1

A brief history of genomics

From time to time new scientific breakthroughs and technologies arise that forever change scientific practice. During the last 50 years, several advances stand out in our minds that – coupled with advances in the computational and computer sciences – have made genomic studies possible. In the brief history of genomics presented here we review the circumstances and consequences of these relatively recent technological revolutions.

Our brief history begins during the years immediately following World War II. It can be argued that the enzyme period that preceded the modern era of molecular biology was ushered in at this time by a small group of physicists and chemists, R. B. Roberts, P. H. Abelson, D. B. Cowie, E. T. Bolton, and J. R. Britton in the Department of Terrestrial Magnetism of the Carnegie Institution of Washington. These scientists pioneered the use of radioisotopes for the elucidation of metabolic pathways. This work resulted in a monograph titled *Studies of Biosynthesis in* Escherichia coli that guided research in biochemistry for the next 20 years and, together with early genetic and physiological studies, helped establish the bacterium *E. coli* as a model organism for biological research [1]. During this time, most of the metabolic pathways required for the biosynthesis of intermediary metabolites were deciphered and biochemical and genetic methods were developed to identify and characterize the enzymes involved in these pathways.

Much in the way that genomic DNA sequences are paving the way for the elucidation of global mechanisms for genetic regulation today, the biochemical studies initiated in the 1950s that were based on our technical abilities to create isotopes and radiolabel biological molecules paved the way for the discovery of the basic mechanisms involved in the regulation of metabolic pathways. Indeed, these studies defined the biosynthetic pathways for the building blocks of macromolecules such as proteins and

nucleic acids and led to the discovery of mechanisms important for metabolic regulation such as end product inhibition, allostery, and modulation of enzyme activity by protein modifications. However, major advances concerning the biosynthesis of macromolecules awaited another breakthrough, the description of the structure of the DNA helix by James D. Watson and Francis H. C. Crick in 1953 [2]. With this information, the basic mechanisms of DNA replication, protein synthesis, gene expression, and the exchange and recombination of genetic material were rapidly unraveled.

During the enzyme period, geneticists around the world were using the information provided by biochemists to develop model systems such as bacteria, fruit flies, yeast, and mice for genetic studies. In addition to establishment of the basic mechanisms for protein-mediated regulation of gene expression by F. Jacob and J. Monod in 1961 [3], these genetic studies led to fundamental discoveries that were to spawn yet another major change in the history of molecular biology. This advance was based on studies designed to determine why *E. coli* cells once infected by a bacteriophage were immune to subsequent infection. These seemingly esoteric investigations led by Daniel Nathans and Hamilton Smith [4] resulted in the discovery of new types of enzymes, restriction endonucleases and DNA ligases, capable of cutting and rejoining DNA at sequence-specific sites. It was quickly recognized that these enzymes could be used to construct recombinant DNA molecules composed of DNA sequences from different organisms. As early as 1972 Paul Berg and his colleagues at Stanford University developed an animal virus, SV40, vector containing bacteriophage lambda genes for the insertion of foreign DNA into *E. coli* cells [5]. Methods of cloning and expressing foreign genes in *E. coli* have continued to progress until today they are fundamental techniques upon which genomic studies and the entire biotechnology industry are based.

The recent history of genomics also has been driven by technological advances. Foremost among these advances were the methodologies of the polymerase chain reaction (PCR) and automated DNA sequencing. PCR methods allowed the amplification of usable amounts of DNA from very small amounts of starting material. Automated DNA sequencing methods have progressed to the point that today the entire DNA sequence of microbial genomes containing several million base pairs can be obtained in less than one week. These accomplishments set the stage for the human genome project.

As early as 1984 the small genomes of several microbes and bacteriophages had been mapped and partially sequenced; however, the modern era

of genomics was not formally initiated until 1986 at an international conference in Santa Fe, New Mexico sponsored by the Office of Health and Environmental Research[1] of the US Department of Energy. At this meeting, the desirability and feasibility of implementing a human genome program was unanimously endorsed by leading scientists from around the world. This meeting led to a 1988 study by the National Research Council titled *Mapping and Sequencing the Human Genome* that recommended the United States support a human genome program and presented an outline for a multiphase plan. In that same year, three genome research centers were established at the Lawrence Berkeley, Lawrence Livermore, and Los Alamos national laboratories. At the same time, under the leadership of Director James Wyngaarden, the National Institutes of Health established the Office of Genome Research which in 1989 became the National Center for Human Genome Research, directed by James D. Watson. The next ten years witnessed rapid progress and technology developments in automated sequencing methods. These technologies led to the establishment of large-scale DNA sequencing projects at many public research institutions around the world such as the Whitehead Institute in Boston, MA and the Sanger Centre in Cambridge, UK. These activities were accompanied by the rapid development of computational and informational methods to meet challenges created by an increasing flow of data from large-scale genome sequencing projects.

