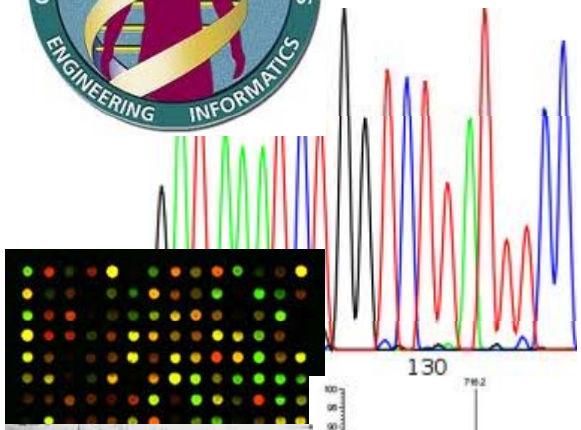
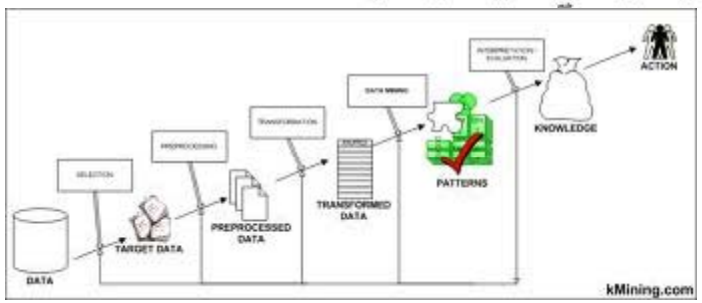
Curso de Posgrado: Alimentos Funcionales

BIONFORMATICA EN EL AMBITO DE LA NUTRIGENOMICA

Norma Paniego
Instituto de Biotecnología, CICVyA, CNIA INTA Castelar





CONCEPTOS GENERALES

NUTRIGENOMICA

NUTRIGENETICA

GENOMICA

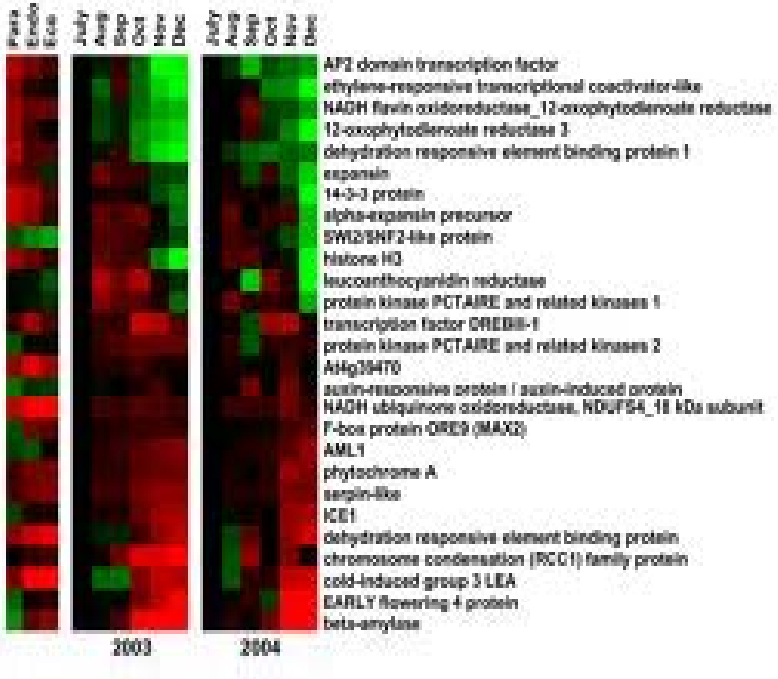
POST-GENOMICA

BASES DE DATOS

EXTRACCIÓN DEL CONOCIMIENTO

Nutrigenómica & Nutrigenética

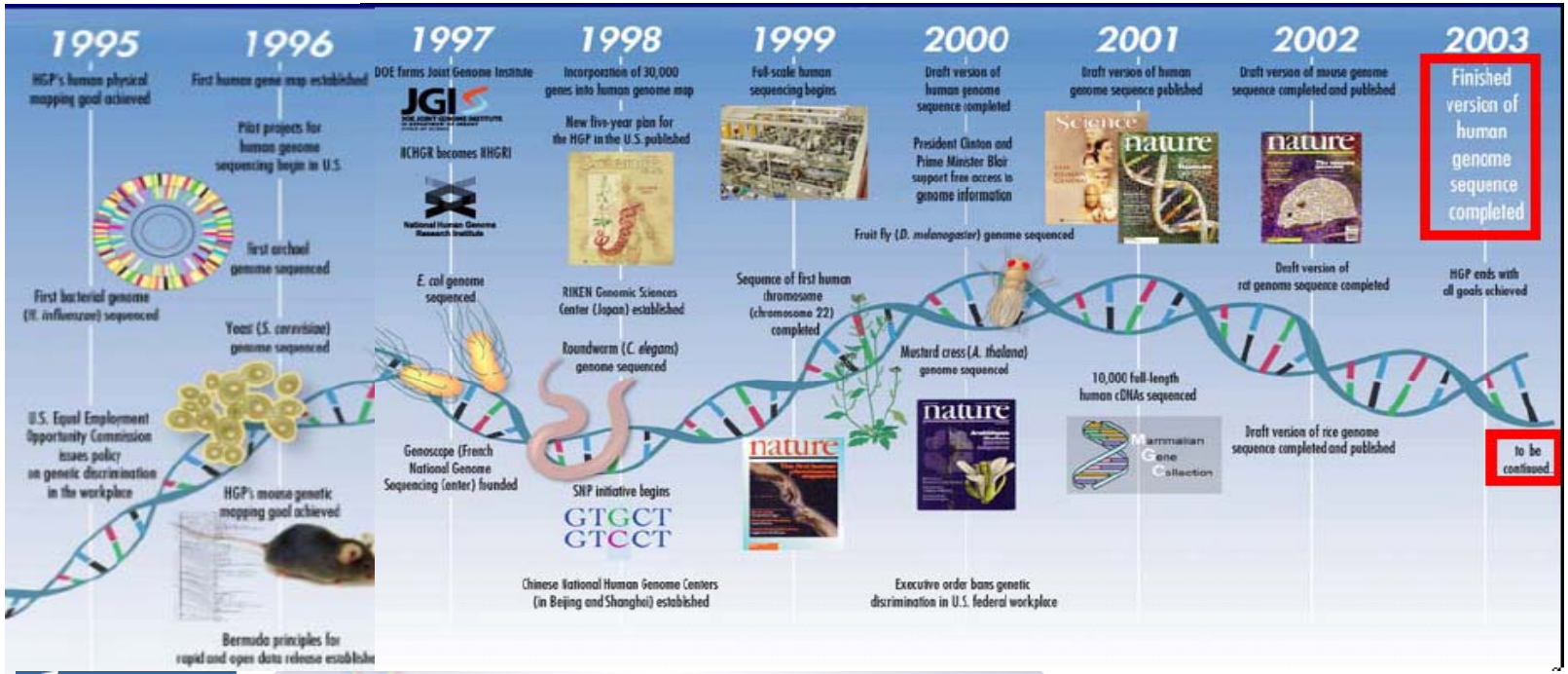
La **nutrigenómica** es la ciencia que estudia la **expresión de los genes** en relación con la nutrición y el desarrollo de enfermedades asociadas a dicha expresión.



La **nutrigenética** es la ciencia que estudia como el **genotipo** del individuo interacciona con ingesta de nutrientes.

Indian Journal of Biotechnology
Vol7, Jan 2008, pp 32-40

Proyectos genómicos



NCBI ENTREZ Genome Project connection information discovery

Entrez | Genome | Gene | Nucleotide | Protein | PopSet | PubMed | Taxonomy

Search Genome Project Go Clear

Properties of Eukaryotic Genome Sequencing Projects

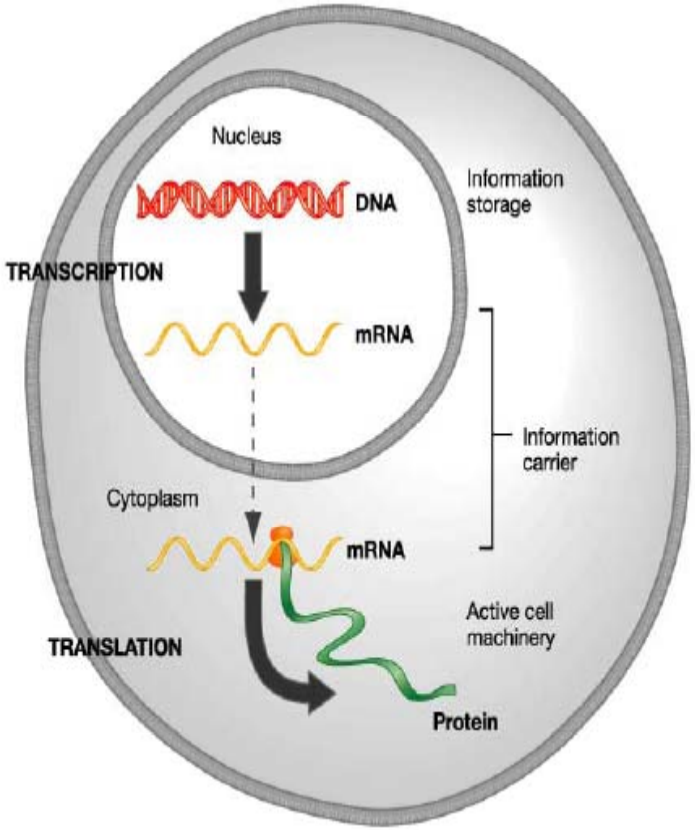
Organism Group: All Eukaryotes Sequencing Status: All Sequencing Method: All Go Reset

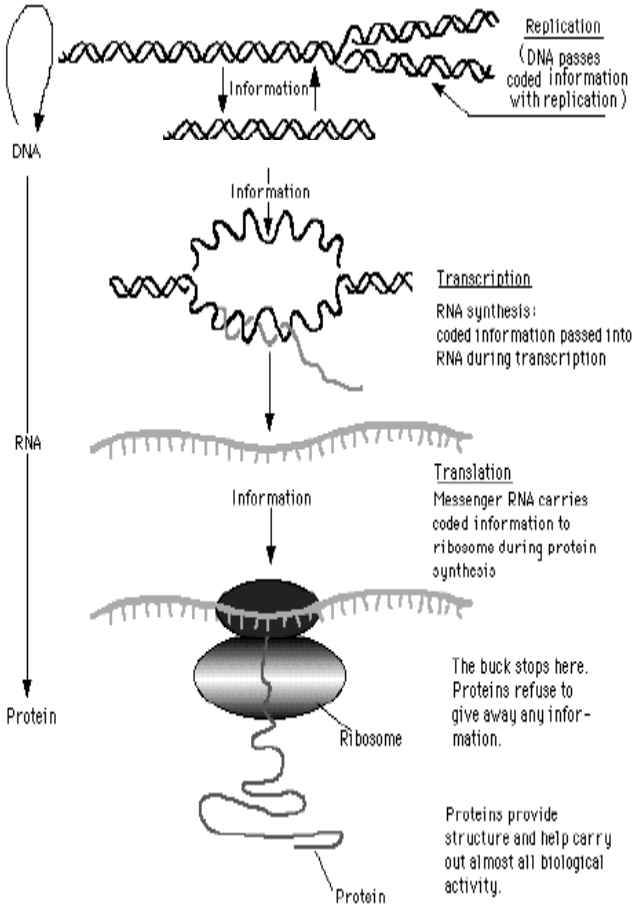
Abbreviations: GB - GenBank Accessions; PM - PubMed; R - RefSeq Accessions; G - Entrez Gene; T - Trace Archive; B - BLAST; M - Map Viewer; F - FTP Sites

