

Biophysics 205,  
“Computational and Functional Genomics”

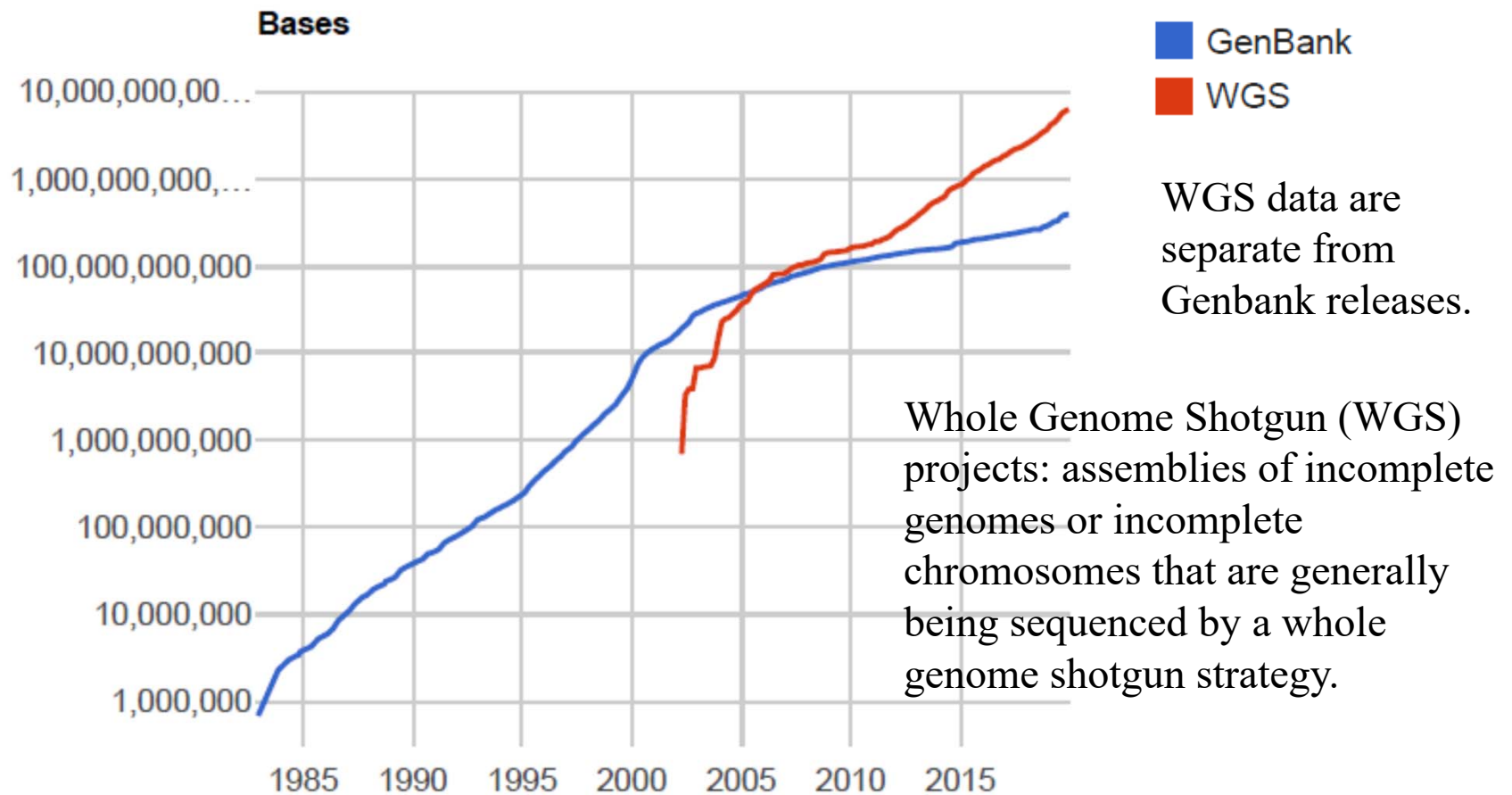
Lecture on Genome Sequencing

Prof. Martha L. Bulyk

January 27, 2020

From 1982 to early 2000s, the number of bases in GenBank doubled approximately every 18 months.

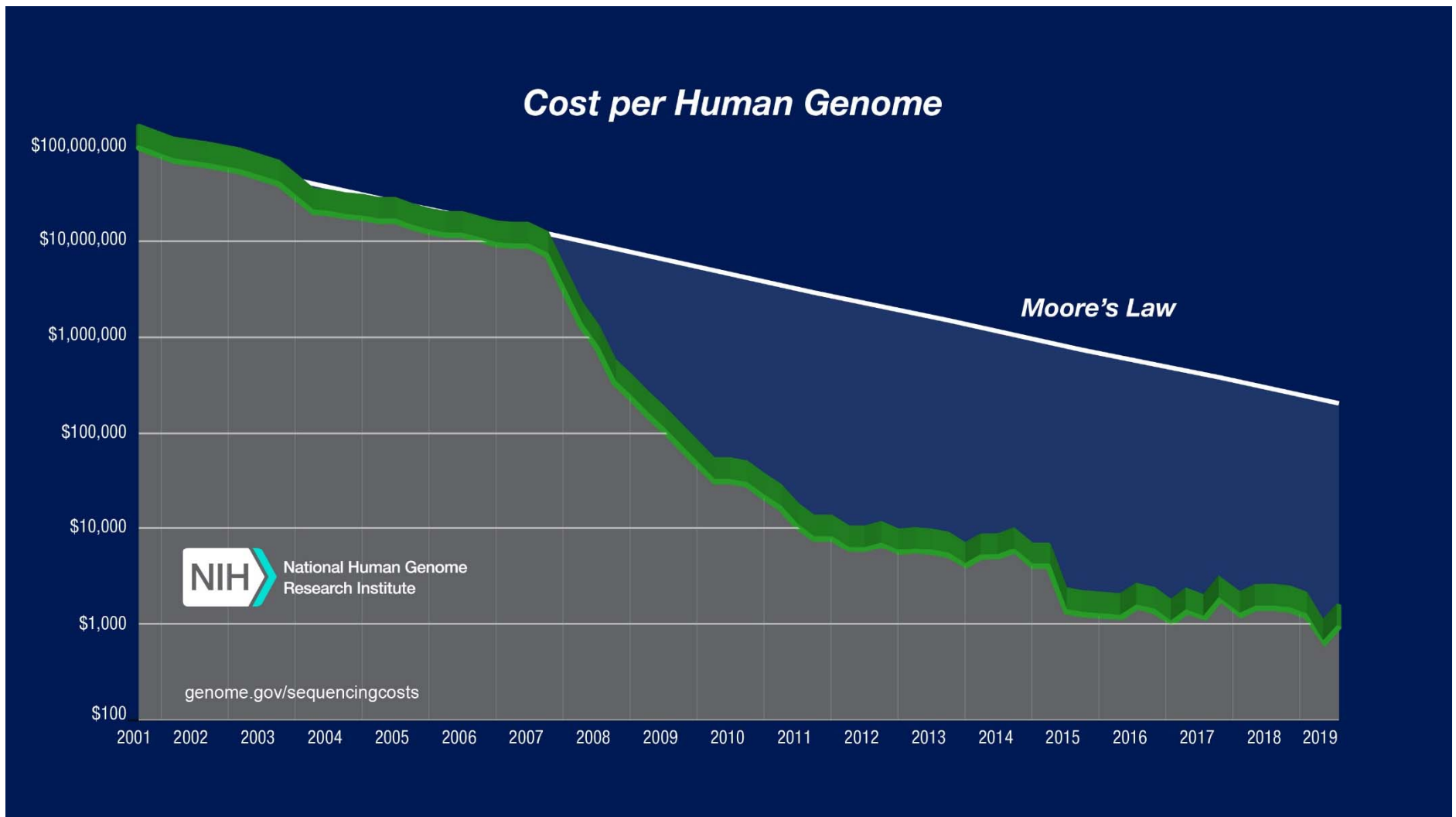
Next-gen sequencing technologies changed the landscape.



