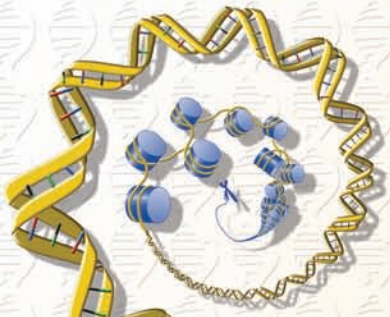
The Evolution of Sequencing Technology



As the basic forms of life began to evolve on earth, so too did a means to carry the essential information required for the replication and reproduction of these increasingly complex organisms. DNA, besides acting as this blueprint, can also provide us with information about our phylogeny, ancestry, environment, susceptibility to disease, and–some claim–our character and possibly our fate. It is no wonder that we as humans find the process and challenge of decoding our genetic information such an irresistible endeavor.

The first serious attempt to decode the human genome was the Human Genome Sequencing Project, started in 1990. Although it had its initial detractors, this effort turned out to be one of the most successful collaborative ventures in scientific research to date. While cooperation and hard work undoubtedly played a critical role, the core engine that drove this project was a string of engineering breakthroughs that allowed for the collection and collation of data at an unprecedented rate. Today, the capacity of those approaches is being far surpassed by technologies that cut sequencing times and costs by several orders of magnitude. Superior detection methods, massive multiplexing, and much-reduced sample size can yield complete microbial genomes in a day and human genomes in only weeks.

It is the scientific foundation that has enabled this extraordinarily rapid progress that we are attempting to capture in this poster. As our knowledge of DNA–both chemical and functional–has grown, so have sequencing technologies evolved and improved, each discovery building upon the previous as we proceed along the uncoiling DNA strand. This poster provides a snapshot of the current state of the art, as well as giving the reader a broad–and necessarily abbreviated–overview of the development of sequencing technologies over the last century and a half. If the speed of advancement in this area continues apace, even some of the more formidable challenges, such as single strand sequencing and the $1,000 genome, might be overcome–not just in our lifetimes, but within the foreseeable future.

Writer: John Hodgson
Design: Lewis Long
Editor: Sean Sanders
Commercial Editor, *Science*

The Evolution of Sequencing Technology

1865
Mendel shows inheritance patterns in plants

1868
Friedrich Miescher identifies "nuclein"

1879
Walther Flemming identifies distinct chromosomes

1903
Cytosine, the last of the DNA bases, is chemically characterized (A, T, and G done in the 1880s)

1928
Genetic transformation observed in pneumococci

1929
Phoebus Levene characterizes deoxyribose and phosphate-sugar-base nature of nucleotides

1934
DNA shown to be a polymer (previously considered a 10-mer)

1944
Avery, MacLeod, McCarthy propose DNA is carrier of genetic information; confirmed in 1952 by Hershey and Chase

1951
5-methylcytosine recognized in nucleic acids

UV spectrometry used to analyze composition of DNA by Chargaff lab

1953
High quality X-ray crystallography data from Rosalind Franklin and Maurice Wilkins allow Watson and Crick to elucidate double helix structure of DNA

1965
First nucleic acid sequence, of Yeast Ala-tRNA, reported

1970
First restriction enzyme, Endonuclease R, isolated from *Haemophilus influenzae*

1972
First recombinant DNA molecule: SV40 combined with λ

1973
Boyer and Cohen first to grow *Escherichia coli* containing recombinant plasmid

1977
Era of DNA sequencing begins: Maxam/Gilbert and Sanger describe sequencing techniques

1977
First genome sequenced: Phi-X174 phage (5,386 bases)

1980
"Shotgun sequencing" coined; technique used by Sanger and colleagues for sequencing and assembly of overlapping reads

1981
Hitachi and Akiyoshi Wada develop high throughput robotic detection; later incorporated into ABI sequencing products

1982
GenBank founded

1983
Kary Mullis and colleagues develop PCR

1986
First commercial DNA sequencer (ABI Prism 370A) launched by Applied Biosystems, Inc.; output = 1,000 bases per day

1990
BLAST algorithm developed at the NIH's NCBI

1991
EST strategy for expressed genes developed

1993
Human genome YAC map derived using highly automated sample handling

1995
ThermoSequenase for cycle sequencing released by Amersham

Applied Biosystems releases first capillary electrophoresis sequencer (Prism 310); output = 5,000-15,000 bases per day

Sequence of first free-living organism: *Haemophilus influenzae*

1996
First eukaryotic genome completed: *Saccharomyces cerevisiae*

1997
Molecular Dynamics MegaBACE 1000 capillary electrophoresis sequencer released; output = 250,000-500,000 bases per day