482 Eukaryotic Genome Sequencing Projects Selected: Complete - 23, Assembly - 231, In Progress - 228 save

731 Eukaryotic Genome Sequencing Projects Selected: Complete - 23, Assembly - 327, In Progress - 381 save

El Dogma Central de la Biología Molecular





Niveles básicos de Información biológica

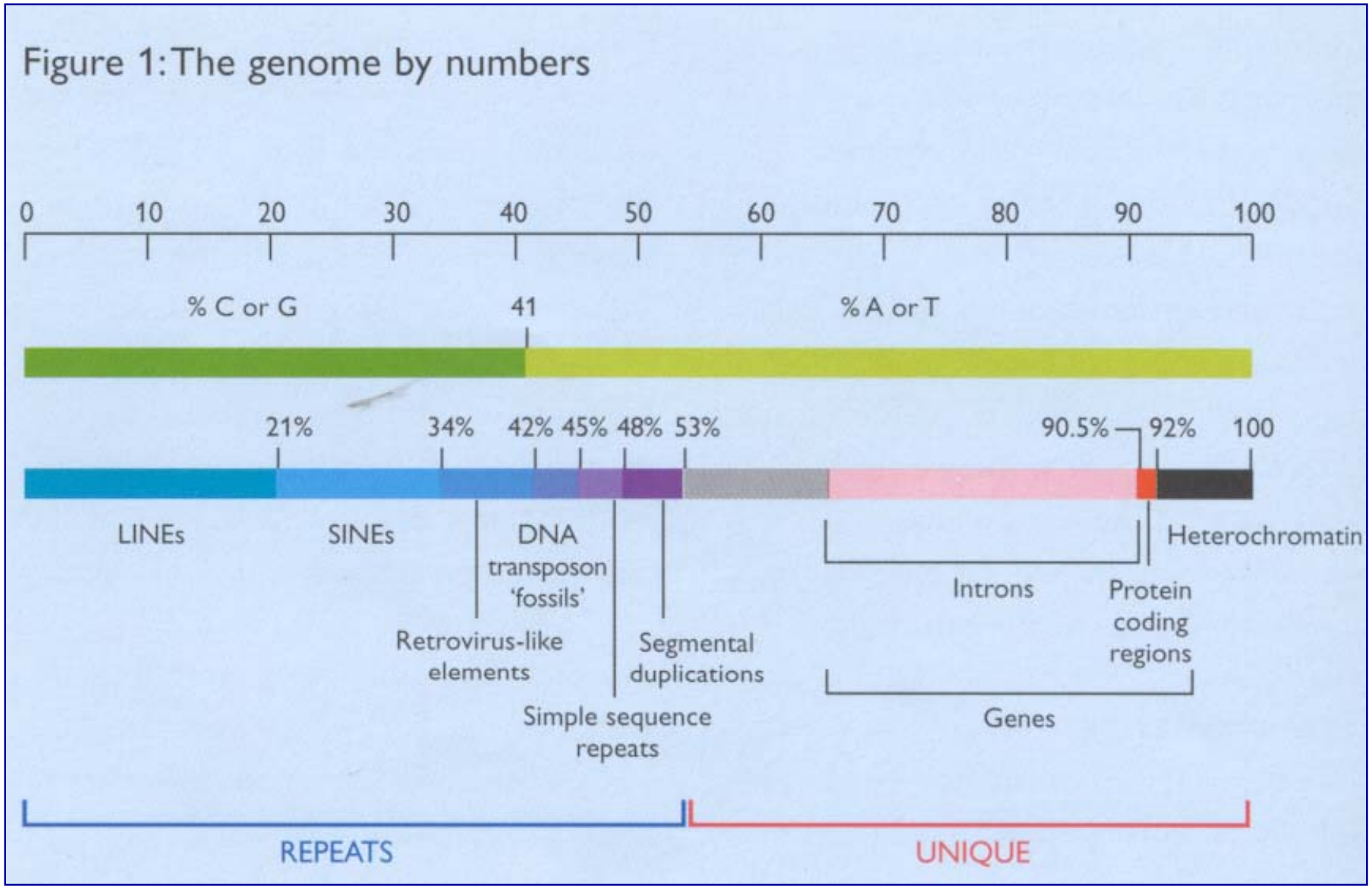
- **Genoma**: es la totalidad de la información genética que posee un organismo en particular.
- **Transcriptoma**: la parte del genoma que se expresa en una célula en una etapa específica de su desarrollo.
- **Proteoma**: la totalidad de las proteínas codificadas por un genoma.
- **Metaboloma**: la totalidad de pequeñas moleculas o metabolitos que se pueden encontrar en una muestra biológica, tal como un organismo.

La era de la genómica

La genómica se ha desarrollado como consecuencia de los avances en Biología Molecular e Informática.

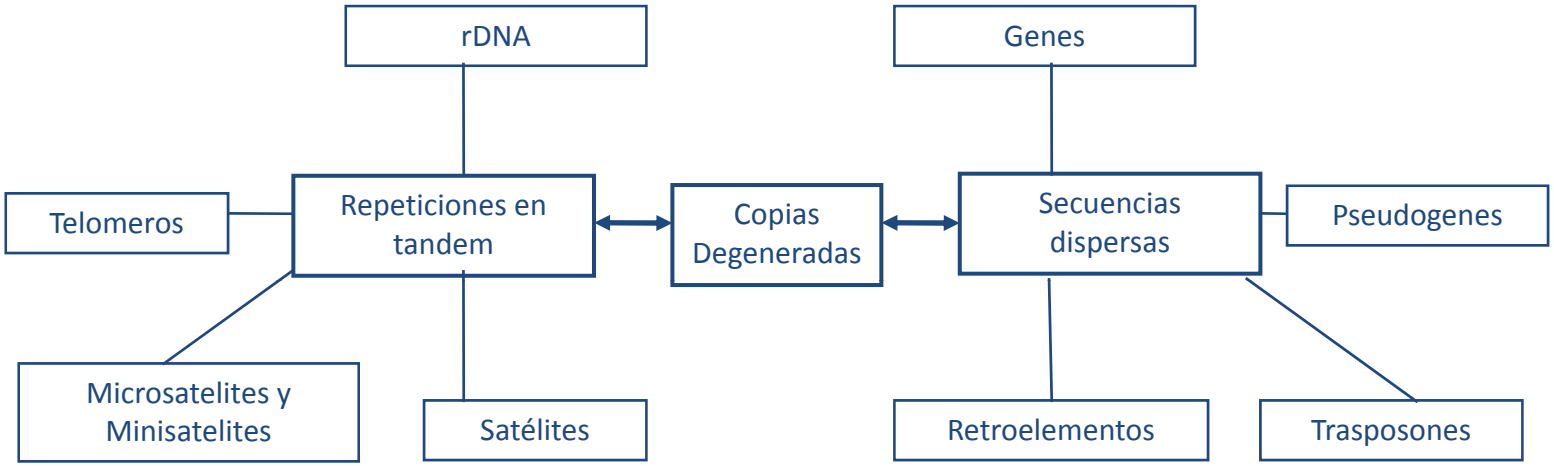
La introducción y popularización de las tecnologías de alta procesividad ha cambiado drásticamente la manera en que se abordan los problemas biológicos y se prueban las hipótesis.

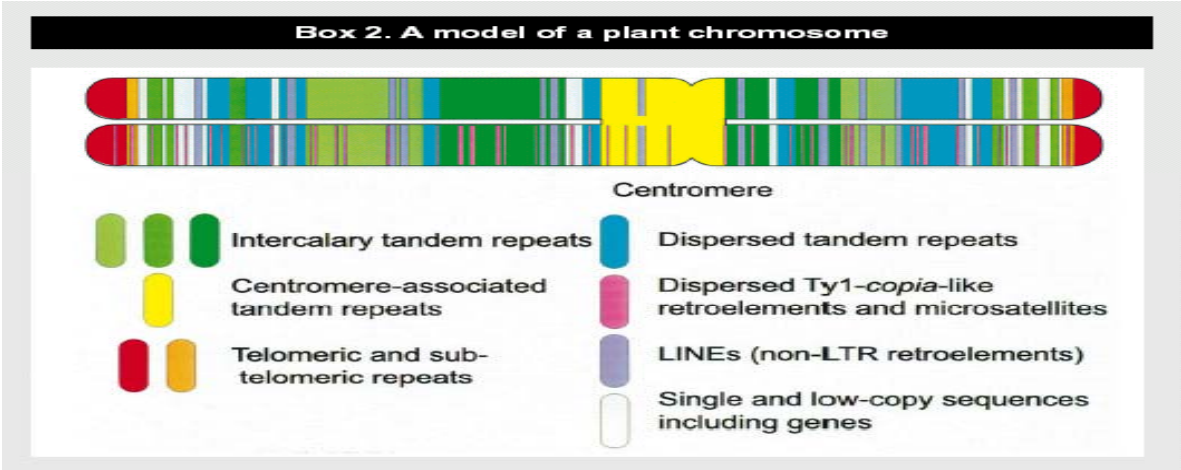
Inicialmente los proyectos genómicos se focalizaban en la secuencia completa del genoma, actualmente incluyen el análisis de la expresión y función de genes, proteínas, metabólitos entre otros.



The Genome by Numbers, the Welcome Trust.

Contenido de secuencias de ADN en el genoma vegetal





Schmid T y Heslop-Harrison JS (1998) Trends in Plant Sci., 3, 195-199

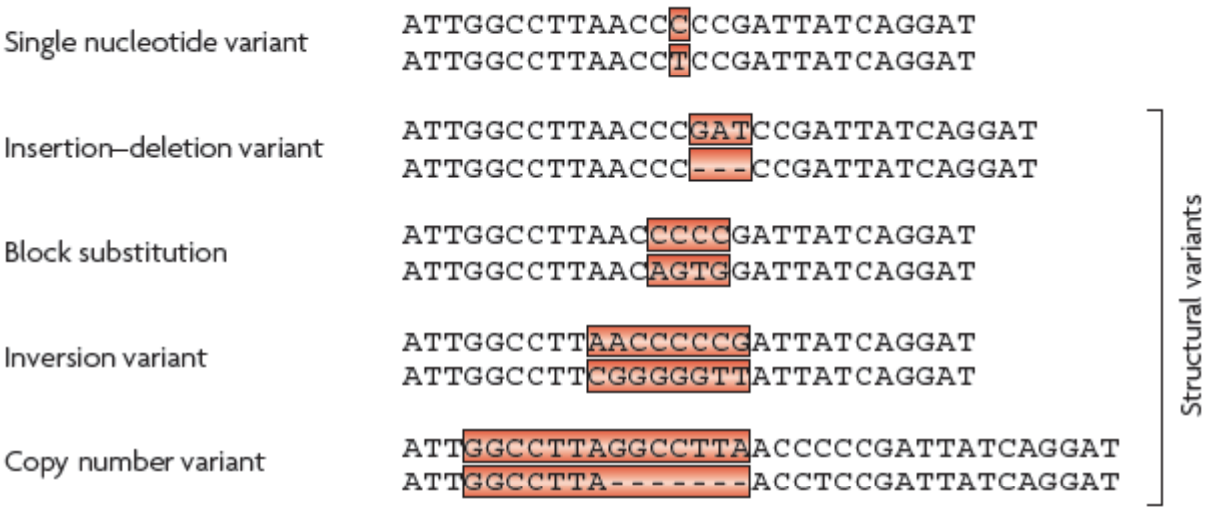
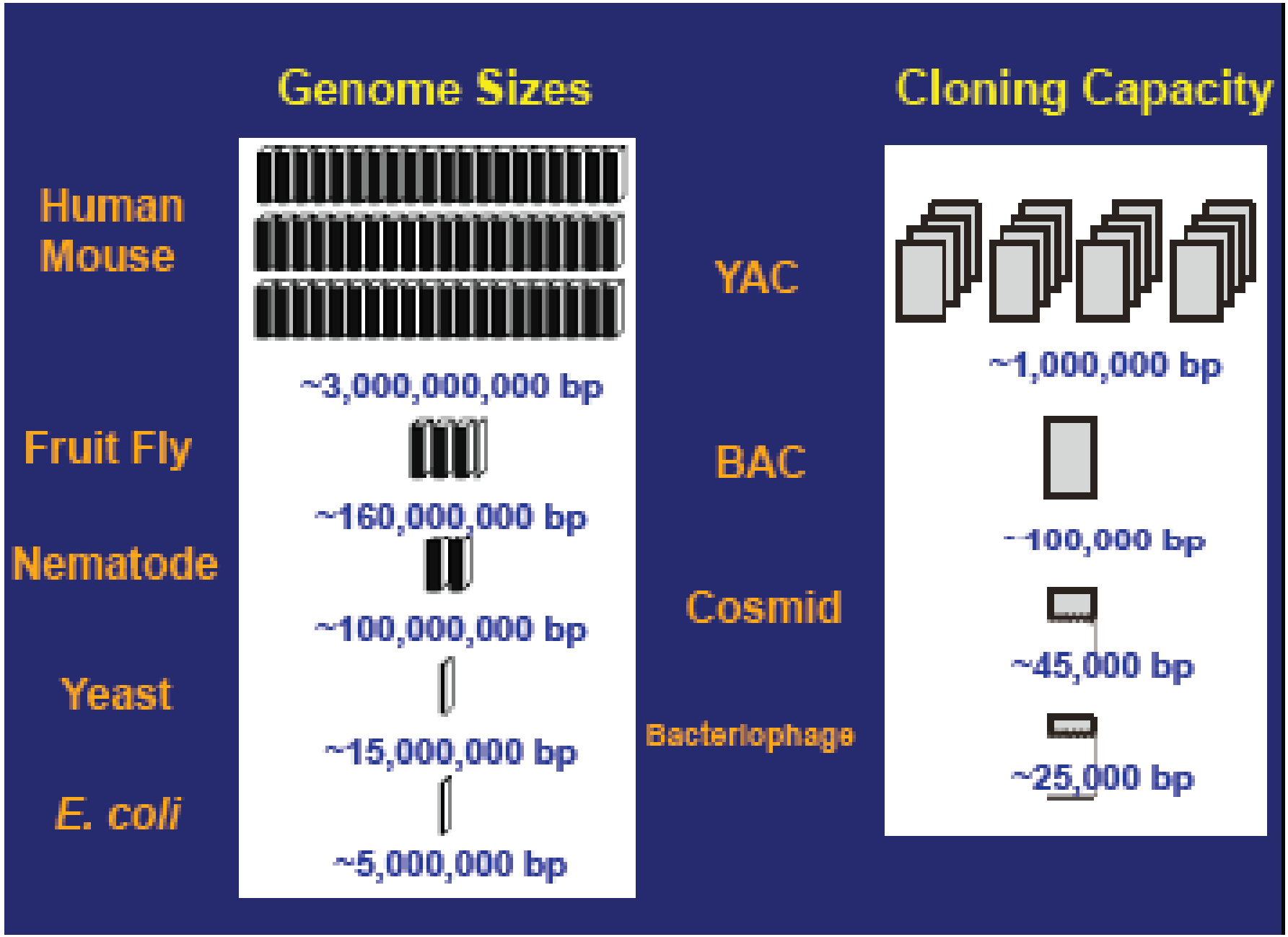