<http://www.ncbi.nlm.nih.gov/genbank/statistics>



“draft” human genome  
(Celera, Public Consortium)



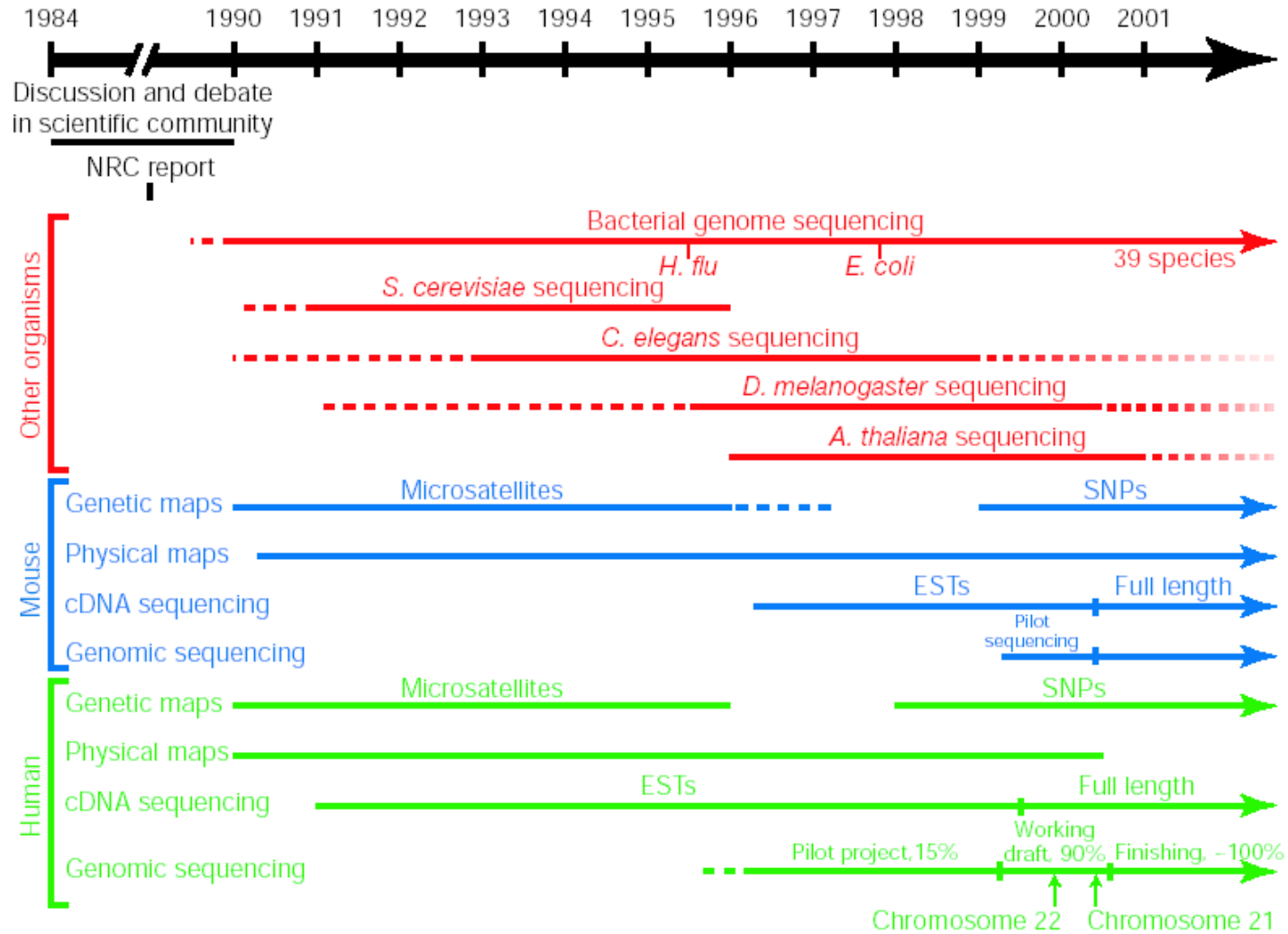
<https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>

- Cost of initial 'draft' human genome sequence production: ~**\$300 million** worldwide  
~\$2.7B if include costs of technology development, physical and genetic mapping, model organism genome mapping and sequencing, bioethics research, and program management
- Cost for advancing to 'finished' genome sequence in 2003: ~**\$150 million** worldwide

# Genome Sequencing

- brief history of Human Genome Project
- human genome sequencing
  - >500,000 human genomes have been sequenced*
- ENCODE Project
- “next-generation” sequencing (NGS) technologies
- sequencing ancient DNA (paleogenomics)

# Large-Scale Genome Sequencing Timeline



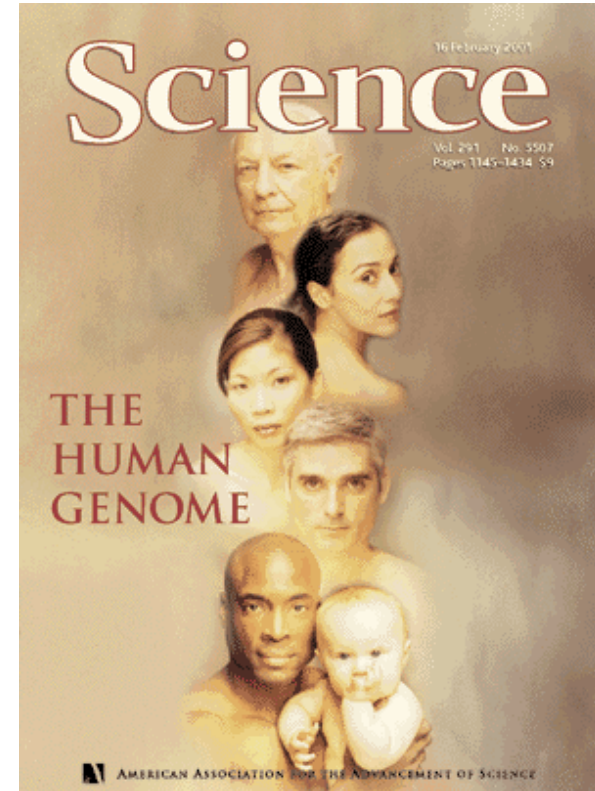
(from The Public Consortium, *Nature* '01)

## International Public Consortium



(Feb. 15, 2001; 409:860)

## Celera



(Feb. 16, 2001; 291:1304)

## *Whose sequence was it?*



(Feb. 15, 2001; 409:860)

Public Consortium:  
large number of  
anonymous donors



(Feb. 16, 2001; 291:1304)

Celera:  
5 individuals,  
including:



## 'Draft' human genome sequence in 2001:

- covered 90 percent of the genome at an error rate of one in 1,000 base pairs
- >150,000 gaps
- only 28% of the genome had reached the 'finished' standard (<1 error in 10,000 bp)

<https://www.genome.gov/human-genome-project/Completion-FAQ>



*Nature* (Oct. 21, 2004): International Public Consortium

“Finishing the euchromatic sequence of the human genome”

- 2.85 billion nucleotides (haploid genome)
- ~99% of the euchromatic portion
- 341 gaps
- error rate of 1 in 100,000 bases
- 20,352 protein-coding genes and 22,259 noncoding genes  
(Pertea *et al.*, *Genome Biol.*, 2018)

What’s left?

~ 7% of the human genome (~1% of euchromatin)

- tandem repeats, homopolymer runs
- large (>100s kb), nearly identical segmental duplications in centromeric, pericentromeric, and subtelomeric regions
  - control of replication, chrom. condensation & dynamics;
  - ribosomal DNA (rDNA) encoding ribosomal RNAs

## Some Other Human Genome Highlights:

- ~45% → different types of repetitive elements
- ~95-99% non-protein-coding
- ~40% GC
- ~40% of genes were of unknown function
- average gene size: 27 kb (average coding length: 1340 bp)
- average # exons per mRNA: ~9
- most common protein domains: 1) Ig; 2) C2H2 ZnF

## Repetitive, Transposable Elements in the Human Genome







DNA transposons: DNA intermediate; transposase enzyme catalyzes excision & re-insertion (“cut & paste”)

Retrotransposons: RNA intermediate is reverse transcribed to DNA by a transposon-encoded reverse transcriptase (“copy & paste”)