1998
Pyrosequencing developed; eliminates need for electrophoresis

PE Biosystems releases Prism 3700 multiple capillary sequencer; output = 500,000 to 1 million bases per day

First multicellular eukaryotic genome sequenced: *Caenorhabditis elegans*

1999
First human chromosome (22) sequence published

2000
First plant genome sequenced: *Arabidopsis thaliana*

2001
Drafts of human genome sequence published – 31,000 predicted genes, 95% of genome is noncoding

2004
Euchromatin sequence of human genome completed; predicted number of genes dropped to 24,000

The discovery of enzymes that reverse histone methylation

2005
Launch of Genome Sequencer 20 System by 454 Life Sciences based on pyrosequencing technology; output = 20 million bases per run

2006
62,250 bases of Neanderthal genome sequenced; demonstrates ability to obtain useful sequence from ancient DNA for comparison to modern humans

2007
First complete sequence of single named human; sevenfold coverage completed within four months

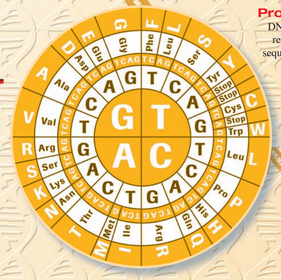
Sponsored by:

454 LIFE SCIENCES

Sponsored by:

Roche

Encoding of Sequence Information



Protein-coding Sequence
DNA that is transcribed and spliced to remove introns, leaving only exonic sequences, and is then translated to form functional proteins.

Pseudogenes
Sequences that have lost protein-coding ability or are otherwise no longer expressed in the cell because of mutation.

Promoters
DNA needed for transcription; promoters include an upstream binding site for RNA polymerase, a transcription start site, and regulatory elements.

Enhancers
200-1000 bp elements upstream or downstream of promoters that influence the transcription rates by binding activator/repressor proteins. Can be located many kilobases from the promoter they influence.

Sites of Methylation/CpG Islands
300-3000 bp sequences in which the frequency of the dinucleotide CG is over 5% rather than the 1% found in the rest of the genome. Methylation of cytosine moiety in CpG islands near promoters is associated with gene repression, and demethylation with activation.

Origins of Replication
Sites (many per chromosome) at which DNA replication is initiated; often associated with genome regions of high (>70%) A+T content.

Nuclease Hypersensitive Sites
Regions exposed to nuclease activity because they are accessible to other DNA-binding proteins such as transcription factors or RNA polymerase.

Histone Modifications

Modification	Enzyme	Illustrative Effects of Modification	Histones Involved
Acetylation	Histone acetyl transferases	Maintenance of active chromatin; acceleration of histone modification; transcriptional activation and silencing; dosage compensation; cell cycle progression; chromosome translocation	H1B, H2A, H2B, H3, H4
Deacetylation	Histone deacetylases	Transcription repression; trigger cell differentiation, growth arrest, apoptosis	
Methylation	Arginine methyltransferase	Transcription activation or repression; cell growth and proliferation; activation/repression of methylation, acetylation, and ubiquitination of other histones	H3, H4
	Lysine methylase	Transcription activation and repression; heterochromatin formation; chromosome loss; marks euchromatin for cytosine methylation (DNA imprinting)	H1B, H2B, H3, H4
Demethylation	Deiminase; lysine demethylase	Transcription activation or repression	H1B, H2B, H3, H4
Phosphorylation	Histone kinases	Initiates chromosome condensation; activates transcription by promoting acetylation; response to DNA damage	H1B, H2A, H2B, H2AX, H3, H4
Ubiquitination	Ubiquitin ligases	Transcription activation; stimulation of transcription-activating methylation of H3	H2A, H2B

Minimal Eukaryotic Promoters

Class or Promoter	Transcription Factor Binding Site	Transcription Start Point
TATA-box	TATAAAA	30 bp downstream of TATA-box
TATA-less	PyPyAN[TA]PyPy	Controlled by initiator
TATA-less with downstream promoter element (DPE)		Around 30 bp upstream of DPE

Scaffold/Matrix Attachment Sites
Generally AT-rich regions with long poly(A) sequences (over 100 bp), that create a rigid narrow minor groove structure bound by enzymes such as DNA topoisomerase II.

Histone Modifications
Acetylation, methylation, phosphorylation, and/or ubiquitination of DNA-packing histones represents a potential epigenetic code.

