windows (several bases in length) to patterns that the program has 'learned' in a given genome. To learn, the program is trained on a part of the sequence for which gene information has already been determined experimentally.

BANANASTOCK

Genie was one of the first programs developed to mine the human genome, and it identified up to 85% of the known protein-coding bases — a performance on par with that of other programs available in 1996. More recently, an abundance of expressed sequence data has improved prediction accuracy. FGENESH and a related suite of programs use an algorithm similar to that of Genie on a first pass; then, a BLAST comparison (see Milestone 15) to databases of exon products is used to mark confirmed exons.

The latest gene-annotation tools, such as Ensembl and Gnomon, integrate an even greater amount of information, including protein homology, cDNA and expressed sequence tag data. Ensembl was written to analyze the draft human genome, and since then has been used to annotate various vertebrate genomes, including zebrafish and human among others. Gnomon was developed to analyse other genomes, and was recently used for honey bee gene annotation.

Today, we take it as a given that the raw genomic data are presented in next-generation programs like Arachne, Atlas and PCAP, each using different algorithms.

Chris Gunter, Senior Editor, Nature

ORIGINAL RESEARCH PAPERS Ewing, B. & Green, P. Basecalling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998) | Ewing, B., Hillier, L., Wendl, M. & Green, P. Basecalling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998) | Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998) | Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 868–877 (2000)

FURTHER READING International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001) | Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001) | She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004)

WEB SITES

National Center for Biotechnology Information assembly information: http://www.ncbi.nlm.nih.gov/genome/guide/Assembly/Assembly.shtml

Phrap: http://www.phrap.org

a meaningful way, as annotated sequences. Yet more is to come, as with increasing amounts of data to base their models on, computational biologists can train programs to make more accurate predictions and to mine the genome at an ever-increasing level of detail.

Irene Kaganman, Senior Copy Editor, Nature Methods

ORIGINAL RESEARCH PAPERS Kulp, D. *et al.* A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **4**, 134–142 (1996) | Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997) | Delcher, A. L. *et al.* Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999) | Salamov, A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000) | Curwen, V. *et al.* The Ensembl automatic gene annotation system. *Genome Res.* **14**, 942–950 (2004)

WEB SITES

Ensembl: http://www.ensembl.org

FGENESH: http://www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=gfs

Genie: http://www.fruitfly.org/seq_tools/genie.html

Genscan: http://genes.mit.edu/GENSCAN.html

GLIMMER: http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi

Gnomon: http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.html

MILESTONE 22

Vive les differénces

Even before DNA sequencing was developed, we knew from enzyme polymorphisms that people had a few genetic differences that could be linked to physical traits. We now know the most easily identifiable variations are single nucleotide polymorphisms (SNPs).

In the early 1990s, coincident with the development of microarrays (see Milestone 16) and discussions of sequencing the human genome, several groups demonstrated that PCR-based or ligation-based sample-preparation methods combined with array-based technologies could readily identify SNPs. At the same time, it became clear that linkage-based studies to find disease genes were limited in power; they worked when rare variants had large effects on health, but complex diseases were due to a combination of common variants with each contributing a smaller effect. In 1996, Neil Risch and Kathleen Merikangas made the controversial proposal that statistical association-based techniques were the method of choice for complex diseases. In order to scan for disease genes, however, markers had to be discovered across the entire genome, requiring massive effort and resources from an international collaboration. The resulting race to develop technologies for SNP detection on a large scale revolutionized (and monetized) human genetics.

By November 2000, the SNP database dbSNP contained over 1.42 million SNPs across the human genome. In 2001, Mark Daly and colleagues showed that the SNPs in the human genome existed in a block-like structure, in which all SNPs in the kilobase-length blocks were linked together into only a few combinations or haplotypes. One simply needed to genotype a few of the SNPs in each block to learn the status of all the others, greatly reducing the complexity of the process. In 2005, a large consortium published the first haplotype map — the genotype of 1 million SNPs in 269 samples from essentially three population groups — setting off a frenzy of whole-genome association studies for common disease-susceptibility variants.