- LINEs, SINEs, LTR-containing elements

# Repetitive, Transposable Elements in the Human Genome

Classes of interspersed repeat in the human genome

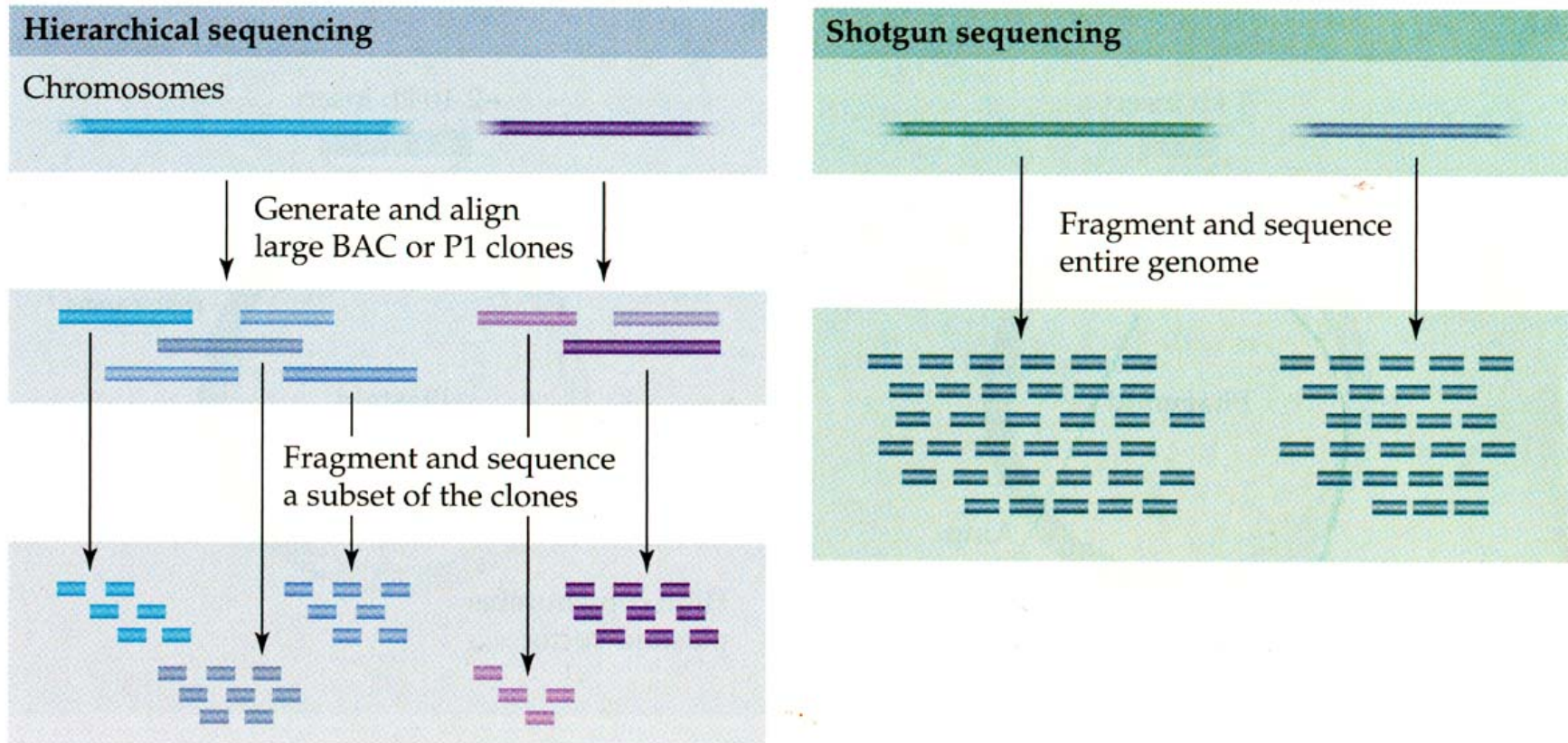
			Length	Copy number	Fraction of genome
LINEs	Autonomous		6-8 kb	850,000	21%
SINEs (ex. Alu)	Non-autonomous		100-300 bp	1,500,000	13%
Retrovirus-like elements (LTRs)	Autonomous		6-11 kb	450,000	8%
	Non-autonomous		1.5-3 kb		
DNA transposon fossils (ex. Mariner)	Autonomous		2-3 kb	300,000	3%
	Non-autonomous		80-3,000 bp		

Total: ~45%

(from The Public Consortium, *Nature* '01)

# Hierarchical (Public) versus Shotgun Sequencing (Celera)

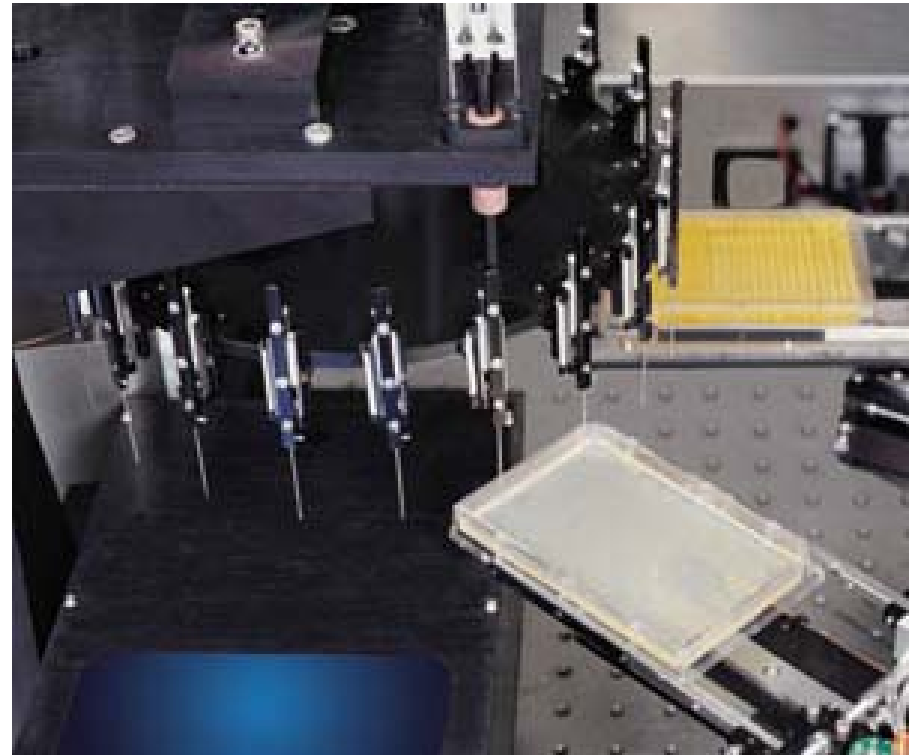
genomic DNA: partial digestion or shearing by sonication



“top-down”, “map-based”,  
or “clone-by-clone”

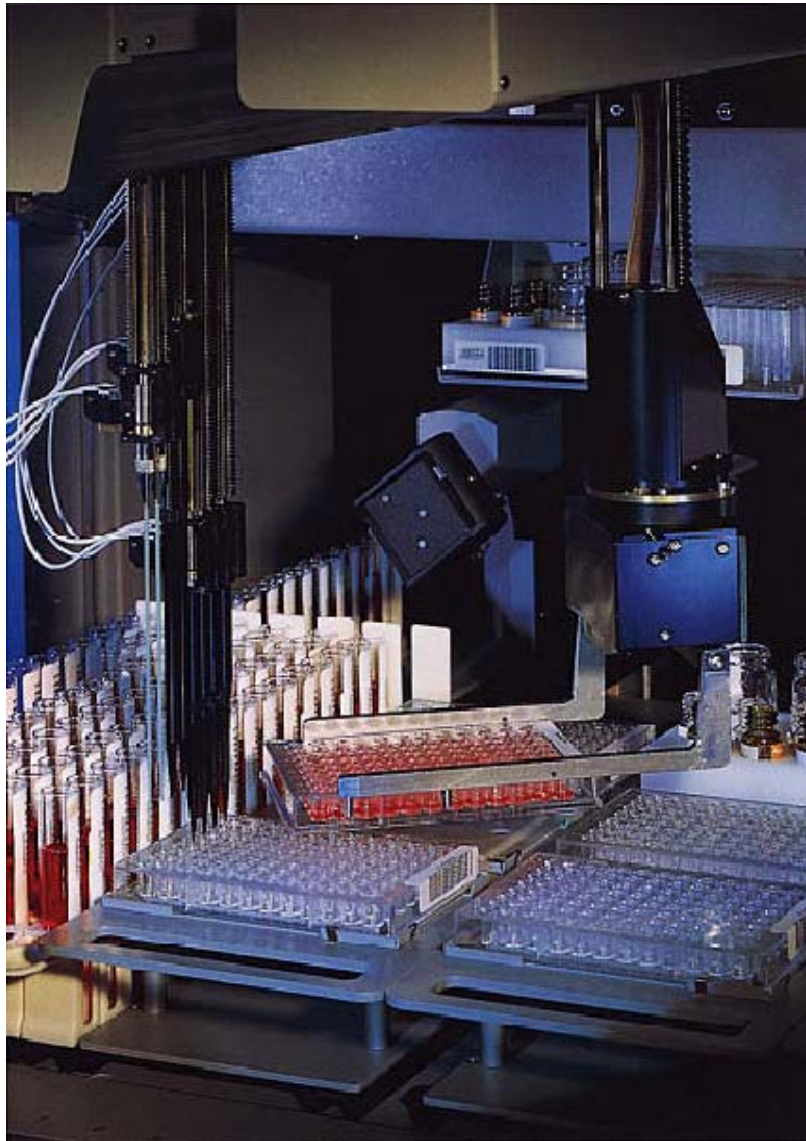
(Gibson & Muse, '02)