In 1991 Craig Venter at the National Institutes of Health developed a way of finding human genes that did not require sequencing of the entire human genome. He relied on the estimate that only about 3 percent of the genome is composed of genes that express messenger RNA. Venter suggested that the most efficient way to find genes would be to use the processing machinery of the cell. At any given time, only part of a cell's DNA is transcriptionally active. These "expressed" segments of DNA are converted and edited by enzymes into mRNA molecules. Using an enzyme, reverse transcriptase, cellular mRNA fragments can be transcribed into complementary DNA (cDNA). These stable cDNA fragments are called expressed sequence tags, or ESTs. Computer programs that match overlapping ends of ESTs were used to assemble these cDNA sequences into longer sequences representing large parts, or all, of many human genes. In 1992, Venter left NIH to establish The Institute for Genomic Research, TIGR. By 1995 researchers in public and private institutions had isolated over 170000

[1] Changed in 1998 to the Office of Biological and Environmental Research of the Department of Energy.

ESTs, which were used to identify more than half of the then estimated 60000 to 80000 genes in the human genome.[2] In 1998, Venter joined with Perkin-Elmer Instruments (Boston, MA) to form Celera Genomics (Rockville, MD).

With the end in sight, in 1998 the Human Genome Program announced a plan to complete the human genome sequence by 2003, the 50th anniversary of Watson and Crick's description of the structure of DNA. The goals of this plan were to:

- Achieve coverage of at least 90% of the genome in a working draft based on mapped clones by the end of 2001.
- Finish one-third of the human DNA sequence by the end of 2001.
- Finish the complete human genome sequence by the end of 2003.
- Make the sequence totally and freely accessible.

On June 26, 2000, President Clinton met with Francis Collins, the Director of the Human Genome Program, and Craig Venter of Celera Genomics to announce that they had both completed "working drafts" of the human genome, nearly two years ahead of schedule. These drafts were published in special issues of the journals *Science* and *Nature* early in 2001 [6, 7] and the sequence is online at the National Center for Biotechnology Information (NCBI) of the Library of Medicine at the National Institutes of Health

As of this writing, the NCBI databases also contain complete or in progress genomic sequences for ten *Archaea* and 151 bacteria as well as the genomic sequences of eight eukaryotes including: the parasites *Leishmania major* and *Plasmodium falciparum*; the worm *Caenorhabditis elegans*; the yeast *Saccharomyces cerevisiae*; the fruit fly *Drosophila melanogaster*; the mouse *Mus musculus*; and the plant *Arabidopsis thaliana.* Many more genome sequencing projects are under way in private and public research laboratories that are not yet available on public databases. It is anticipated that the acquisition of new genome sequence data will continue to accelerate. This exponential increase in DNA sequence data has fuelled a drive to develop technologies and computational methods to use this information to study biological problems at levels of complexity never before possible.

[2] At the present time (September 2001) the estimate of the number of human genes has decreased nearly twofold.

REFERENCES

1. Roberts, R. B., Abelson, P. H., Cowie, D. B., Bolton, E. B., and Britten, J. R. *Studies of Biosynthesis in* Escherichia coli. 1955. Carnegie Institution of Washington, Washington, DC.
2. Watson, J. D., and Crick, F. H. C. A structure for deoxyribose nucleic acid. 1953. *Nature* 171:173.
3. Jacob, F., and Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. 1961. *Journal of Molecular Biology* 3:318–356.
4. Nathans, D., and Smith, H. O. A suggested nomenclature for bacterial host modification and restriction systems and their enzymes. 1973. *Journal of Molecular Biology* 81:419–423.
5. Jackson, D. A., Symons, R. H., and Berg, P. Biochemical method for inserting new genetic information into DNA of simian virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. 1972. *Procedings of the National Academy of Sciences of the USA* 69:2904–2909.
6. *Science* Human Genome Issue. 2001. 16 February, vol. 291.
7. *Nature* Human Genome Issue. 2001. 15 February, vol. 409.

2

DNA array formats

Array technologies monitor the combinatorial interaction of a set of molecules, such as DNA fragments and proteins, with a predetermined library of molecular probes. The currently most advanced of these technologies is the use of DNA arrays, also called DNA chips, for simultaneously measuring the level of the mRNA gene products of a living cell. This method, gene expression profiling, is the major topic of this book.