Genome Sizes

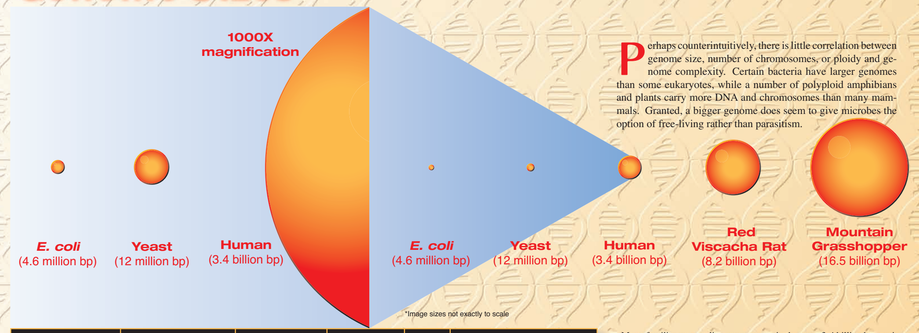


*Image sizes not exactly to scale

Perhaps counterintuitively, there is little correlation between genome size, number of chromosomes, or ploidy and genome complexity. Certain bacteria have larger genomes than some eukaryotes, while a number of polyploid amphibians and plants carry more DNA and chromosomes than many mammals. Granted, a bigger genome does seem to give microbes the option of free-living rather than parasitism.

Species	Common Name	Genome Size (billion bp)	Chromosome Number	Ploidy	Notes
Encephalitozoon intestinalis	Parasitic microsporidium	0.00225	10	1-2	Smallest eukaryal genome
Escherichia coli	-	0.0046	-	1	Typical bacterial genome size
Pneumocystis carinii (human)	-	0.0075	16	1	Smallest fungal genome
Saccharomyces cerevisiae	Baker's yeast	0.012	16	1-2	Common laboratory organism
Trichoplax adhaerens	-	0.054	6	1-2	Smallest animal genome
Caenorhabditis elegans	Worm	0.097	12	2	Common laboratory model
Fragaria viridis	Green strawberry	0.098	14	2	Smallest plant genome
Caenocholax fenyesi texensis	Twisted-wing parasite	0.108	24	2	Smallest insect genome
Arabidopsis thaliana	Mustard weed	0.125	5	1	Common laboratory model
Drosophila melanogaster	Fruit fly	0.18	8	2	Common laboratory model
Anopheles gambiae	Malaria mosquito	0.278	6	2	Familiar insect
Xenopus tropicalis	Western clawed frog	1.7	20	2	Smallest amphibian chromosome number
Danio rerio	Zebrafish	1.7	50	2	Common laboratory model
Canis familiaris	Domestic dog	2.5	78	2	Familiar animal
Zea mays	Maize	2.5	20	4	Familiar plant
Mus musculus	Mouse	2.7	40	2	Familiar animal
Xenopus laevis	South African clawed frog	3.0	36	4	Common laboratory model
Ornithorhynchus anatinus	Duck-billed platypus	3.0	54	2	Familiar animal
Homo sapiens	Human	3.4	46	2	Familiar animal
Bos taurus	Domestic cow	3.7	50	2	Familiar animal
Pan troglodytes	Chimpanzee	3.8	48	2	Familiar animal
Xenopus ruwenzoriensis	Uganda clawed frog	7.8*	108	12	Largest amphibian chromosome number
Tympanoctomys barrerae	Red viscacha rat	8.21	102	4	Largest mammalian genome
Podisma pedestris	Mountain grasshopper	16.5	22-24	2	Largest insect genome
Ambystoma mexicanum	Axotoil or Mexican salamander	21.9-48	28	2	Demonstrates regrowth of body parts
Ophioglossum petiolatum	Stalked adder's tongue	64	1020*	32-34*	Highest chromosome number and ploidy
Necturus lewisi	Neuse River waterdog	118	38	2	Largest amphibian genome
Fritillaria assyriaca	Assyrian fritillary	125	48	4	Largest plant genome
Protopterus aethiopicus	Marbled lungfish	130	Unknown	2	Largest animal genome

*estimated

Many familiar mammalian genomes are in the same 3-4 billion base pairs range as human, but a number of other mammals—bats and muntjak-like deer, for instance—have genomes under 2 billion bp. At the other extreme, the red viscacha rat (*Tympanoctomys barrerae*)—the only known mammalian tetraploid—has over twice the DNA of *Homo sapiens*, and more than double the chromosome number. The extremes of vertebrate genome sizes belong to fish: *Tetraodon nigroviridis*, the spotted green pufferfish, has a tiny 0.35 billion bp genome while the genome of *Protopterus aethiopicus*, the marbled lungfish, is approximately 130 billion bp.