# Colony Picking Robot



- Gene Machines Mantis:
- capacity: 72 plates
  - can run 14 hrs unattended
  - can pick 27,000 colonies

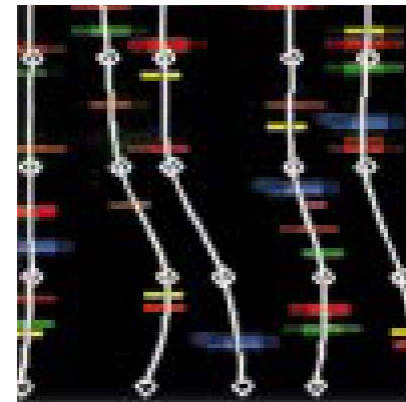
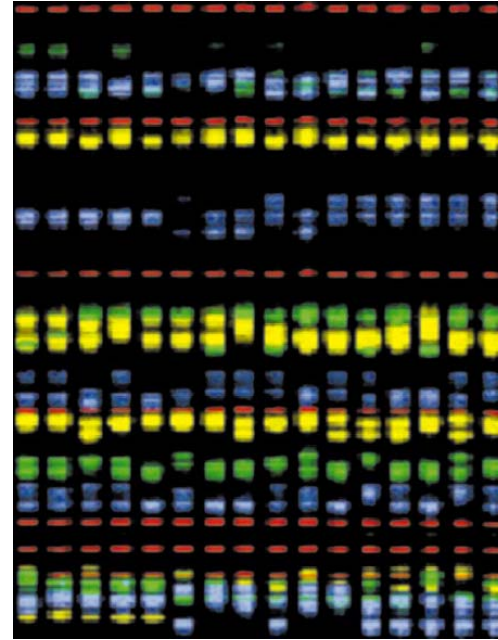
# Robot for Setting up Sequencing Reactions





# ABI377 DNA Sequencer

samples loaded into  
96-well vertical slab gel



automated  
lane-  
tracking

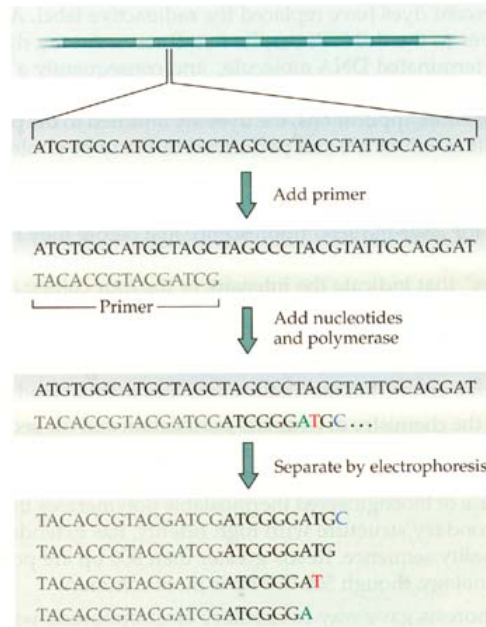


## ABI377 DNA Sequencer

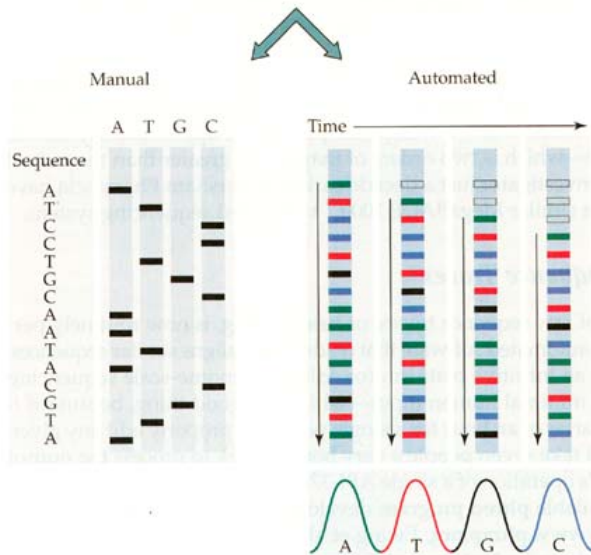
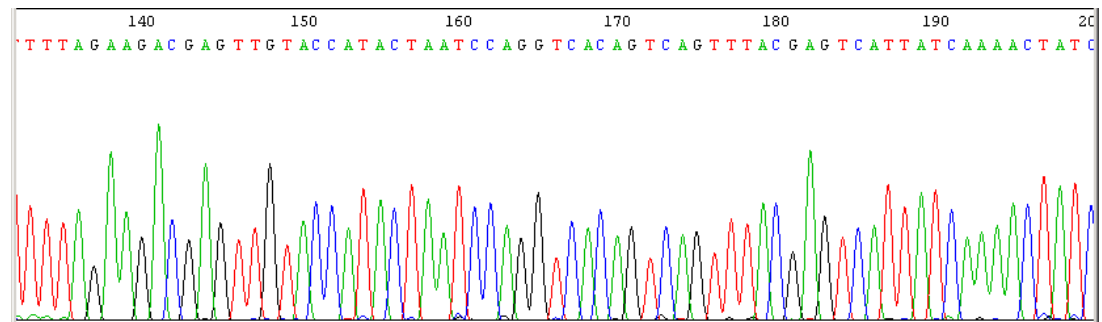
- up to 200 nt per hour
- typical read length: ~500-700 bp
- ABI 3700 capillary electrophoresis sequencer
  - no gels to pour
  - ~0.5 Mb sequence per day

→ At peak production, Int'l Consortium: 1,000 nt/sec

# Sanger Sequencing (“1<sup>st</sup> generation sequencing”)



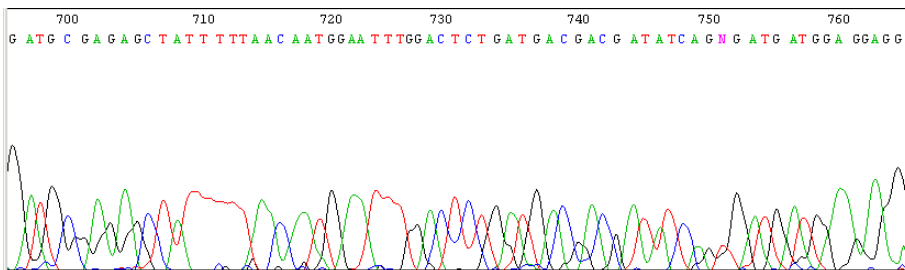
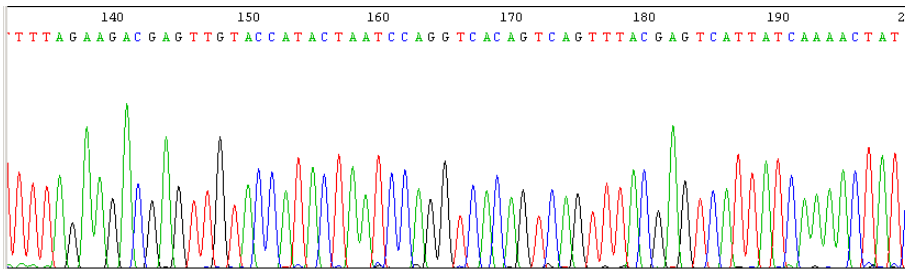
## DNA sequence trace



(Gibson & Muse, '02)