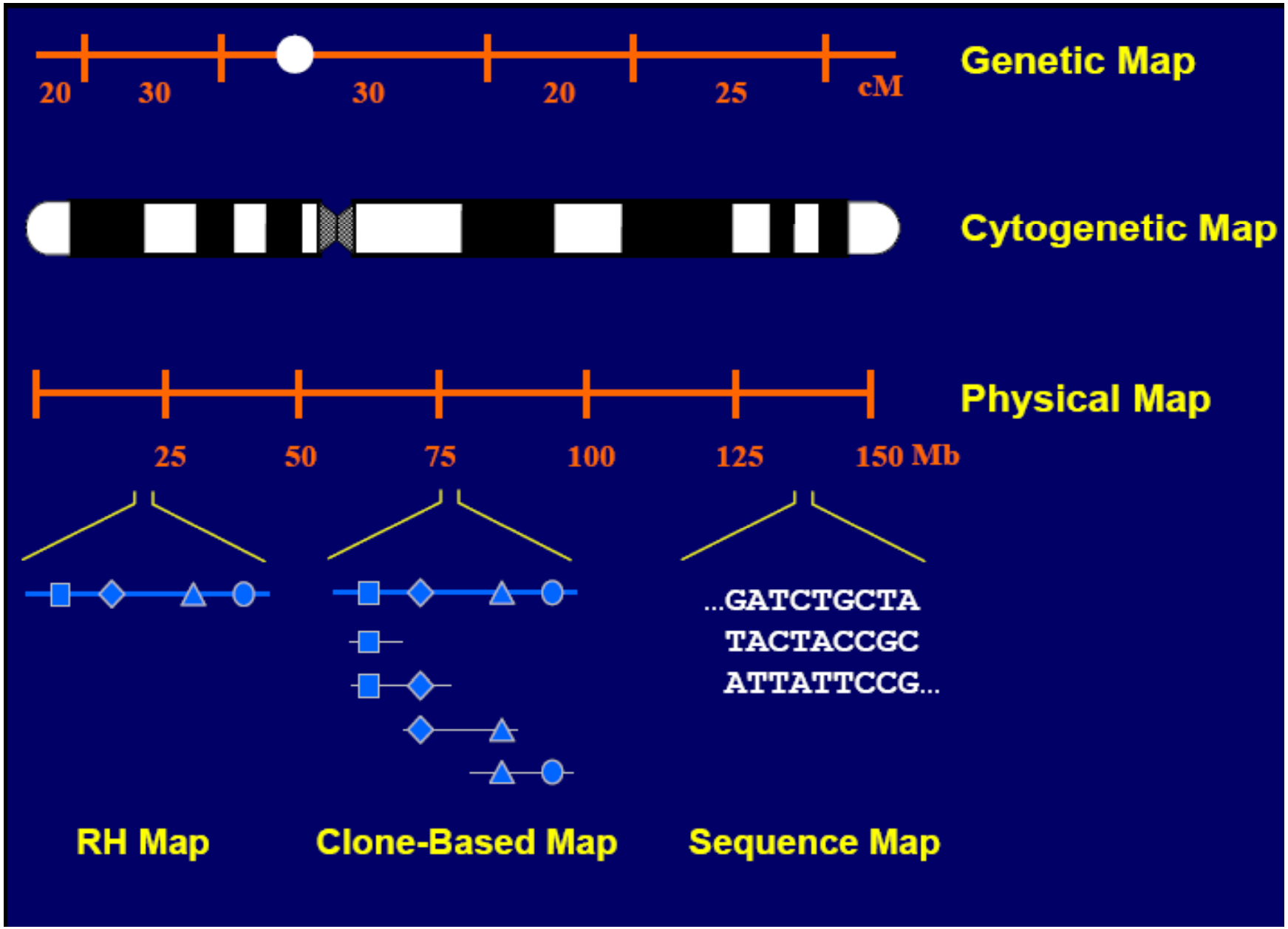
Figure 1 | **Classes of human genetic variants.** The nomenclature used to describe the various types of structural variants is not yet standard[121]. Here, the terminology used aims to describe the nucleotide composition of the variant and distinguish it from other types of variants. Single nucleotide variants are DNA sequence variations in which a single nucleotide (A, T, G or C) is altered. Insertion–deletion variants (indels) occur when one or more base pairs are present in some genomes but absent in others. They are generally composed of only a few bases but can be greater than 80 kb in length[11]. Block substitutions describe cases in which a string of adjacent nucleotides varies between two genomes. An inversion variant is one in which the order of the base pairs is reversed in a defined section of a chromosome. A well-characterized inversion variant that has been described in humans involves a section of chromosome 17 in which a ~900 kb interval is in the reverse order in approximately 20% of individuals with Northern European ancestry[122]. Copy number variants occur when identical or nearly identical sequences are repeated in some chromosomes but not others. The largest copy number variant identified in the Venter genome[11] was almost 2 Mb in length.

Fraser et al, 2009 Nat Rev Genet 10:241-251



The Genomic Landscape: ***circa 2010, Eric Green***

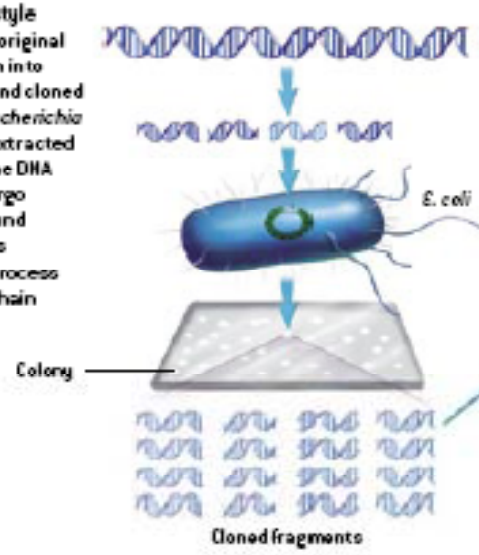
Como se estudian los genomas: técnicas de mapeo



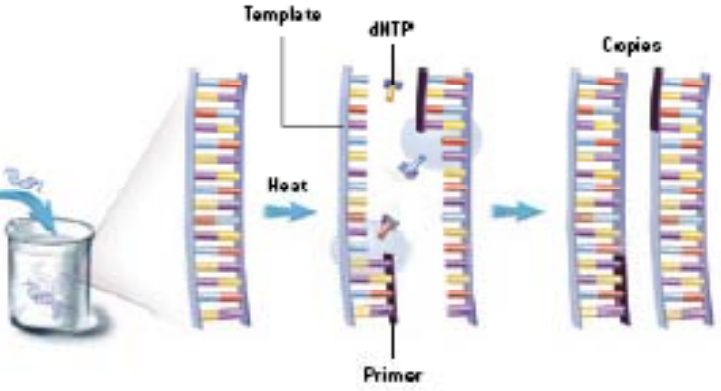
The Genomic Landscape: ***circa 2010, Eric Green***

Método de secuenciación

1 Before Sanger-style sequencing, an original DNA strand is broken into smaller fragments and cloned within colonies of *Escherichia coli* bacteria. Once extracted from the bacteria, the DNA fragments will undergo another massive round of copying, known as amplification, by a process called polymerase chain reaction (PCR).



2 During PCR, fragments are heated so they will separate into single strands. A short nucleotide sequence called a primer is then annealed to each original template. Starting at the primer, polymerase links free-floating nucleotides (called dNTPs) into new complementary strands. The process is repeated over and over to generate millions of copies of each fragment.



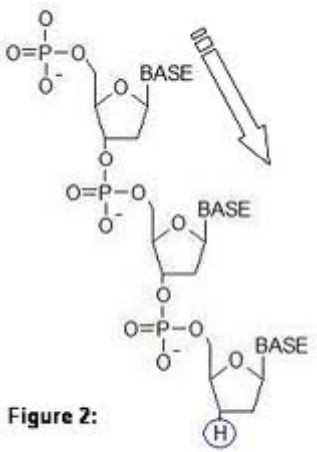


Figure 2:

3 Single-stranded fragments are next tagged in a process similar to PCR but with fluorescently labeled terminator nucleotides (ddNTPs) added to the mixture of primers, polymerase and dNTPs. Complementary strands are built until by chance a ddNTP is incorporated, halting synthesis. The resulting copy fragments have varying lengths and a tagged nucleotide at one end.

4 Capillary electrophoresis separates the fragments, which are negatively ch drawing them toward a positively charged pole. Because the shortest fragm fastest, their order reflects their size and their ddNTP terminators can thus be "r template's base sequence. Laser light activates the fluorescent tags as the frag detection window, producing a color readout that is translated into a sequence.

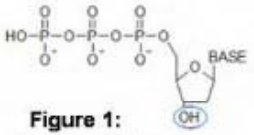
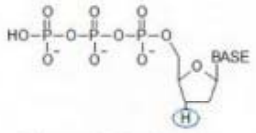
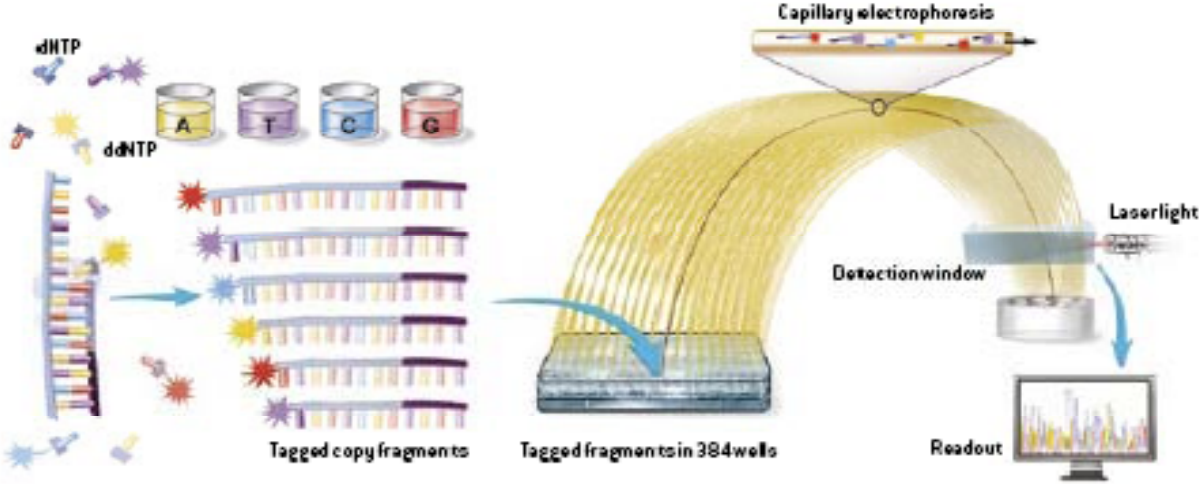


Figure 1:

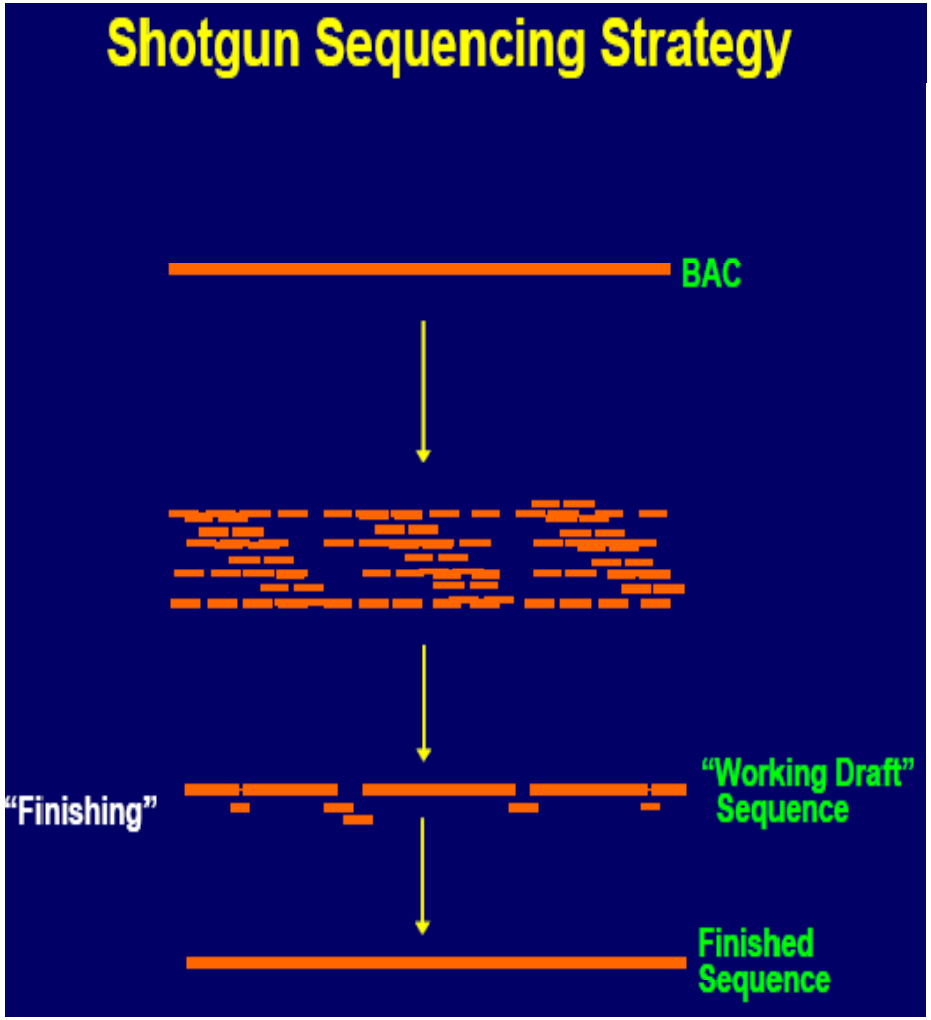
deoxynucleotide triphosphate

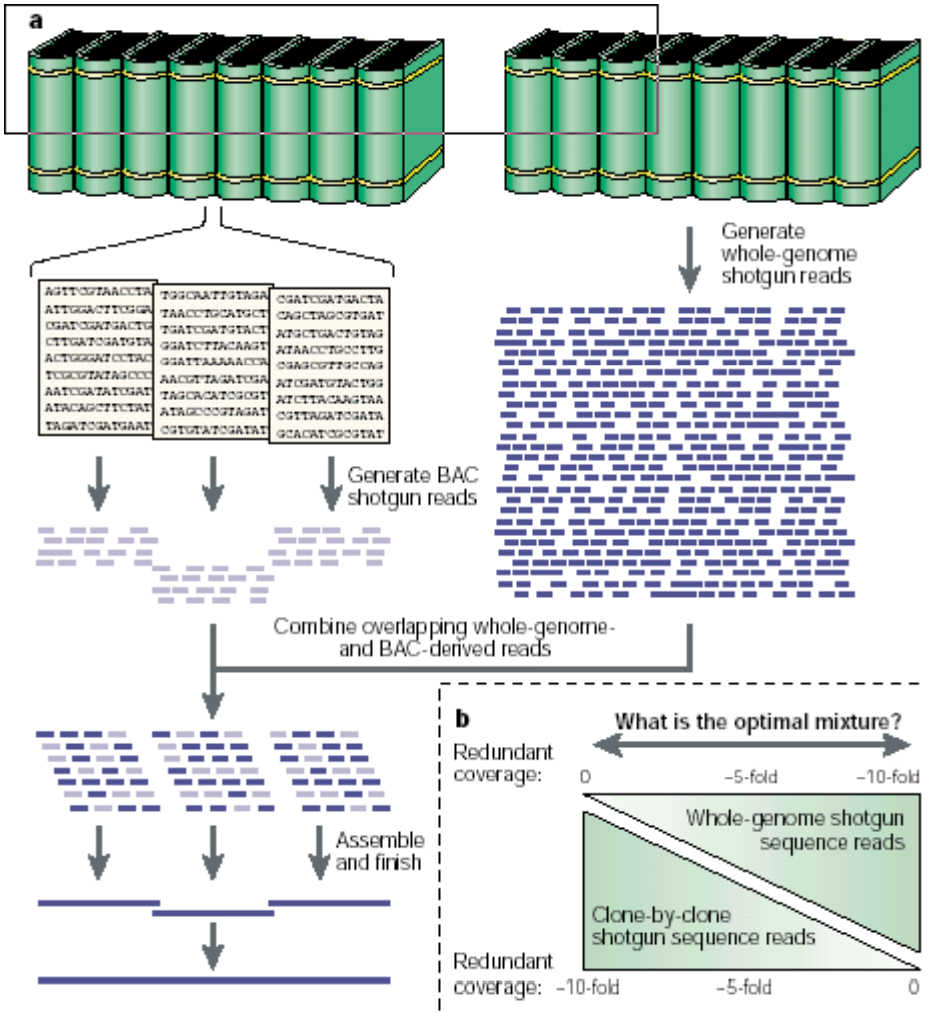


dideoxynucleotide triphosphate



Estrategias alternativas de secuenciación

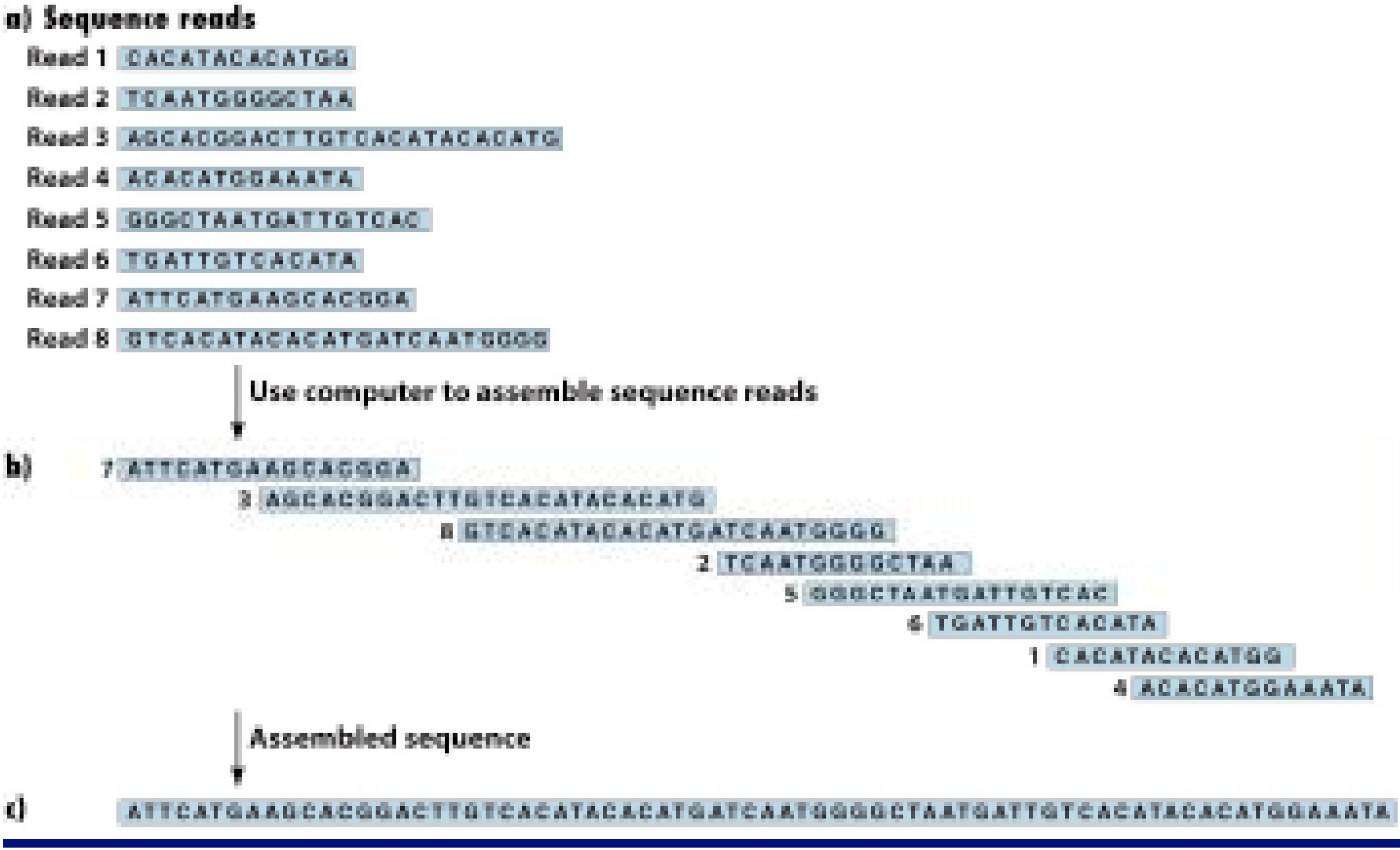




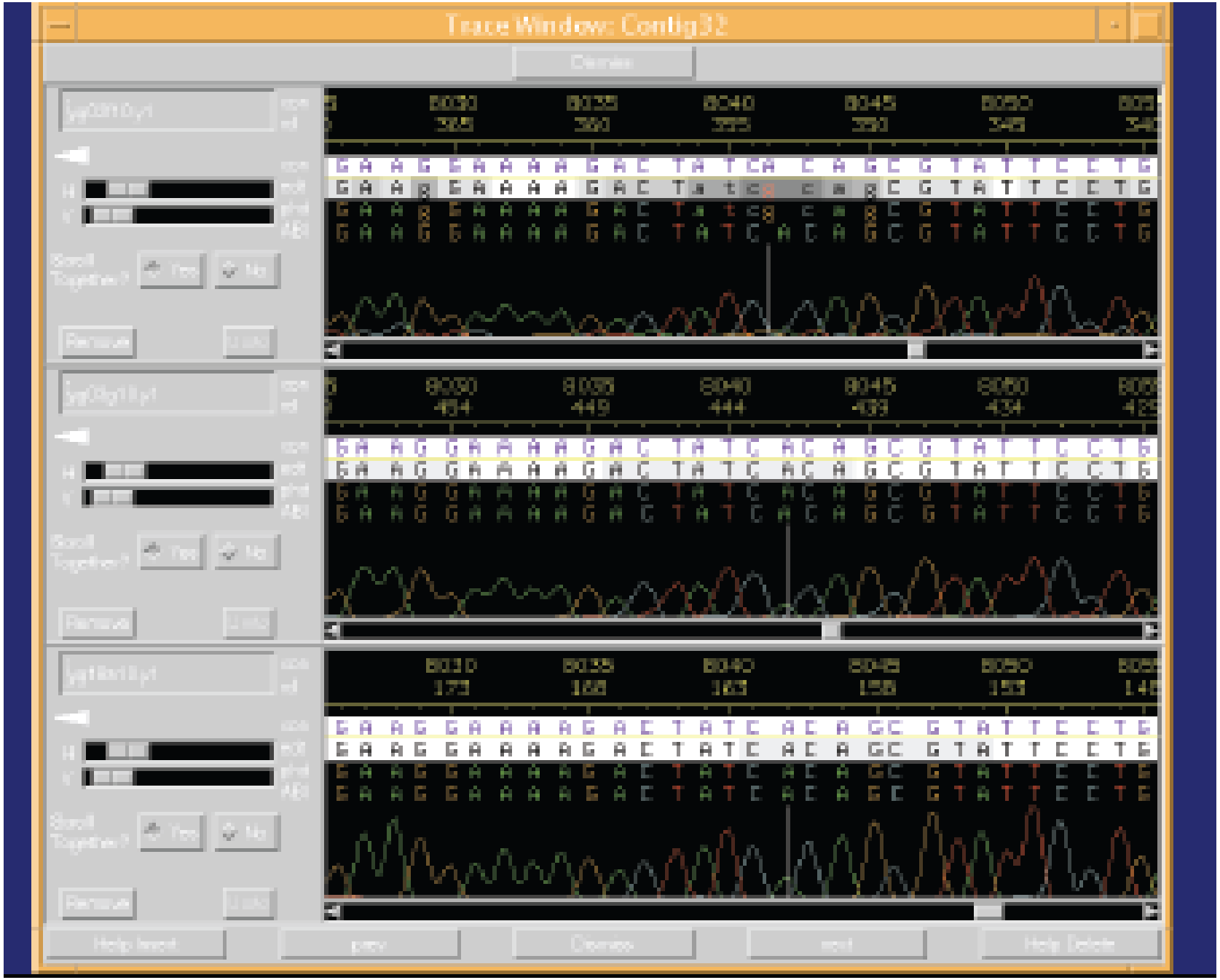
The Genomic Landscape: ***circa 2010, Eric Green***