In its most simple sense, a DNA array is defined as an orderly arrangement of tens to hundreds of thousands of unique DNA molecules (probes) of known sequence. There are two basic sources for the DNA probes on an array. Either each unique probe is individually synthesized on a rigid surface (usually glass), or pre-synthesized probes (oligonucleotides or PCR products) are attached to the array platform (usually glass or nylon membranes). The various types of DNA arrays currently available for gene expression profiling, as well as some developing technologies, are summarized here.

In situ synthesized oligonucleotide arrays

The first *in situ* probe synthesis method for manufacturing DNA arrays was the photolithographic method developed by Fodor *et al.* [1] and commercialized by Affymetrix Inc. (Santa Clara, CA). First, a set of oligonucleotide DNA probes (each 25 or so nucleotides in length) is defined based on its ability to hybridize to complementary sequences in target genomic loci or genes of interest. With this information, computer algorithms are used to design photolithographic masks for use in manufacturing the probe arrays. Selected addresses on a photo-protected glass surface are illuminated through holes in the photolithographic mask, the glass surface is flooded with the first nucleotide of the probes to be synthesized at the

selected addresses, and photo-chemical coupling occurs at these sites. For example, the addresses on the glass surface for all probes beginning with guanosine are photo-activated and chemically coupled to guanine bases. This step is repeated three more times with masks for all addresses with probes beginning with adenosine, thymine, or cytosine. The cycle is repeated with masks designed for adding the appropriate second nucleotide of each probe. During the second cycle, modified phosphoramidite moieties on each of the nucleosides attached to the glass surface in the first step are light-activated through appropriate masks for the addition of the second base to each growing oligonucleotide probe. This process is continued until unique probe oligonucleotides of a defined length and sequence have been synthesized at each of thousands of addresses on the glass surface (Figure 2.1).

Several companies such as Protogene (Menlo Park, CA) and Agilent Technologies (Palo Alto, CA) in collaboration with Rosetta Inpharmatics (Kirkland, WA) of Merck & Co. Inc. (Whitehouse Station, NJ) have developed *in situ* DNA array platforms through proprietary modifications of a standard piezoelectric (ink-jet) printing process that unlike the manufacturing process for Affymetrix GeneChips™, does not require photolithography. These *in situ* synthesized oligonucleotide arrays are fabricated directly on a glass support on which oligonucleotides up to 60 nucleotides are synthesized using standard phosphoramidite chemistry. The ink-jet printing technology is capable of depositing very small volumes – picoliters per spot – of DNA solutions very rapidly and very accurately. It also delivers spot shape uniformity that is superior to other deposition methods.

Researchers in the Nano-fabrication Center at the University of Wisconsin have developed yet another method for the manufacture of *in situ* synthesized DNA arrays that also does not require photolithographic masks [2]. This technology known as MAS for maskless array synthesizer capitalizes on existing electronic chips used in overhead projection known as digital light processors (DLPs). A DLP is an array of up to 500000 tiny aluminum mirrors arranged on a computer chip. By electronic manipulation of the mirrors, light can be directed to specific addresses on the surface of a DNA array substrate, thus eliminating the need for expensive photolithographic masks. This technology is being implemented by NimbleGen Systems, LLC (Madison, WI). DNA arrays containing over 307000 discrete features are currently being synthesized and plans are under way to synthesize a second-generation MAS array containing over 2 million discrete features. The Wisconsin researchers claim that this method will greatly reduce the time and cost for the manufacture of high-density *in situ*

Figure 2.1. The Affymetrix method for the manufacture of *in situ* synthesized DNA microarrays (courtesy of Affymetrix). (1) A photo-protected glass substrate is selectively illuminated by light passing through a photolithographic mask. (2) Deprotected areas are activated. (3) The surface is flooded with a nucleoside solution and chemical coupling occurs at photo-activated positions. (4) A new photolithographic mask pattern is applied. (5) The coupling step is repeated. (6) This process is repeated until the desired set of probes is obtained.

This figure is available in colour for download from www.cambridge.org/9780521176354