# Phred: base-calling program for DNA sequence traces

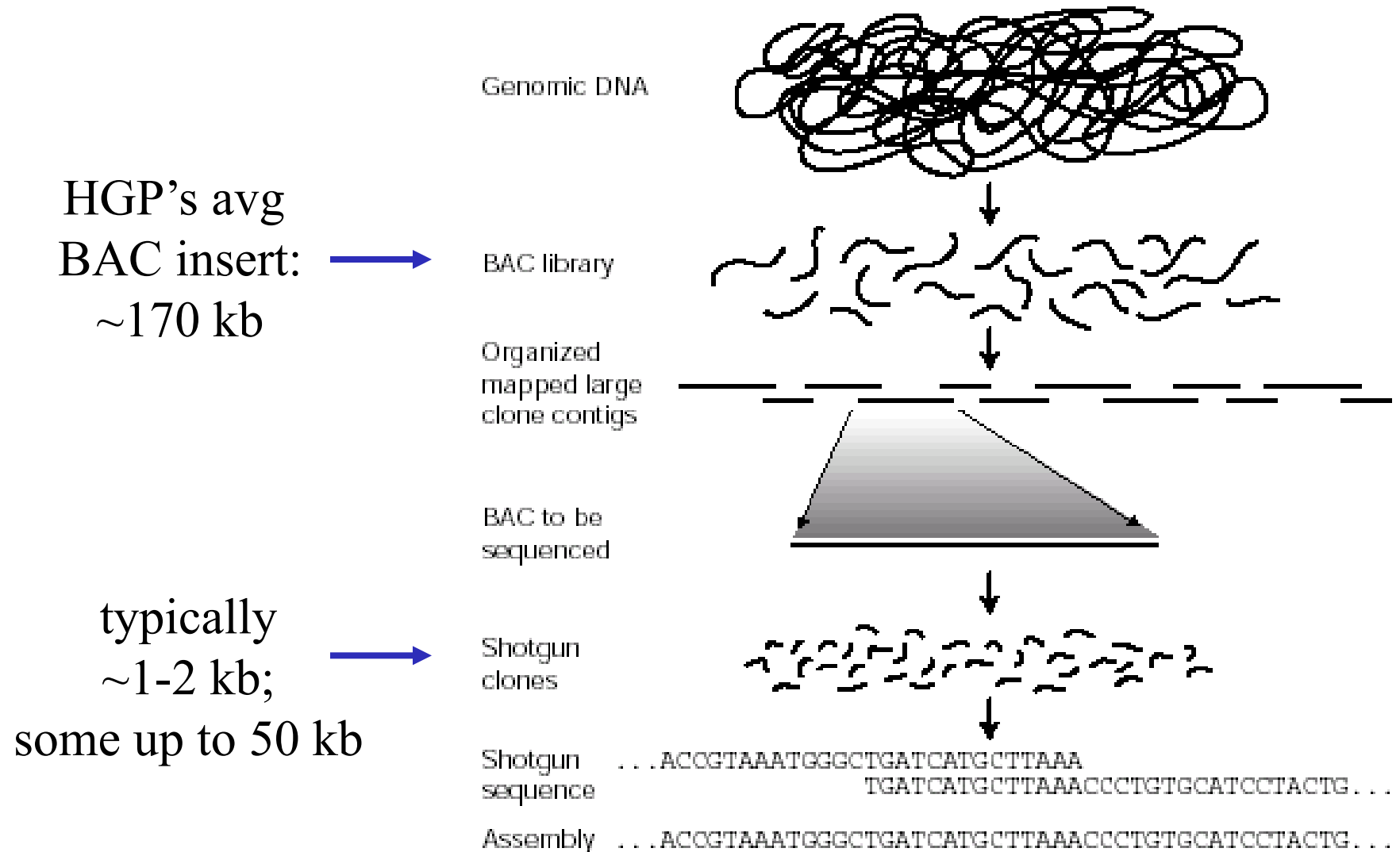
- assigns probability score to accuracy of each base call
- outputs sequence in FASTA format



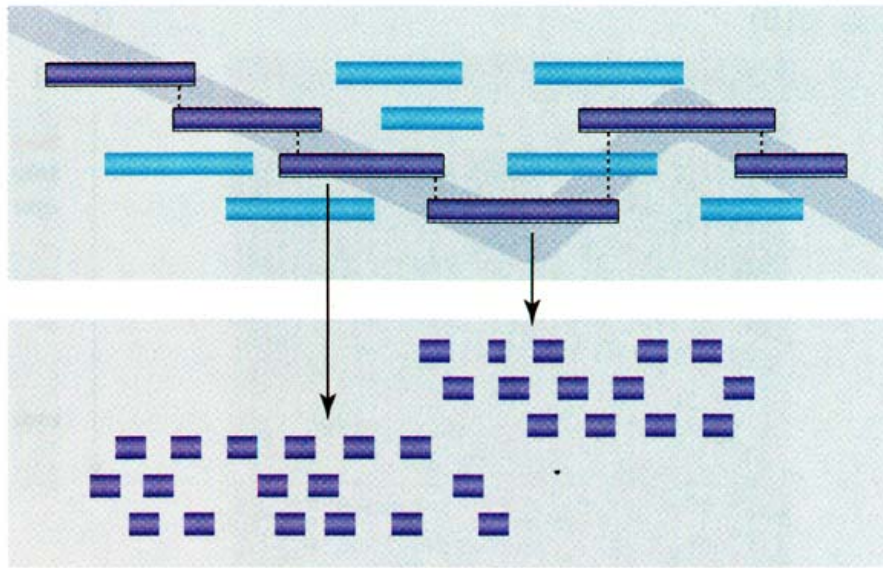
- 1) Calculate mean distance between peaks to locate predicted peaks.
- 2) Identify peaks with sufficient area.
- 3) Match observed/predicted peaks.
- 4) Find missing peaks.
- 5) Assign error probabilities to peaks.

(details: Gibson & Muse, p.69)