Equipamiento automatizado

The Genomic Landscape: ***circa 2010, Eric Green***

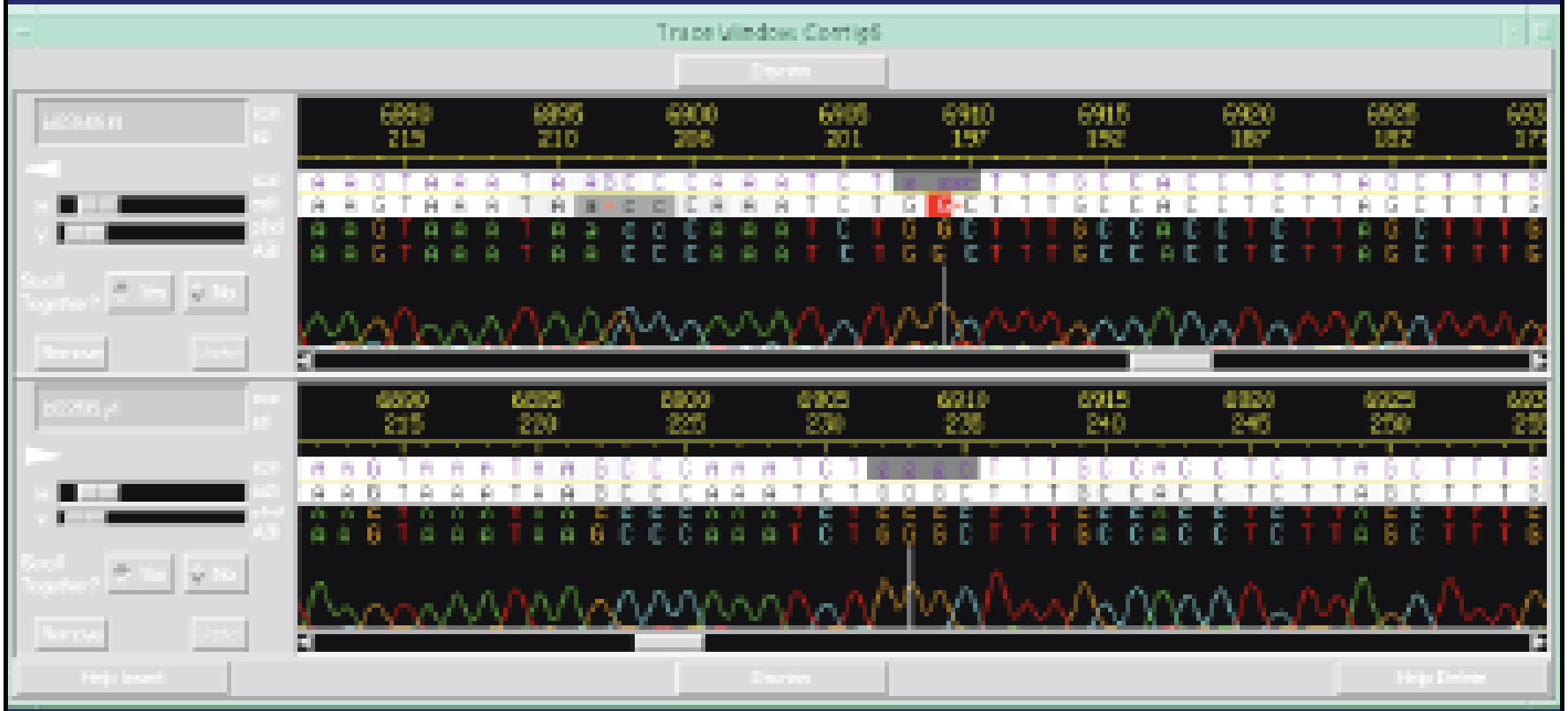


The Genomic Landscape: ***circa 2010, Eric Green***



The Genomic Landscape: ***circa 2010, Eric Green***

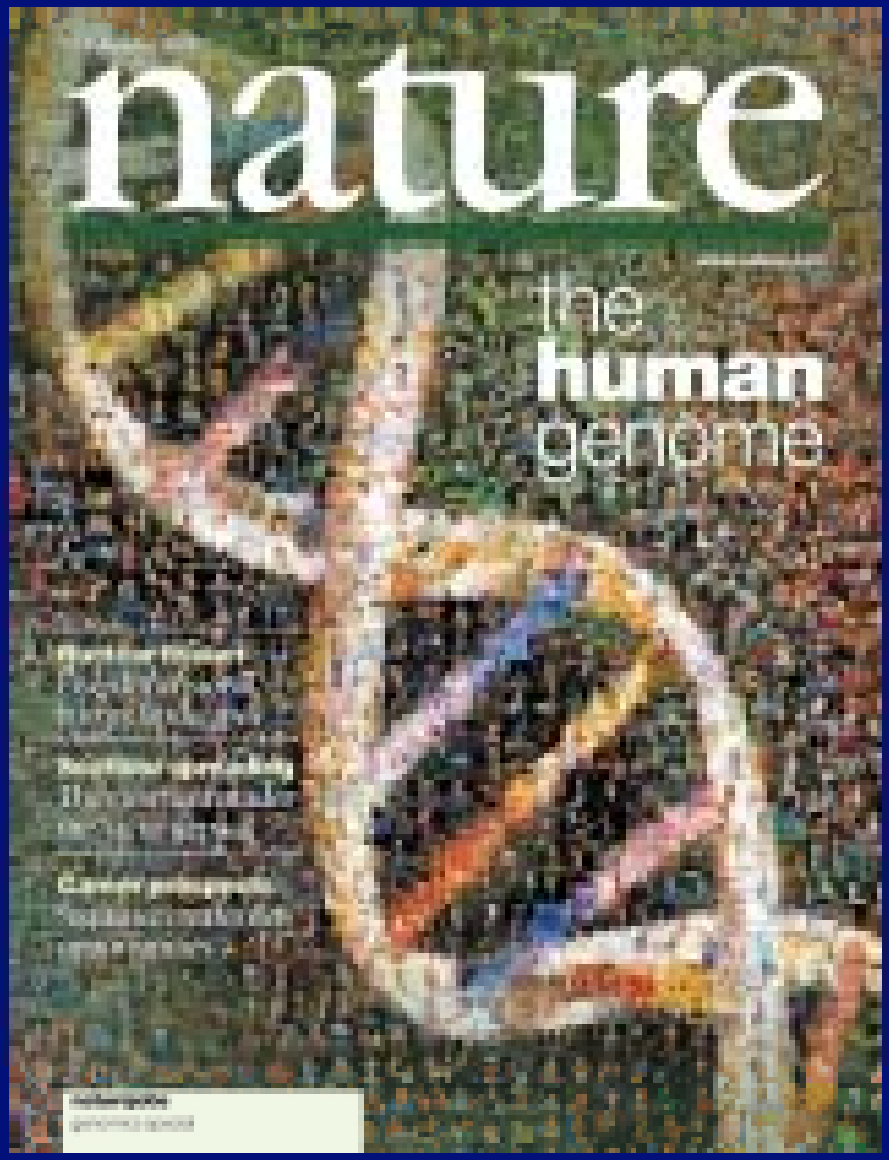
Sequence Finishing: Resolving Ambiguities



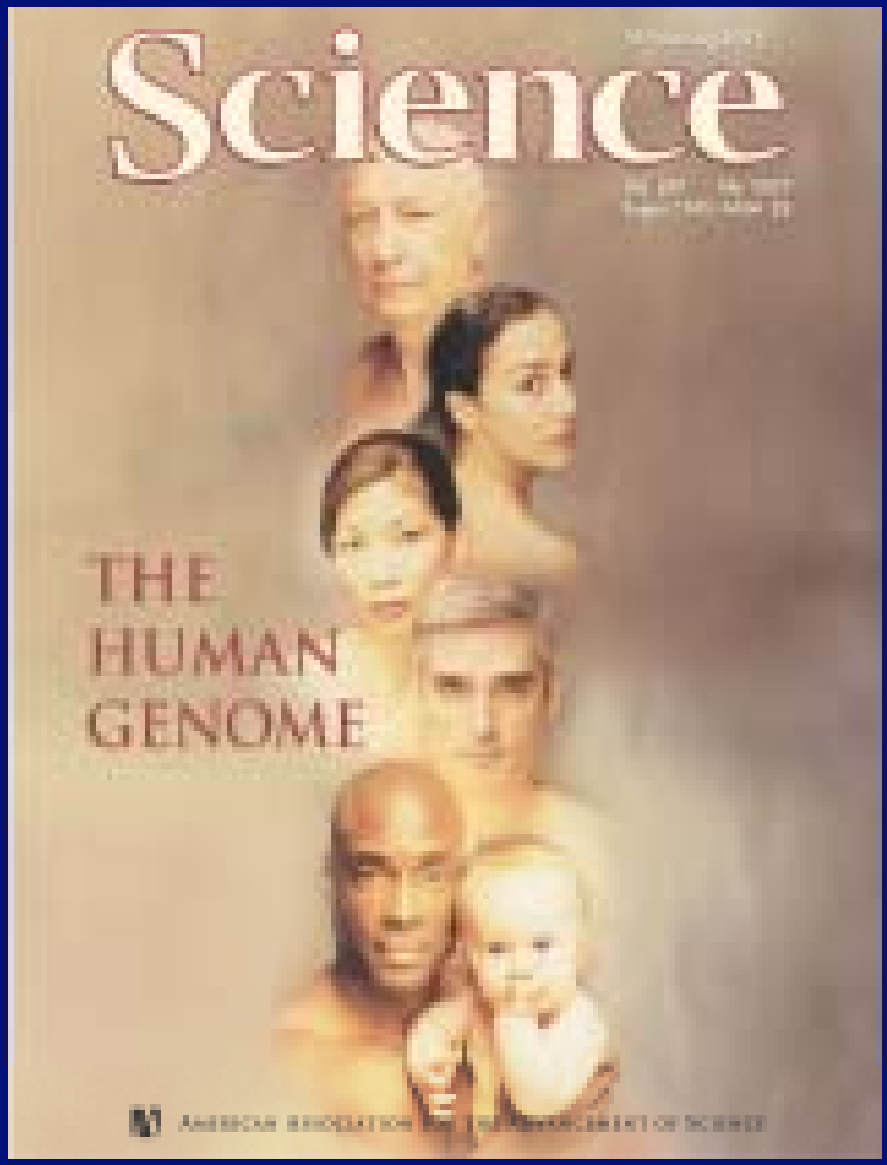
*** Sequence Finishing: Remains Relatively Expensive ***

The Genomic Landscape: ***circa 2010, Eric Green***

February, 2001 Draft Sequence



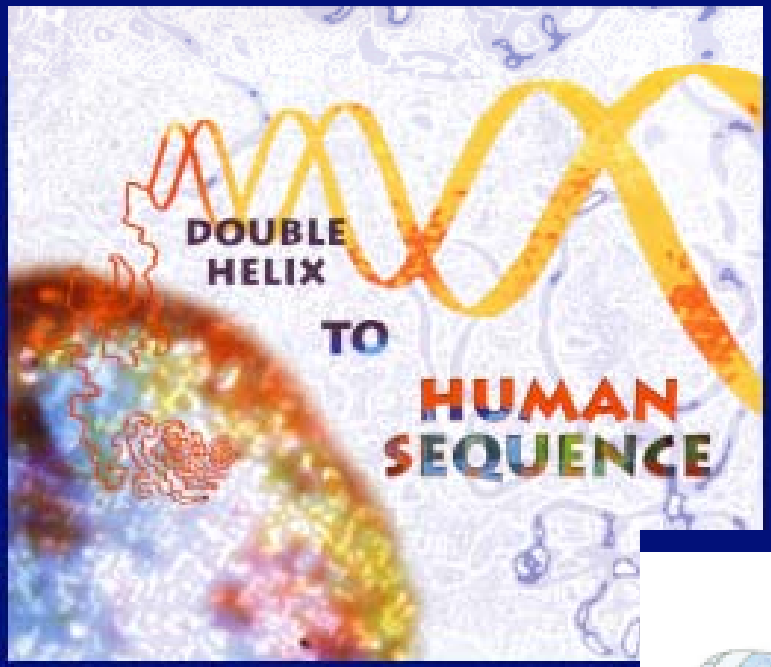
International Human Genome Sequencing Consortium (2001)

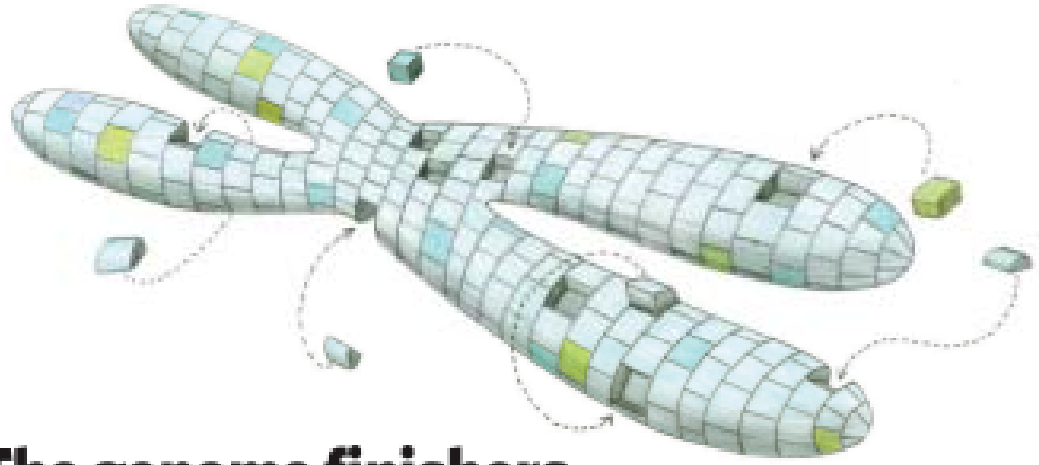


Venter et al. (2001)