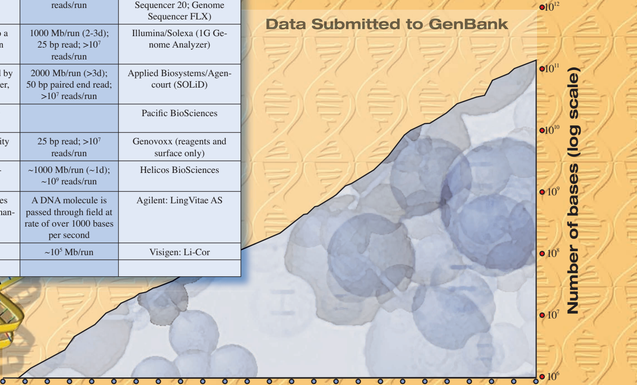
A small proportion of mammals, birds, amphibians, and plants have more than 100 chromosomes. No insects (so far) have such high numbers. The top of the tree from the perspective of genome organization is the stalked adder's tongue, a 32- to 34-ploid plant with an estimated 1,020 chromosomes.

For more information, go online to www.genomesonline.org/gold.cgi and egg.ebi.ac.uk/services/cogent/

Sequencing Technologies

Sequencing principles – now and the future

Class	Subclass	Description	Optimum Sequencing Performance	Companies (and Commercially Available Systems)
Synthetic chain-terminator chemistry (Sanger method)		1977; DNA polymerase synthesizes a set of DNA fragments each one base longer than the other. Size-specific separation of the fragments (by electrophoresis, for instance) and base-specific tagging of each fragment (by fluorescence, for instance) allows sequence to be deduced.	67,000 - 96,000 bp/run (1-3h), 700-1000 bp read	Applied Biosystems (3730; 3730xl)
Sequencing-by-hybridization		1987; target sequence is annealed to a fixed array of oligonucleotide probes (8-10 bases). Sequence is deduced from hybridization pattern.	25 bp read (probe size)	Affymetrix/Perlegen (GeneChip CustomSeq Resequencing Arrays); Illumina (Beadchip); Premier Biosoft (AlleleID)
Cyclic sequencing on amplified DNA		Enzymatic methods are multiplexed in systems with a large number of addressable locations.		
Available next generation technologies	Pyrosequencing	1996; parallel sequencing-by-synthesis. Amplified tethered target sequences fixed in distinct physical locations (e.g., wells). Cycle adds each deoxynucleotide in turn with an enzyme cocktail and wells "light up" when the correct deoxynucleotide is present.	100 Mb/run (7-8h), ~250 bp read; >400,000 reads/run	Roche Applied Science/454 Life Sciences (Genome Sequencer 20; Genome Sequencer FLX)
	Clonal single molecule array	2001; parallel sequencing-by-synthesis. Random fragments of target DNA tethered to a flow cell surface are amplified in situ. Sequencing cycle adds labeled, reversible chain terminators and interrogates all amplified clusters with a laser.	1000 Mb/run (2-3d); 25 bp read; >10^7 reads/run	Illumina/Solexa (1G Genome Analyzer)
	Sequencing-by-ligation	2007; unlabeled, tethered target DNA region defined by an "anchor primer" is probed by fluorescently labeled oligomers, the "query primers." Ligase extends the anchor primer, allowing subsequent bases to be determined in later cycles.	2000 Mb/run (>3d); 50 bp paired end read; >10^7 reads/run	Applied Biosystems/Agencourt (SOLiD)
Single molecule sequencing-by-synthesis (still in development)		Fluorescence at just the active site of a single immobilized DNA polymerase enzyme measured using zero mode waveguide.		Pacific BioSciences
		Sequencing-by-synthesis using random immobilized DNA fragments and high intensity fluorescent emitters.	25 bp read; >10^7 reads/run	Genovoxx (reagents and surface only)
		Fluorescent resonance energy transfer (FRET) to channel energy via GFP-DNA polymerase to bound nucleotides.	~1000 Mb/run (~1d); ~10^6 reads/run	Helicos BioSciences
Direct "reading" of single molecule (largely in early development)		Conductance in nanopores. Differing physiochemical properties of nucleotide residues alter electric field as DNA is drawn through a pore. Protein pores, engineering nanochannels, and nanopore array systems are in development.	A DNA molecule is passed through field at rate of over 1000 bases per second	Agilent: LingVitae AS
		Real-time reading of the DNA polymerase reaction using FRET.	~10^5 Mb/run	Visigen: Li-Cor
		Force spectroscopy to follow the molecular mechanics of DNA synthesis.		



Science
AAAS
AAAS/*Science* Business Office Publication