# Assembly for Hierarchical Sequencing



# Assembly for Hierarchical Sequencing



Align BACs

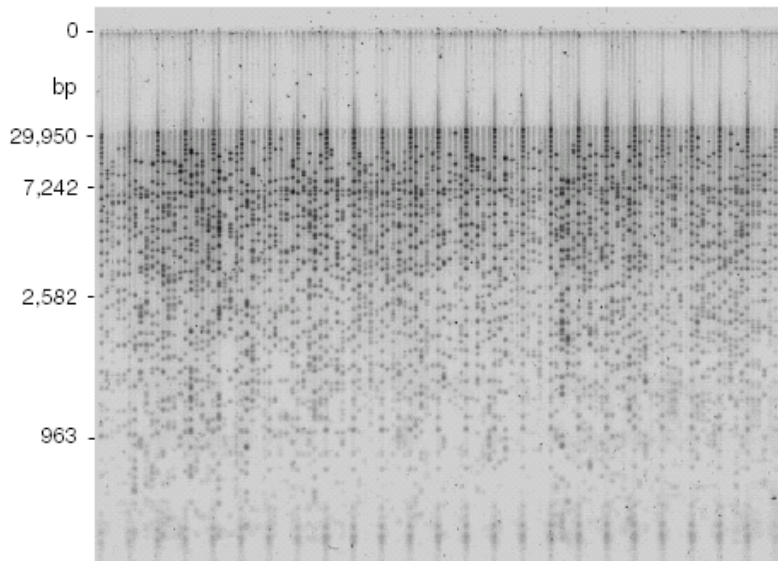
← minimal tiling path

Shotgun Sequencing

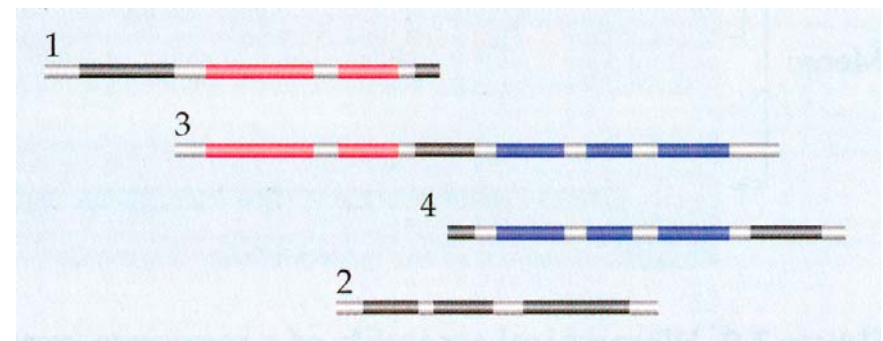
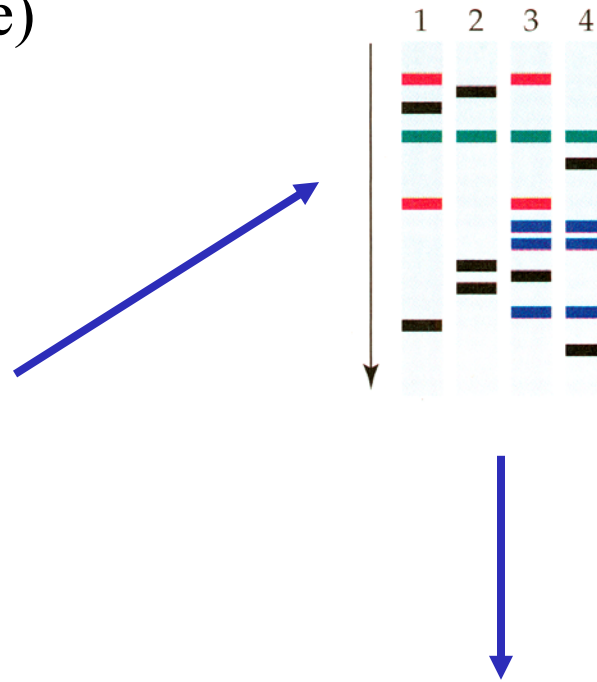
# BAC Fingerprinting

BAC: Bacterial Artificial Chromosome

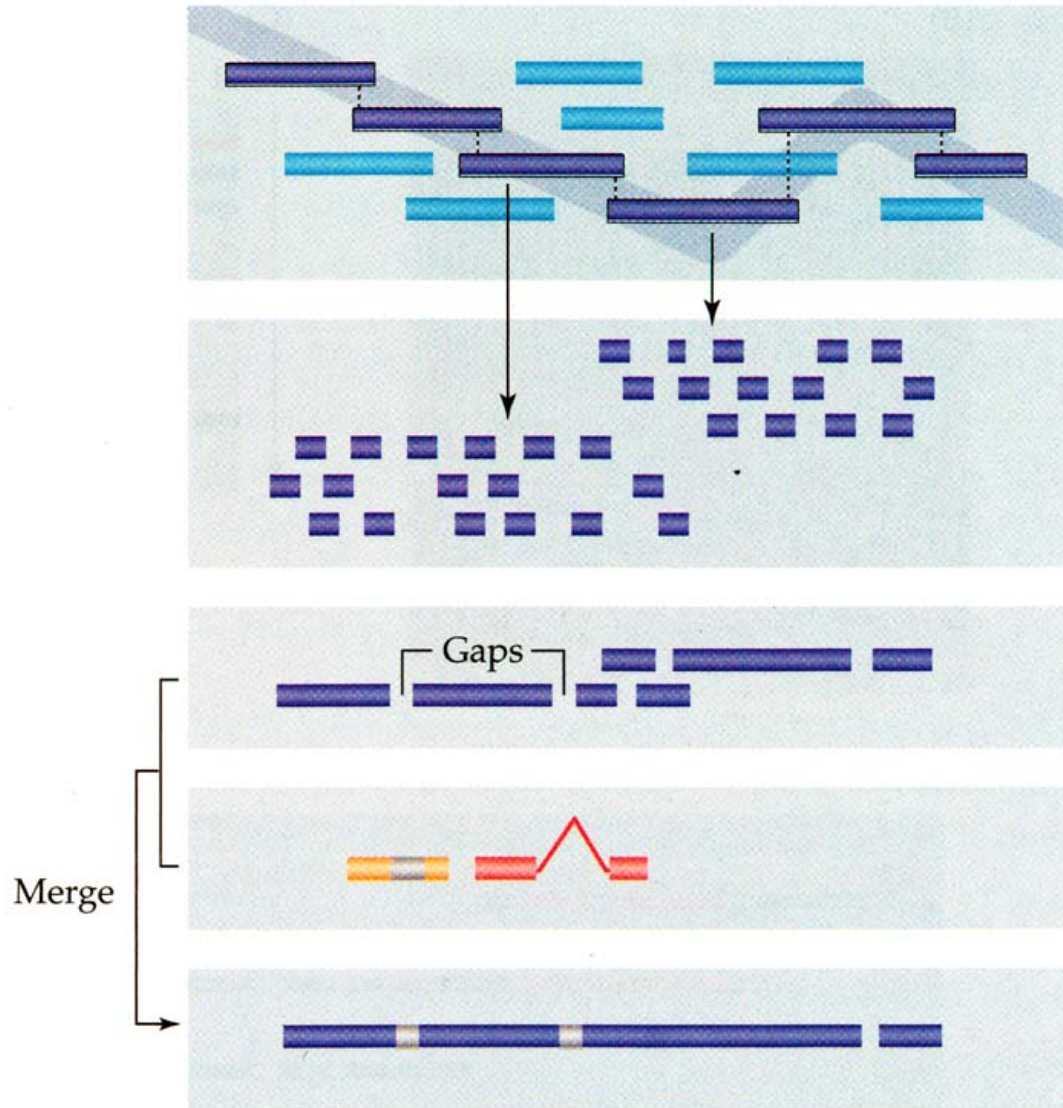
restriction digest (6-bp recog. site)



avg. BAC insert ~170 kb



# Assembly for Hierarchical Sequencing



Align BACs

Align shotgun Sequences

Contigs of  
sequenced clones

Mate-pair and cDNA data  
→ from 50 kb plasmids

Scaffold of  
sequence contigs

(Gibson & Muse, '02)



## Computational Processing of Sequenced Clones:

1. Filtering – remove bacterial, vector sequences
2. Layout – ordering the contigs
3. Merging

➔ **sequence-contig scaffolds** (draft genome sequence)

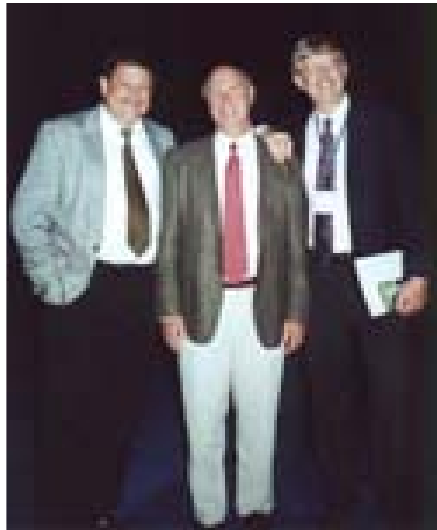




**Celera used the Public sequence in their assembly:**

“Shredding” of the public sequence with 2-fold coverage,  
to create “faux reads”

- “**coverage**”, or “**read depth**”: average number of times that a target base (*e.g.*, in the reference genome) is covered by a high-quality read
- bias in coverage depending on sequence composition



# Suggested readings

## On the sequencing of the human genome

Robert H. Waterston<sup>\*†</sup>, Eric S. Lander<sup>‡</sup>, and John E. Sulston<sup>§</sup>

<sup>\*</sup>Genome Sequencing Center, Washington University, Saint Louis, MO 63108; <sup>†</sup>Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02142; and <sup>‡</sup>Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

Communicated by Aaron Klug, Medical Research Council, Cambridge, United Kingdom, December 21, 2001 (received for review November 7, 2001)

3712–3716 | PNAS | March 19, 2002 | vol. 99 | no. 6

## Whole-genome disassembly

Phil Green<sup>\*</sup>

Howard Hughes Medical Institute and University of Washington, Seattle, WA 98195

PNAS | April 2, 2002 | vol. 99 | no. 7 | 4143–4144

## On the sequencing and assembly of the human genome

Eugene W. Myers<sup>\*</sup>, Granger G. Sutton, Hamilton O. Smith, Mark D. Adams, and J. Craig Venter

Celera Genomics, 45 W. Gude Drive, Rockville, MD 20850

PNAS | April 2, 2002 | vol. 99 | no. 7 | 4145–4146

## More on the sequencing of the human genome

Robert H. Waterston<sup>\*†</sup>, Eric S. Lander<sup>‡</sup>, and John E. Sulston<sup>§</sup>

<sup>\*</sup>Department of Genome Sciences, University of Washington, Seattle, WA 98195; <sup>†</sup>Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02142; and <sup>‡</sup>Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