The Genomic Landscape: *circa 2010, Eric Green*

April, 2003 Completion



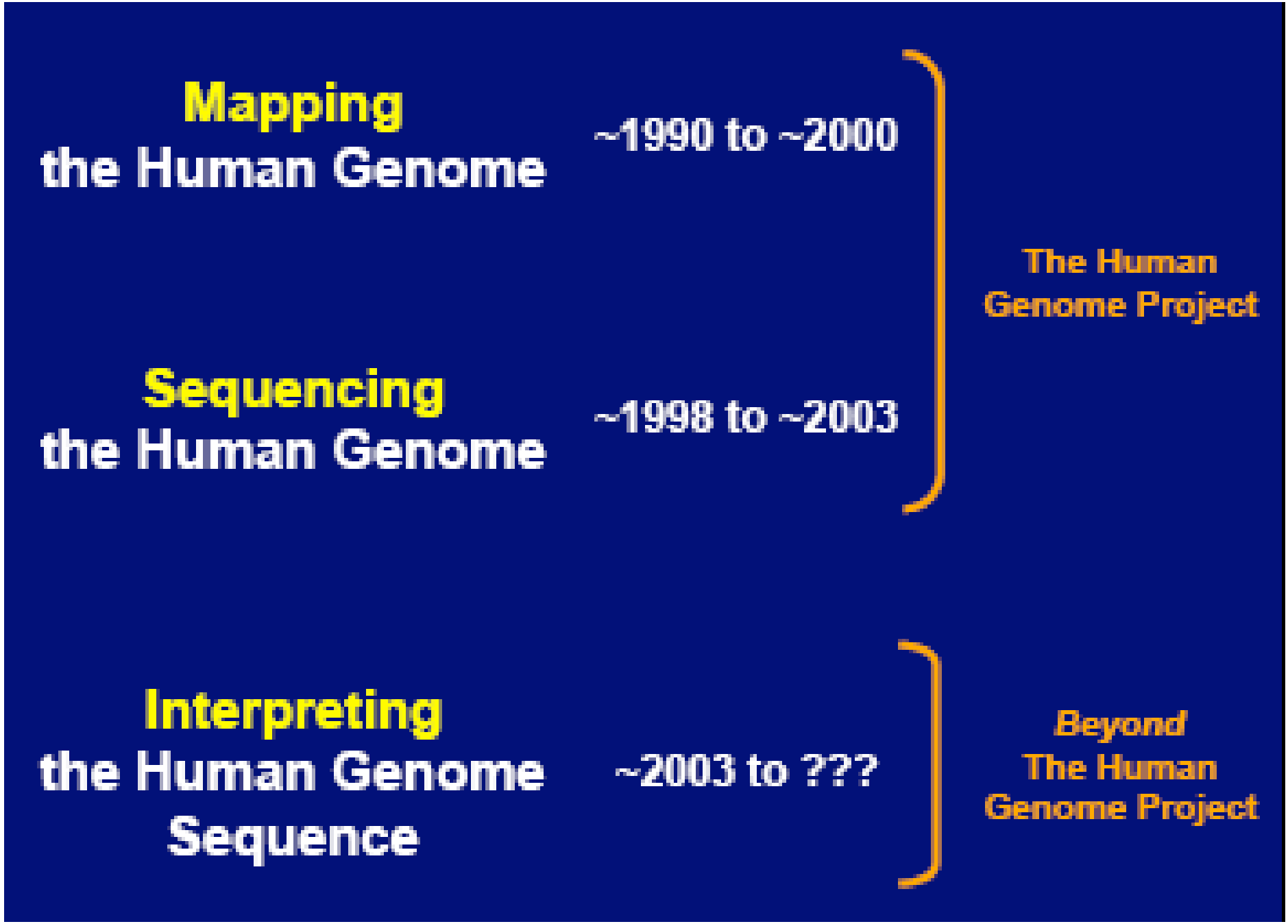


The genome finishers

Dedicated scientists are working hard to close the gaps, fix the errors and finally complete the human genome sequence. **Elie Dolgin** looks at how close they are.

***Nature* (2009)**

The Genomic Landscape: *circa 2010, Eric Green*



The Genomic Landscape: ***circa 2010, Eric Green***

The Human Genome… by the Numbers

~5% of Human Genome Sequence is Constrained Across Mammals (and Presumed Functional)

- 5% of 3B Bases = ~150M Bases
- Do NOT Yet Know the Position of these ~150M Functional Bases
- Lower Bound for the Amount that is Funotional

~1.5% Encodes for Protein (Genes)

- Corresponds to ~18-22K Genes
- Many More than ~22K Different Proteins
- Good Inventory at Present

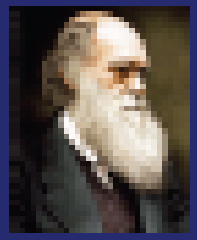
~3.5% Functional But Non-Coding

- Gene Regulatory Elements
- Chromosomal Functional Elements
- Undiscovered Functional Elements (NOT Yet in Textbooks!)
- Poor Inventory at Present

The Genomic Landscape: *circa 2010, Eric Green*

Comparative Sequence Analysis

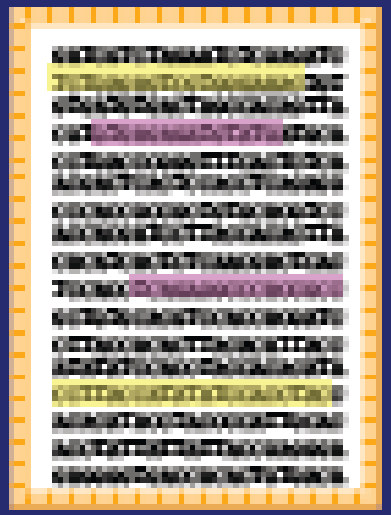
Using the 'Experiments of Evolution' to Decode the Human Genome



Species A

Species B

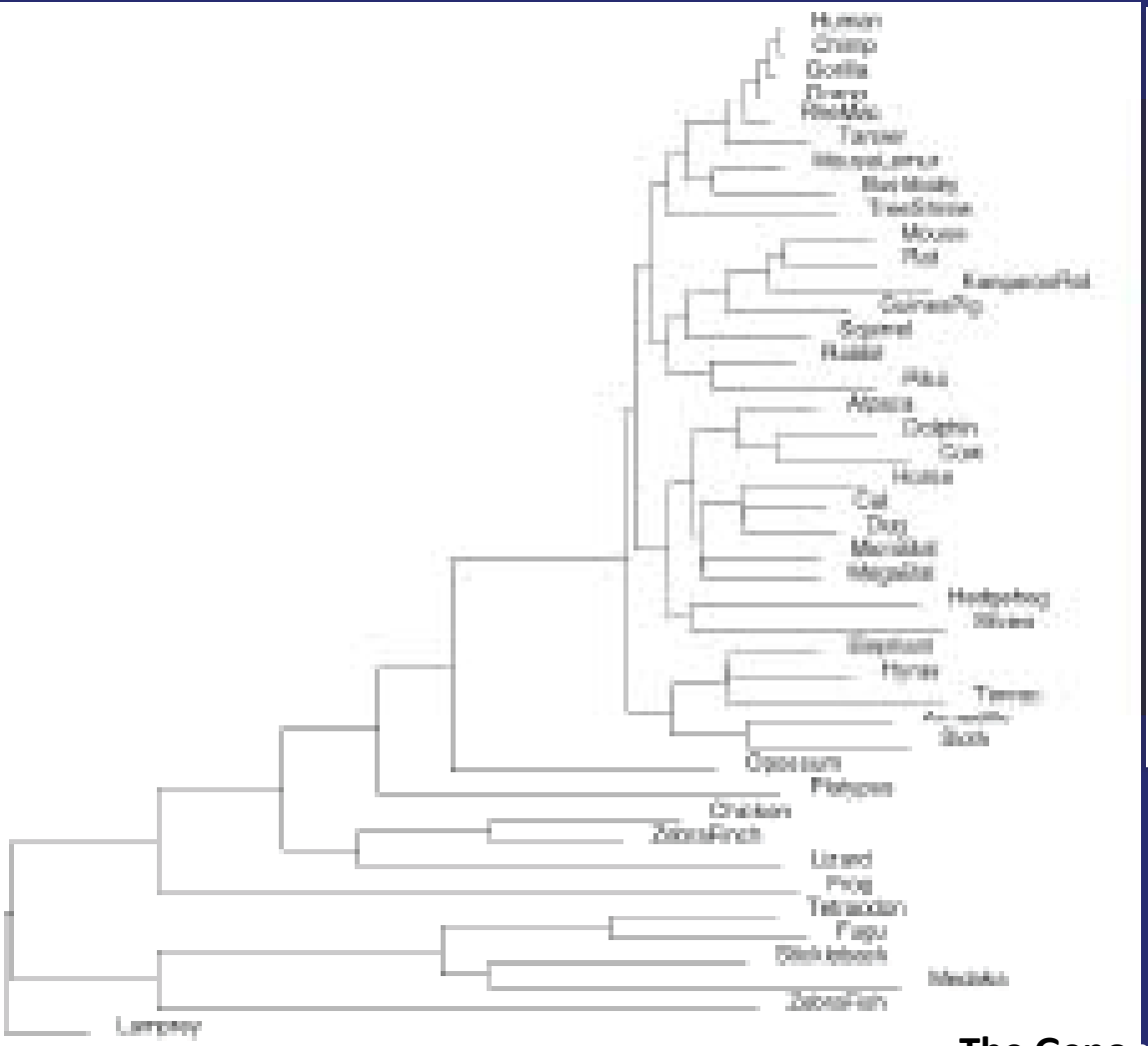
Compare



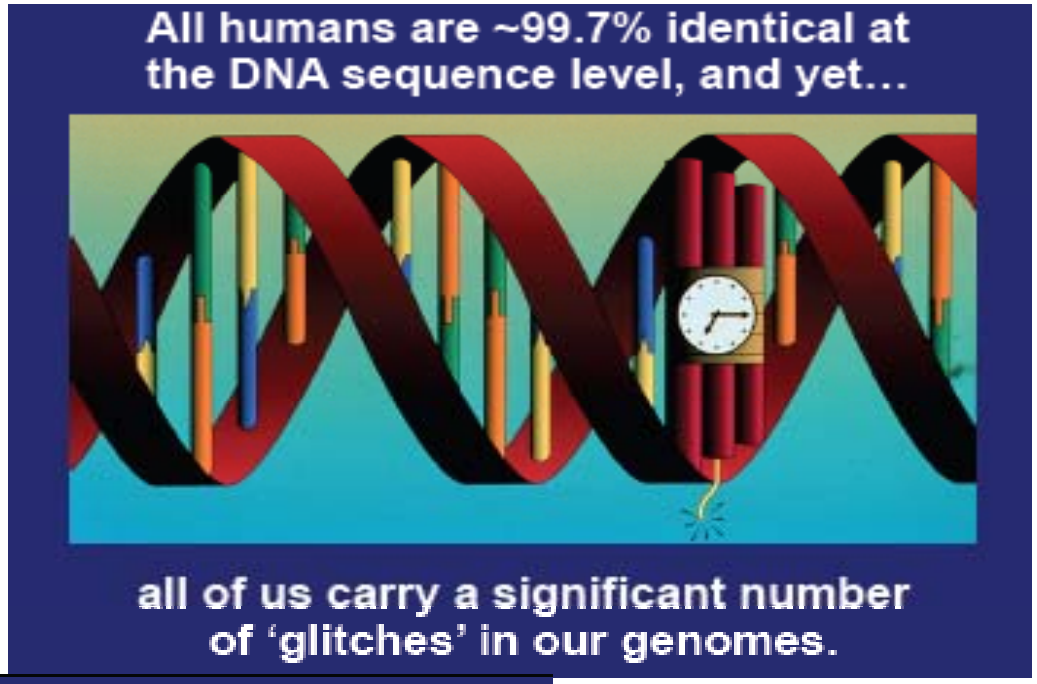
Sequences in Common (i.e., 'Conserved' or 'Constrained')

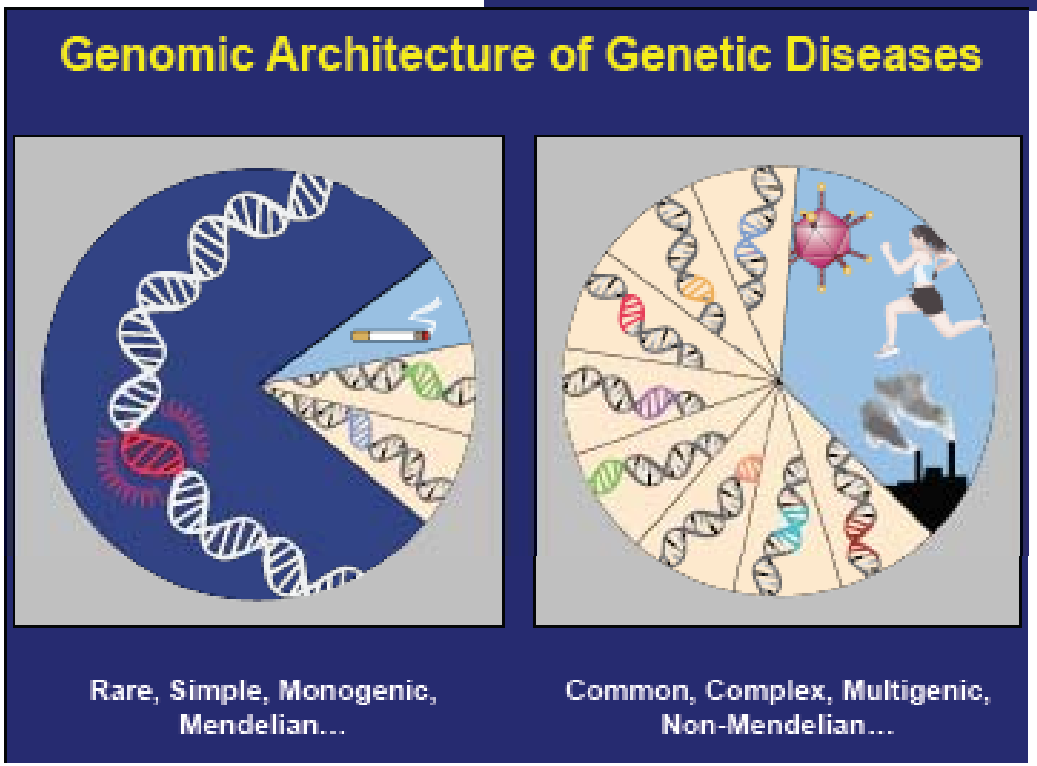
The Genomic Landscape: *circa 2010*, *Eric Green*