Table 2.1. *Commercial sources for DNA arrays*

Company	Nylon filters	Glass slides	Plastic slides	Chips	Web site
Affymetrix[1,2,3,4,5,6,7,8,18]				X	www.affymetrix.com
Agilent Technologies[18]		X			www.chem.agilent.com
AlphaGene[1,18]		X			www.alphagene.com
Clontech[1,2,3,18]	X	X	X		www.clontech.com
Corning[6]		X			www.corning.com/cmt
Eurogentec[5,6,9,11,12,14,15,16,18]	X	X			www.eurogentec.be
Genomic Solutions[1,2,3]		X			www.genomicsolutions.com
Genotech[1,2]	X				www.genotech.com
Incyte Pharmaceuticals[1,2,3,4,9,10,18]	X	X			www.incyte.com
Invitrogen[1,2,3,6]	X			X	www.invitrogen.com
Iris BioTechnologies[1]					www.irisbiotech.com
Mergen Ltd[1,2,3]		X			www.mergen-ltd.com
Motorola Life Science[1,3,18]		X			www.motorola.com/lifesciences
MWG Biotech[3,6,8,18]		X			www.mwg-biotech.com
Nanogen				X	www.nanogen.com
NEN Life Science Products[1]		X		X	www.nenlifesci.com
Operon Technologies Inc.[1,6,18]		X			www.operon.com
Protogene Laboratories[18]					www.protogene.com
Radius Biosciences[18]		X			www.ultranet.com/~radius
Research Genetics[1,2,3,6]	X				www.resgen.com
Rosetta Inpharmatics[18]	X	X			www.rii.com
Sigma-Genosys[1,2,8,11,12,13,18]	X				www.genosys.com
Super Array Inc.[1,2,18]	X				www.superarray.com
Takara[1,2,4, 8,17,18]		X			www.takara.co.jp/english/bio_e

Notes:
[1]Human, [2]Mouse, [3]Rat, [4]*Arabidopsis*, [5]*Drosophila*, [6]*Saccharomyces cerevisiae*, [7]HIV, [8]*Escherichia coli*, [9]*Candida albicans*, [10]*Staphylococcus aureus*, [11]*Bacillus subtilis*, [12]*Helicobacter pylori*, [13]*Campylobacter jejuni*, [14]*Streptomyces lividans*, [15]*Streptococcus pneumoniae*, [16]*Neisseria meningitidis*, [17]Cyanobacteria, [18]Custom.

synthesized DNA mircoarrays, and bring this activity into individual research laboratories.

CombiMatrix (Snoqualmie, WA) and Nanogen (San Diego, CA) are developing electrical addressing systems for the manufacture of DNA arrays on semiconductor chips. The CombiMatrix method involves attaching each addressable site on the chip to an electrical conduit (electrode) applied over a layer of porous material. Each DNA probe is synthesized one base at a time by flooding the porous layer with a nucleoside and activating each electrode where a new base is to be added. Once activated, the electrode causes an electrochemical reaction to occur which produces

chemicals that react with the existing nucleotides, or chains of DNA, at that site for bonding to the probe site or to the next nucleotide base. At present, CombiMatrix has produced DNA arrays with 100 μm features that possess 1024 test sites within less than a square centimeter. Researchers at CombiMatrix believe that by using a standard 0.25-μm semiconductor fabrication process, they can produce a biological array processor with over 1 000 000 sites per square centimeter.

Nanogen uses a similar process to attach pre-synthesized oligonucleotides to electronically addressable sites on a semiconductor chip. To date, Nanogen has only produced a 99 probe array suitable for forensic and diagnostic purposes; however, Nanogen's researchers anticipate electronic arrays with thousands of addresses for genomics applications.

Pre-synthesized DNA arrays

The method of attaching pre-synthesized DNA probes (usually 100–5000 bases long) to a solid surface such as glass (or nylon filter) supports was conceived 25 years ago by Ed Southern and more recently popularized by the Patrick O. Brown laboratory at Stanford University. While the early manufacturing methods for miniaturized DNA arrays using *in situ* probe synthesis required sophisticated and expensive robotic equipment, the glass slide DNA array manufacturing methods of Brown made DNA arrays affordable for academic research laboratories. As early as 1996 the Brown laboratory published step-by-step plans for the construction of a robotic DNA arrayer on the internet. Since that time, many commercial DNA arrayers have become available. Besides the commercially produced Affymetric GeneChips™, these Brown-type glass slide DNA arrays are currently the most popular format for gene expression profiling experiments.

The Brown method for printing glass slide DNA arrays involves the robotic spotting of small volumes (in the nanoliter to picoliter range) of a DNA probe sample onto a $25 \times 76 \times 1$ mm glass slide surface previously coated with poly-lysine or poly-amine for electrostatic adsorption of the DNA probes onto the slide. Depending upon the pin type and the exact printing technology employed, 200 to 10 000 spots ranging in size from 500 to 75 μm can be spotted in a 1-cm^2 area. Many public and private research institutions in the USA and abroad have developed core facilities for the in-house manufacture of custom glass slide DNA arrays. Detailed discussions of the instrumentation and methods for printing glass slide DNA arrays can be found in a book edited by Mark Schena titled *Microarray Biochip Technology* [3].