3022–3024 | PNAS | March 18, 2003 | vol. 100 | no. 6

- “*de novo* sequencing”: assembly of the genome from the sequencing reads
- most genome sequencing now is just “**resequencing**”: aligning reads to a reference genome

## From a reference genome to a pan-genome:

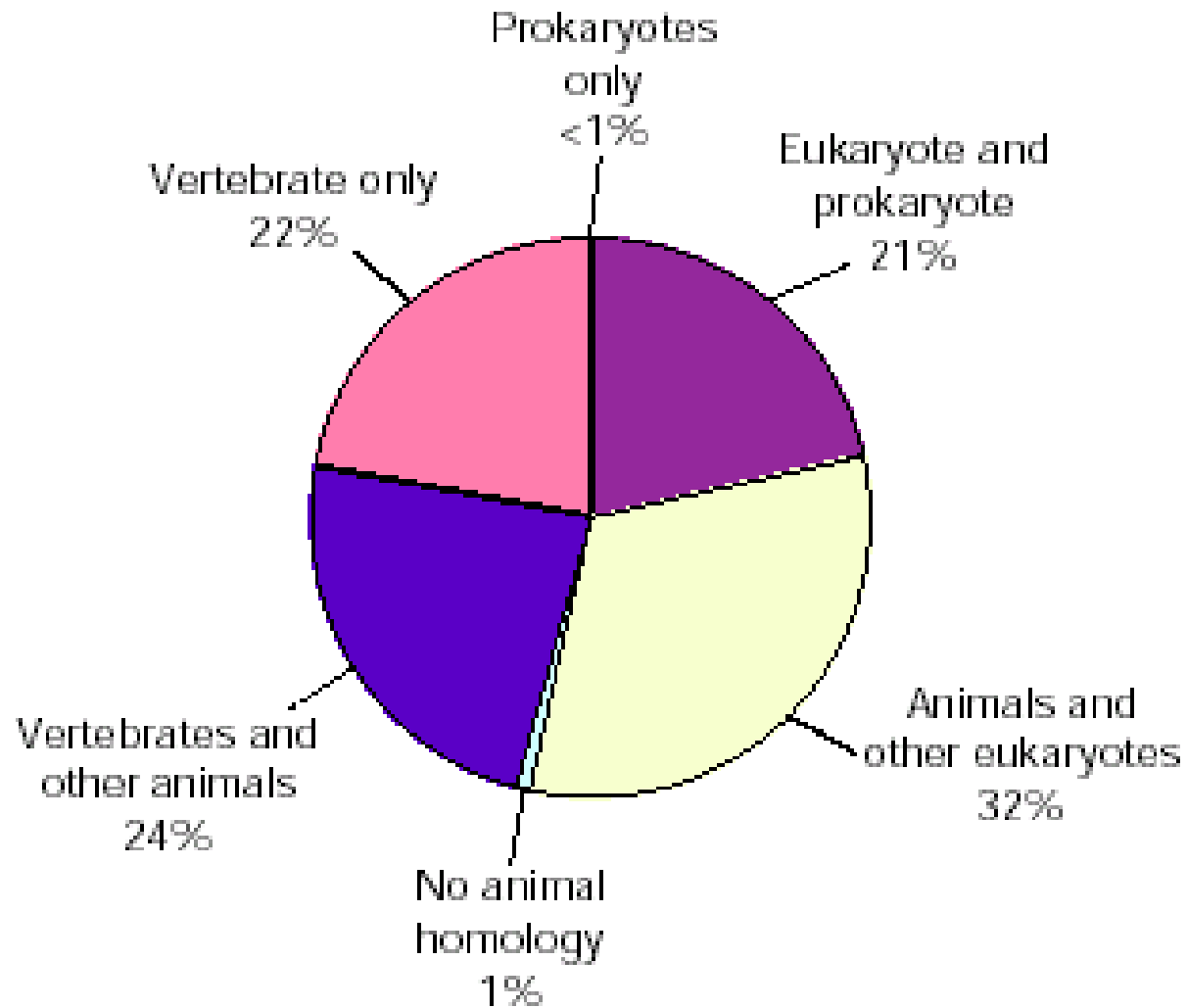
- deep sequencing of 910 individuals of African descent
- identified a set of unique sequences representing regions of the African pan-genome missing from the reference genome
- African pan-genome contains ~10% more DNA (296.5 Mb) than the current human reference genome (GRCh38)

# Genome Annotation

- protein-coding genes
- RNA genes – tRNAs, rRNAs, microRNAs, etc.
- *cis* regulatory elements
- chromosome maintenance
- repeats
- etc.

→ *How do these all function?*

# Homology of Human Genes



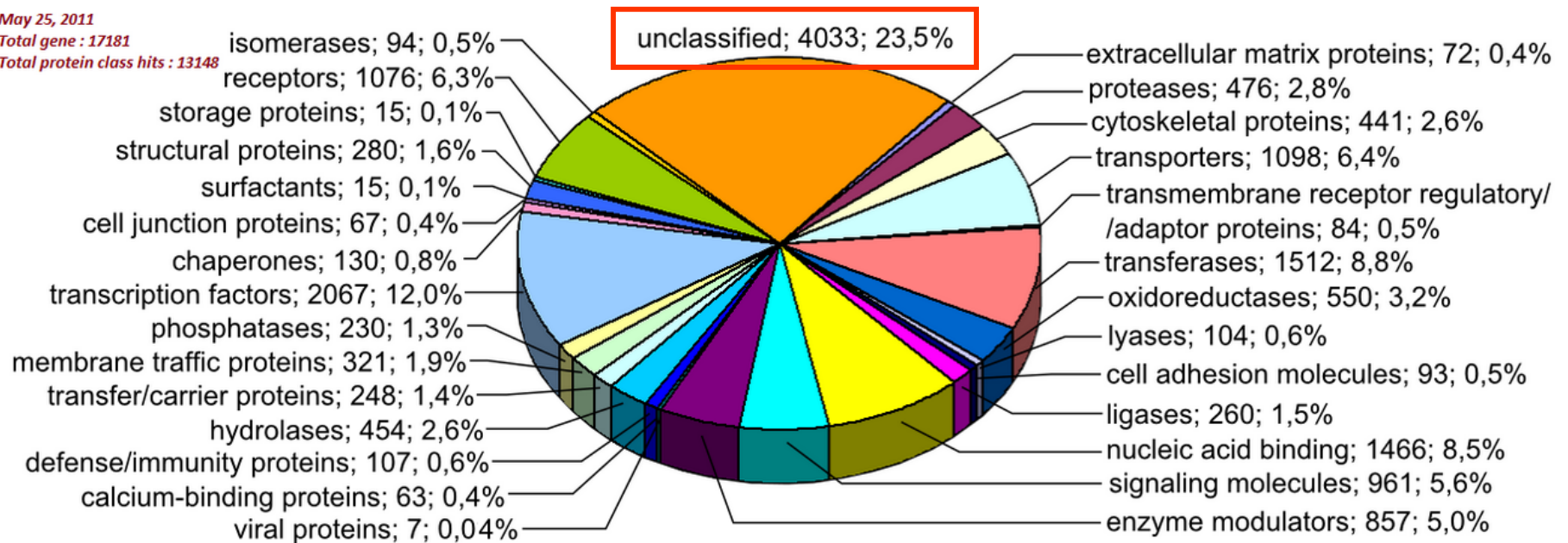
(from The Public Consortium, *Nature* '01)