22 Additional Mammalian Genome Sequences (@ Low Redundancy)



The Genomic Landscape: *circa 2010, Eric Green*





The Genomic Landscape: ***circa 2010, Eric Green***


International
HapMap
Project
International HapMap Project
Home | About the Project | Data | Publications | Tutorial


THE HUMAN VARIOME PROJECT
sharing data · reducing disease
Home
About
Activities
Publications
Recommendations
Meetings
News
Links
Blogs
Members

Nuevas tecnologías de secuenciación de ADN

Eric D. Green, M.D., Ph.D. — Techniques for Genome Mapping & Sequencing

Genome sequencing in microfabricated high-density picolitre reactors

Margulies M et al. (2005)

Solexa Ltd

Bennett (2004)

Toward the $1000 human genome

Bennett et al. (2005)

Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome

Jay Shendure,[1*] Gregory J. Porreca,[1*] Nikos B. Reppas,[1] Xiaoxia Lin,[1] John P. McCutcheon,[2,3] Abraham M. Rosenbaum,[1] Michael D. Wang,[1] Kun Zhang,[1] Robi D. Mitra,[2] George M. Church[1]

Shendure et al. (2005)

Perspective

Emerging technologies in DNA sequencing

Michael L. Metzker

Human Genome Sequencing Center and Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA

Demand for DNA sequence information has never been greater, yet current Sanger technology is too costly, time consuming, and labor intensive to meet this ongoing demand. Applications span numerous research interests, including sequence variation studies, comparative genomics and evolution, forensics, and diagnostic and applied therapeutics. Several emerging technologies show promise of delivering next-generation solutions for fast and affordable genome sequencing. In this review article, the DNA polymerase-dependent strategies of Sanger sequencing, single nucleotide addition, and cyclic reversible termination are di… and potential challenges these technologies face in their development for ultrafa…

Metzker (2005)

SCIENTIFIC AMERICAN

Alternatives to Toxic Tests on Animals

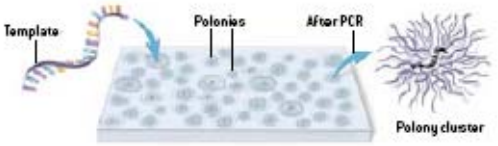
Church (2006)

Know Your DNA

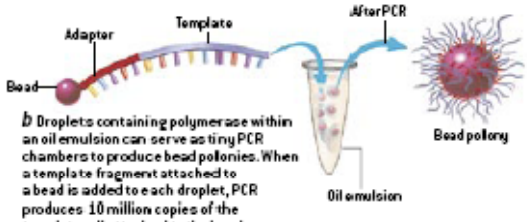
Inexpensive gene readers will soon unlock the secrets in your personal double helix

AMPLIFICATION

Because light signals are difficult to detect at the scale of a single DNA molecule, base-extension or ligation reactions are often performed on millions of copies of the same template strand simultaneously. Cell-free methods (*a* and *b*) for making these copies involve PCR on a miniaturized scale.



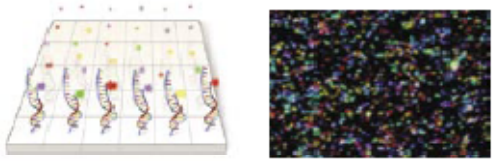
a Polonies—polymerase colonies—created directly on the surface of a slide or gel each contain a primer, which a template fragment can find and bind to. PCR within each polony produces a cluster containing millions of template copies.



b Droplets containing polymerase within an oil emulsion can serve as tiny PCR chambers to produce bead polonies. When a template fragment attached to a bead is added to each droplet, PCR produces 10 million copies of the template, all attached to the bead.

MULTIPLEXING

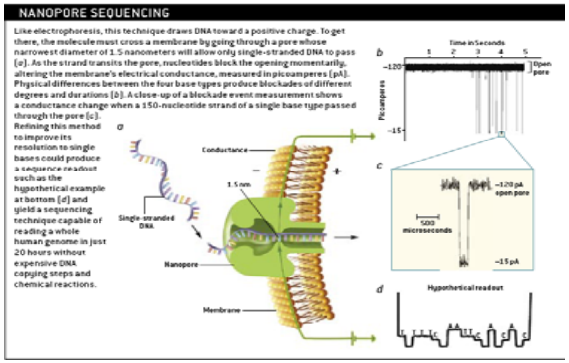
Sequencing thousands or millions of template fragments in parallel maximizes speed. A single-molecule base-extension system using fluorescent-signal detection, for example, places hundreds of millions of different template fragments on a single array (*below left*). Another method immobilizes millions of bead polonies on a gel surface for simultaneous sequencing by ligation with fluorescence signals, shown in the image at right below, which represents 0.01 percent of the total slide area.

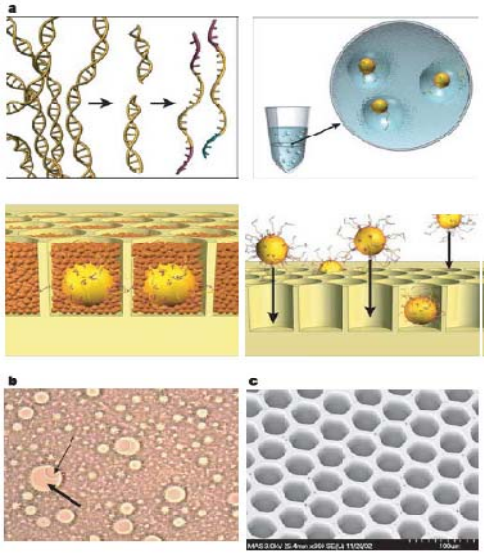


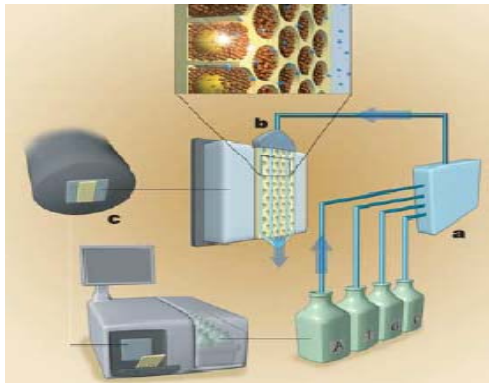
Single-molecule array — Bead polonies

NANOPORE SEQUENCING

Like electrophoresis, this technique draws DNA toward a positive charge. To get there, the molecule must cross a membrane by going through a pore whose narrowest diameter of 1.5 nanometers will allow only single-stranded DNA to pass (*a*). As the strand transits the pore, nucleotides block the opening momentarily, altering the membrane's electrical conductance, measured in picoamperes (pA). Physical differences between the four base types produce blockades of different degrees and durations (*b*). A close-up of a blockade event measurement shows a conductance change when a 150-nucleotide strand of a single base type passed through the pore (*c*). Refining this method to improve its resolution to single bases could produce a sequence readout such as the hypothetical example at bottom (*d*) and yield a sequencing technique capable of reading a whole human genome in just 20 hours without expensive DNA copying steps and chemical reactions.



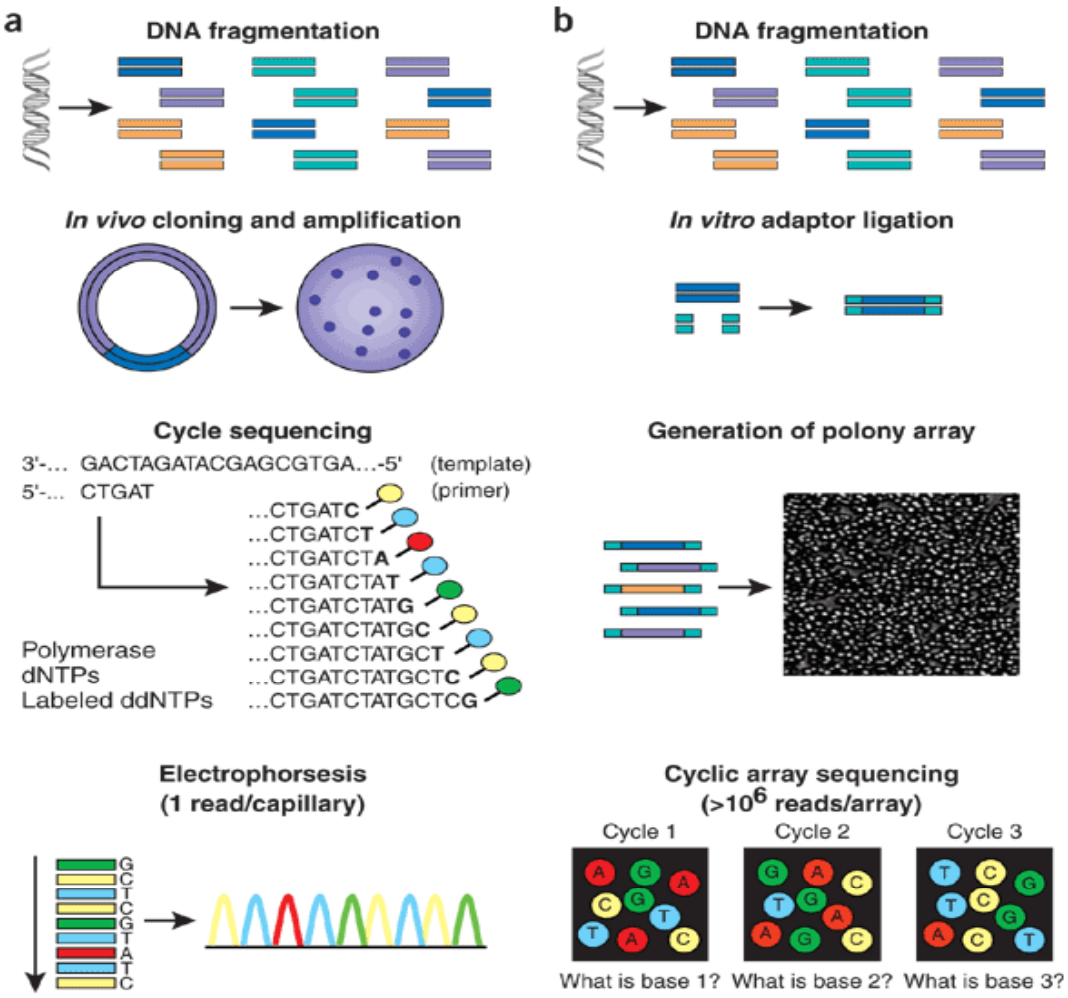




DNA Sequencing Technologies

Method	Feasibility	Read Length	Data Quality	Raw Data Production
Sanger Sequencing	Well Established	Long (800-1200 bases)	+++	+
Stepwise Synthesis	Becoming Established	Short (25-100 bases)	+	+++++
Single Molecule	Far from Established	Long (>1000 bases ?)	???	+++++ ?

Nueva generación de tecnologías de secuenciación



(**a**) With high-throughput shotgun Sanger sequencing, genomic DNA is fragmented, then cloned to a plasmid vector and used to transform *E. coli.* For each sequencing reaction, a single bacterial colony is picked and plasmid DNA isolated. Each cycle sequencing reaction takes place within a microliter-scale volume, generating a ladder of ddNTP-terminated, dye-labeled products, which are subjected to high-resolution electrophoretic separation within one of 96 or 384 capillaries in one run of a sequencing instrument. As fluorescently labeled fragments of discrete sizes pass a detector, the four-channel emission spectrum is used to generate a sequencing trace. (**b**) In shotgun sequencing with cyclic-array methods, common adaptors are ligated to fragmented genomic DNA, which is then subjected to one of several protocols that results in an array of millions of spatially immobilized PCR colonies or 'polonies'15. Each polony consists of many copies of a single shotgun library fragment. As all polonies are tethered to a planar array, a single microliter-scale reagent volume (e.g., for primer hybridization and then for enzymatic extension reactions) can be applied to manipulate all array features in parallel. Similarly, imaging-based detection of fluorescent labels incorporated with each extension can be used to acquire sequencing data on all features in parallel. Successive iterations of enzymatic interrogation and imaging are used to build up a contiguous sequencing read for each array.

***Nature Biotechnology* 26, 1135 - 1145 (2008)**

Table 1 Second-generation DNA sequencing technologies

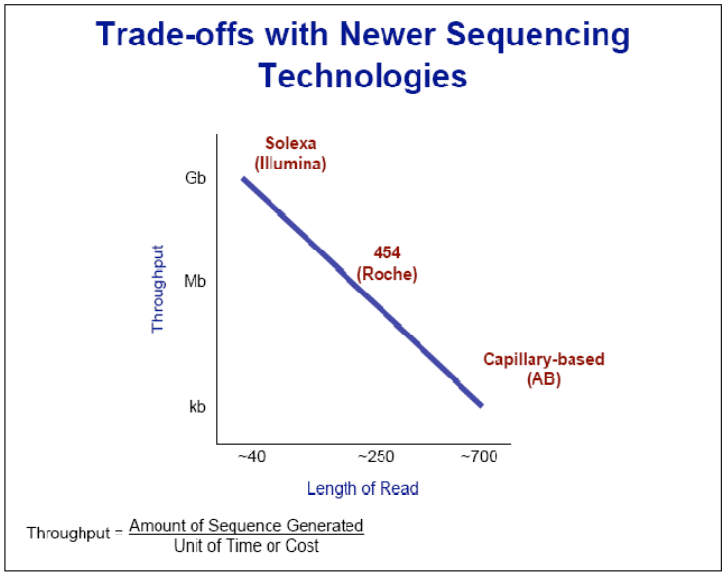
	Feature generation	Sequencing by synthesis	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length	References
454	Emulsion PCR	Polymerase (pyrosequencing)	~$60	$500,000	Yes	Indel	250 bp	14,20
Solexa	Bridge PCR	Polymerase (reversible terminators)	~$2	$430,000	Yes	Subst.	36 bp	17,22
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)	~$2	$591,000	Yes	Subst.	35 bp	13,26
Polonator	Emulsion PCR	Ligase (nonamers)	~$1	$155,000	Yes	Subst.	13 bp	13,20
HeliScope	Single molecule	Polymerase (asynchronous extensions)	~$1	$1,350,000	Yes	Del	30 bp	18,30

The pace with which the field is moving makes it likely that estimates for costs and read-lengths will be quickly outdated. Vendors including Roche Applied Science, Illumina, and Applied Biosystems have major upgrade releases currently in progress. Estimated costs-per-megabase are approximate and inclusive only of reagents. Read-lengths are for single tags. Subst., substitutions; indel, insertions or deletions; del, deletions.