Chris Gunter, Senior Editor, Nature

> “The resulting race to develop technologies for SNP detection on a large scale revolutionized (and monetized) human genetics.”

ORIGINAL RESEARCH PAPERS Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**,1516–1517 (1996) | Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001) | International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**,1299–1320 (2005)

FURTHER READING Landegren, U. Ligation-based DNA diagnostics. *Bioessays* **15**, 761–765 (1993)

WEB SITE

dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP/index.html

MILESTONE 23

Charting the DNA-methylation landscape

In the late 1980s, it became apparent that cytosine methylation within CpG dinucleotides is sufficient to block the binding of transcription factors to DNA, thereby inhibiting transcription. To understand the role of this DNA modification *in vivo*, it became necessary to determine its frequency and location.

The first breakthrough came in 1992, with the development of bisulphite sequencing. This method exploits the fact that sodium bisulphite treatment induces a conversion of unmethylated (but not methylated) cytosine to uracil. DNA is then PCR amplified and sequenced. Because the resulting DNA strands are no longer complementary, PCR primers can be designed to yield strand-specific methylation patterns.

Bisulphite sequencing was especially useful for studies of individual loci, but with the dawn of the genomic era more global approaches were required. Developed in 2005, methylated DNA immunoprecipitation (MeDIP) provided a crucial advance in this regard. Here, an antibody specific for methylated cytosines is used to capture methylated DNA fragments, which can then be analysed in a range of standard ways, including by hybridization to DNA microarrays. Unlike previous restriction-based approaches, methylation detection using MeDIP is unbiased by the restriction enzyme-recognition sequence. In their 2005 paper, Weber *et al.* used arrays of human bacterial artificial chromosome (BAC) clones to generate chromosomal maps of methylation of the human genome, with an average tiling resolution of 80 kb. They also used this approach to carry out a global comparison of CpG island methylation in normal and colon cancer cells, revealing specific sites of hypermethylation in the latter.

Immunoprecipitation-based approaches were subsequently used to characterize genome-wide methylation in *Arabidopsis*. Zhang *et al.* and Zilberman *et al.* used antibodies against methylated cytosines and then hybridized the resulting samples to high-density tiling arrays. The resolution of the arrays (only 35 base pairs between the oligos) used by Zhang *et al.* allowed them to generate a particularly high-resolution genome-wide methylation map for *Arabidopsis*. Together, these two studies generated a wealth of data on the distribution of cytosine methylation in relation to functional elements within the genome, such as open reading frames and promoters. They also revealed a crucial interdependence between methylation and transcription.

Since 2000, the Human Epigenome Project has been identifying, cataloguing and interpreting genome-wide DNA-methylation patterns of human genes in major tissues and cell types. The recently initiated international Alliance for the Human Epigenome and Disease (AHEAD) project will extend this work to other epigenetic marks, and their roles in development and disease, including cancer. As always, biological discoveries will no doubt go hand in hand with further technological breakthroughs.

Magdalena Skipper, Chief Editor,
Nature Reviews Genetics

ORIGINAL RESEARCH PAPERS Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl Acad. Sci. USA* **89**, 1827–1831 (1992) | Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genet.* **37**, 852–862 (2005) | Zhang, X. *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* **126**, 1189–1201 (2006) | Zilberman, D. *et al.* Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genet.* **39**, 61–69 (2007)

CITING THE MILESTONES

The *Nature Milestones* in DNA technologies supplement is published simultaneously by *Nature*, *Nature Methods* and *Nature Review Genetics*.
However, most referencing formats and software do not allow the inclusion of more that one journal name or volume in an article reference. Therefore, should you wish to cite any of the Milestones or Commentaries, please reference the page number (Sxx–Sxx) as a supplement to *Nature Reviews Genetics*. For example, *Nature Rev. Genet.* **8**, Sxx–Sxx (2007). To cite articles from the collection, please use the original citation, which can be found at the start of each article.