Nature Biotechnology 26, 1135 - 1145 (2008)

Advances in DNA sequencing technologies

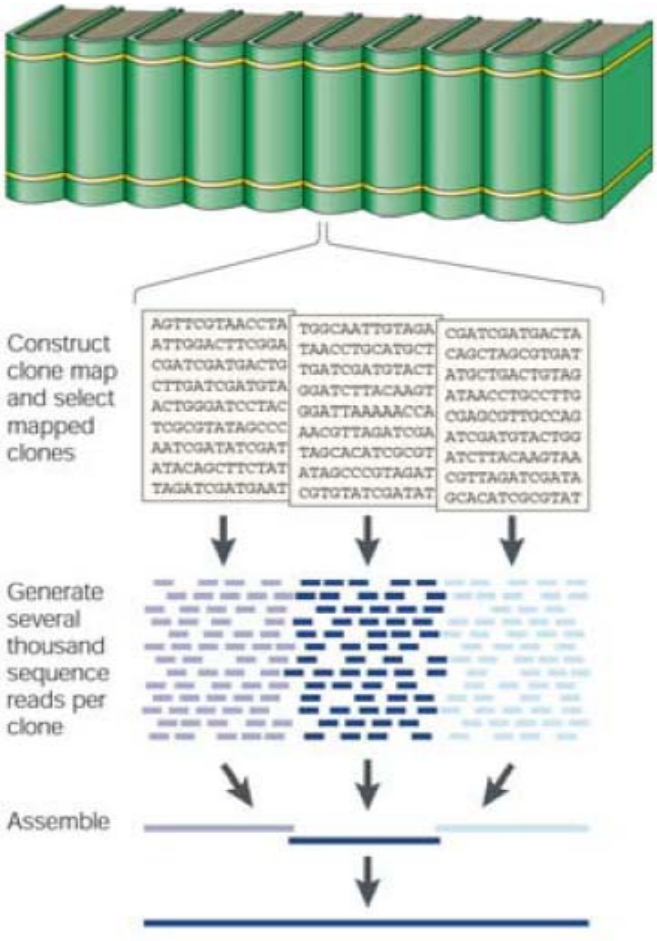
Technology	Approach	Read length	Bp per run	Company name and Web site
Automated Sanger sequencer ABI3730xl	Synthesis in the presence of dye terminators	Up to 900 bp	96 kb	Applied Biosystems www.appliedbyosystems.com
454/Roche FLX system	Pyrosequencing on solid support	200–300 bp	80–120 Mb	Roche Applied Science www.roche-applied-science.com
Illumina/ Solexa	Sequencing by synthesis with reversible terminators	30–40 bp	1 Gb	Illumina, Inc. http://www.illumina.com/
ABI/SOLiD	Massively parallel sequencing by ligation	35 bp	1–3 Gb	Applied Biosystems www.appliedbyosystems.com



R McCombie, 2008. Next Generation Sequencing Symposium.Santa Fe

Genomics 92 (2008) 255–264

Procesamiento de secuencias en proyectos genómicos



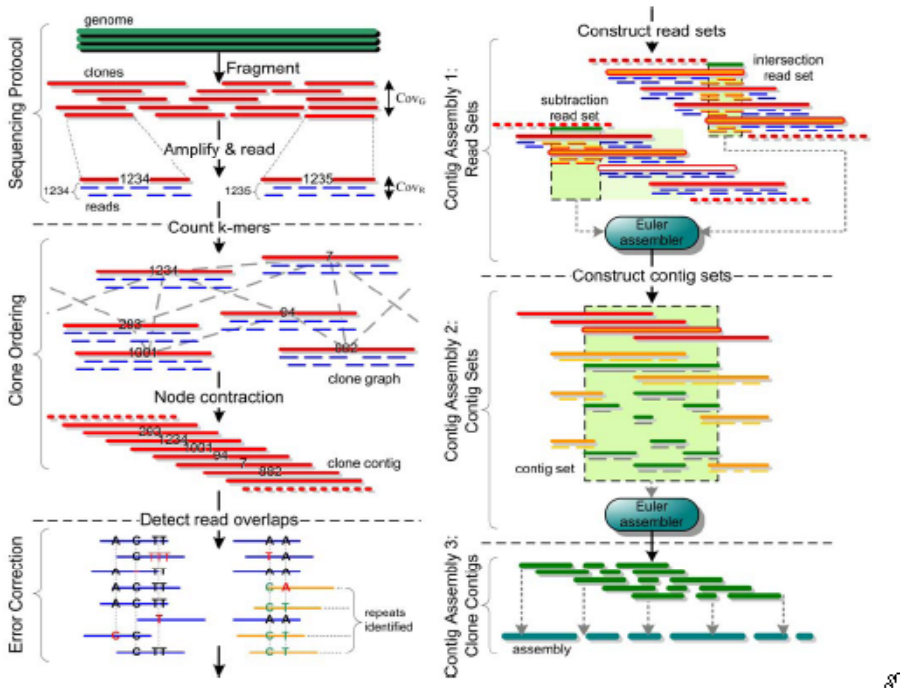
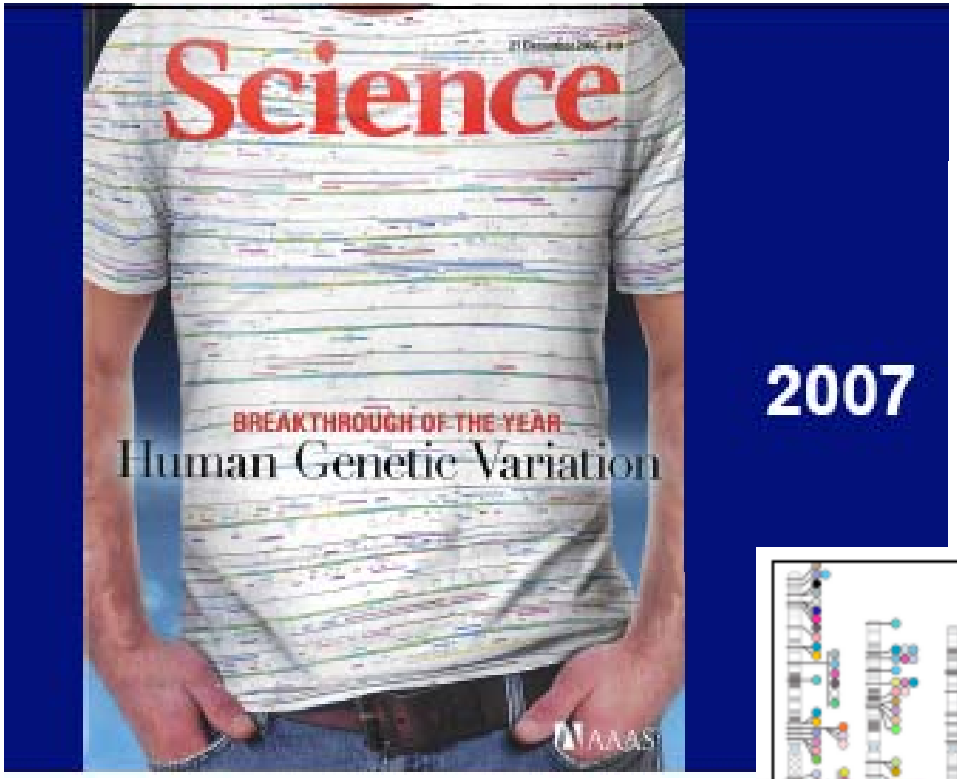


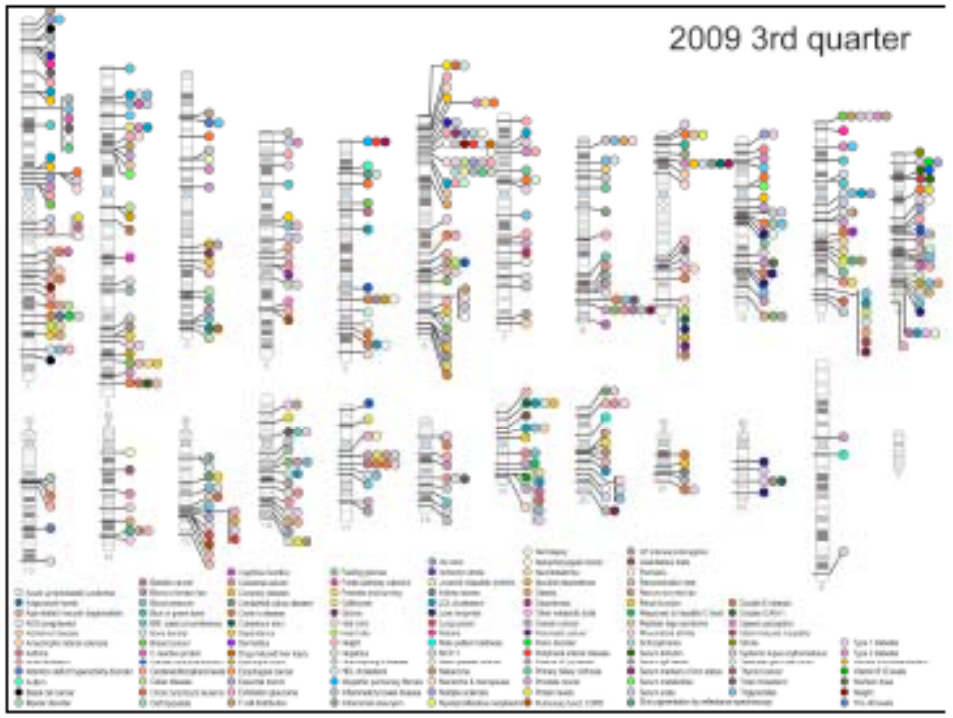
Figure 1. Sequencing protocol and assembly methodology. Reads are obtained in a hierarchical sequencing protocol with high genome-clone coverage and low clone-read coverage. From the *k*-mer content of each clone we construct a clone graph whose edge weights reflect the likely clone proximities, and from this our clone ordering algorithm determines the clone contigs. Next, we find all putative read overlaps by only looking in nearby clones and perform error correction. In three stages of contig assembly we 1) create *read sets* via set operations that consist of reads from multiple overlapping clones within small clone subregions and assemble using *Euler*, 2) combine contigs resulting from the previous stage in clone-sized *contig sets* for assembly, and 3) use a scalable assembler to merge entire clone contigs.
doi:10.1371/journal.pone.0000484.g001

PLoS ONE 2(5): e484.
doi:10.1371/journal.pone.0000484

The Genomic Landscape: *circa 2010, Eric Green*



2007



http://demo.decodeme.com/research-catalog

Human Genome Sequence

>$1,000,000,000

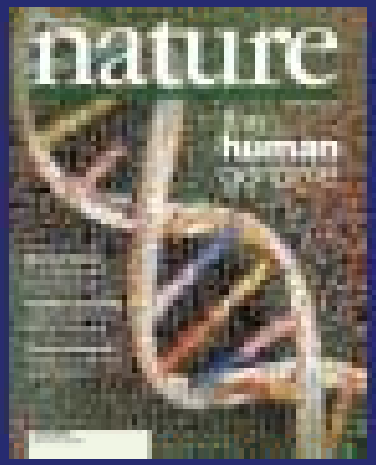
~$1,000

The Genomic Landscape: *circa 2010, Eric Green*

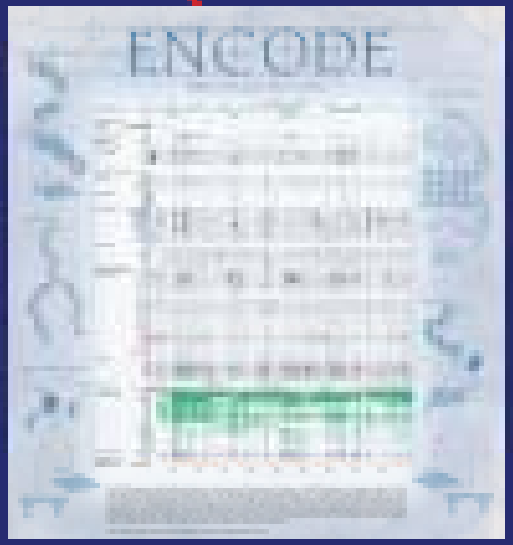
The Pathway to Genomic Medicine

Interpreting the Human Genome Sequence

Implicating Genetic Variants with Human Disease



HGP



Realization of Genomic Medicine

The Genomic Landscape: ***circa 2010, Eric Green***