# Functions of Human Genes

May 25, 2011

Total gene : 17181

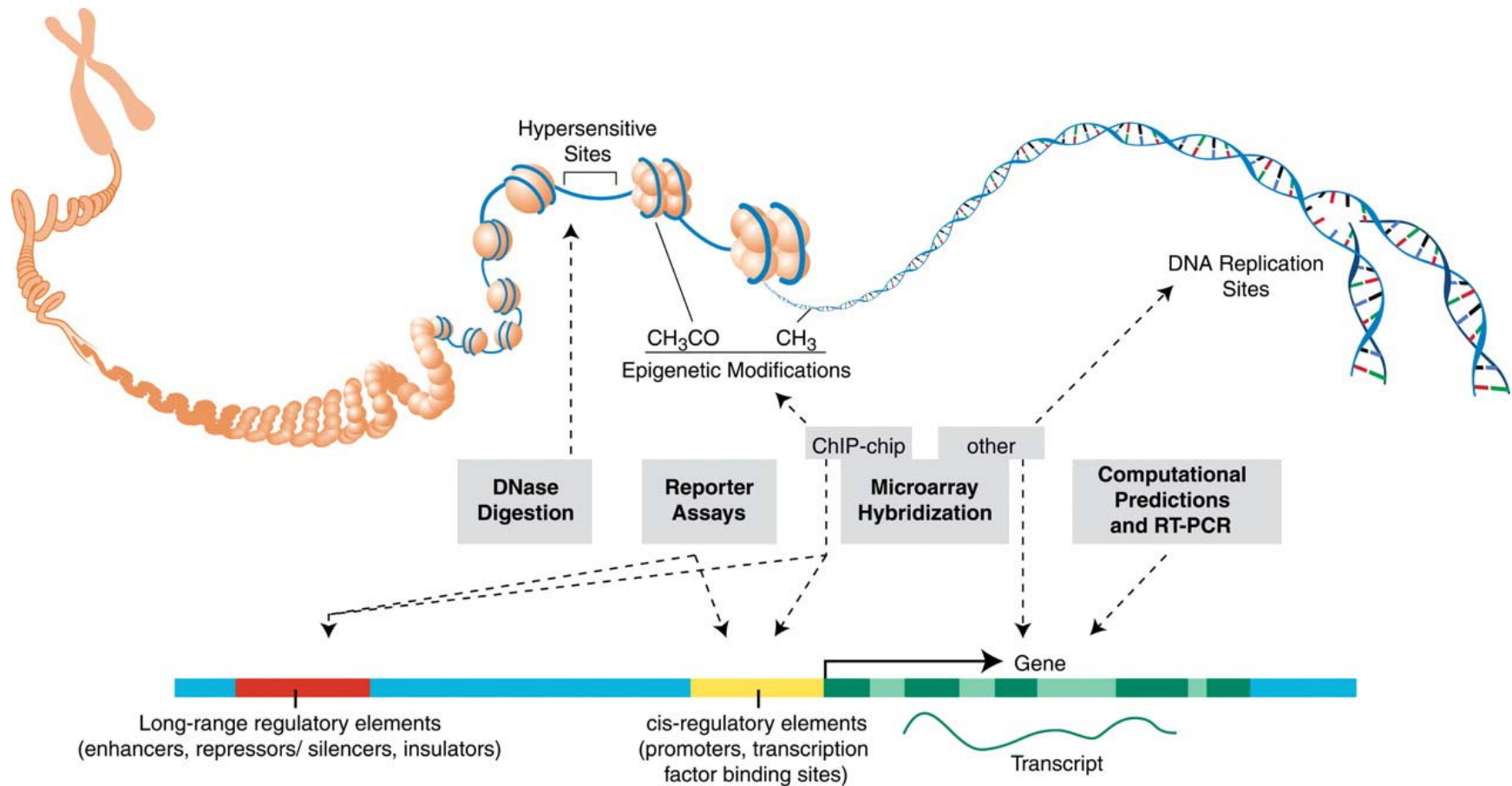
Total protein class hits : 13148



(Panther classifications from May, 2011)



# ENCODE Project: catalog the “parts list” (ENCyclopedia Of DNA Elements)



(*Science*; Oct. 22, 2004; 306:636-640)

## An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium\*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

*Nature* (2012) 489:57; plus numerous companion papers

“biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions”

## modENCODE:

Similar genome ‘parts’ cataloguing efforts, but focused on the model organisms *Drosophila melanogaster* (fruit fly) and *C. elegans* (nematode worm)

Neither ENCODE nor modENCODE cataloguing efforts were comprehensive or complete.

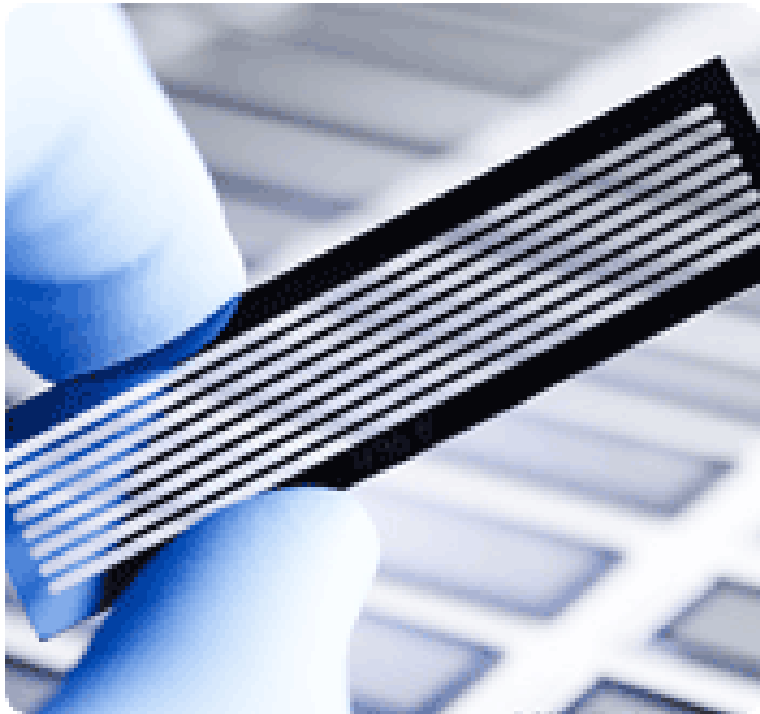
Revealed significant trends, developed technologies, generated much data, but much still remains uncharacterized and biological functions unknown.

## 2<sup>nd</sup> (“Next”) Generation Sequencing

- reaction miniaturization → cost savings
- methods of DNA sample preparation & sequence detection/readout impacts reagent costs, labor, and time required
- re-sequencing vs. *de novo* sequencing
- ..... truly personalized medicine → “precision medicine”

## Most commonly used: Illumina sequencing

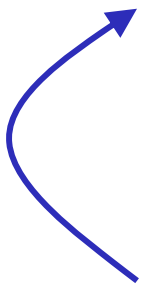
- Attach DNAs to slide surface and do on-slide template amplification to create more than a billion “clusters”, each containing ~1,000 copies of template DNA



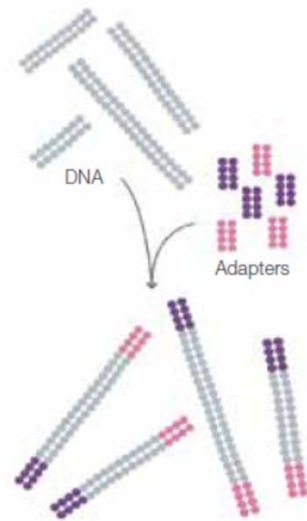
flow cell with  
8 separate “lanes” per run

# Illumina sequencing: sequencing by synthesis

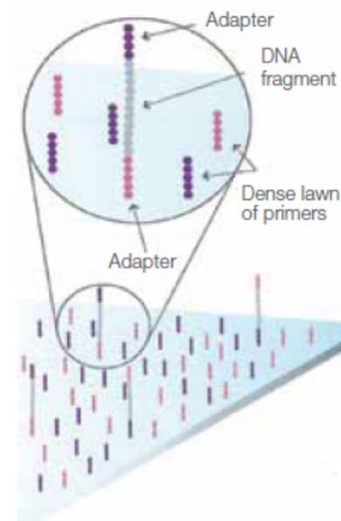
- Attach DNAs to slide surface and do on-slide template amplification to create ~more than a billion “clusters”, each containing ~1,000 copies of template DNA
- Extend with DNA polymerase + reversible fluorescent terminator nucleotides
- 4-color (or 2-color) fluorescence scans
- Remove blocked 3' terminus & fluorophore
- Add polymerase + reversible fluorescent terminators
- 4-color (or 2-color) fluorescence scans
- Repeat.
- Up to 300-nt read lengths (although quality decreases with increasing read lengths); sequencing run-times: hours to days



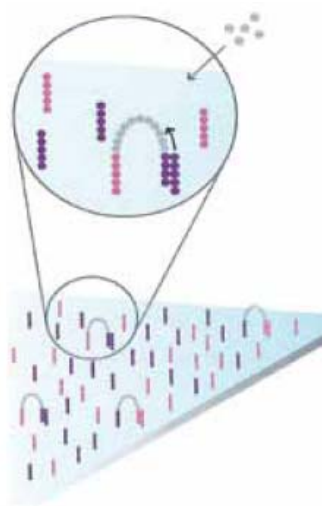
# Attaching & amplifying samples on Illumina slides



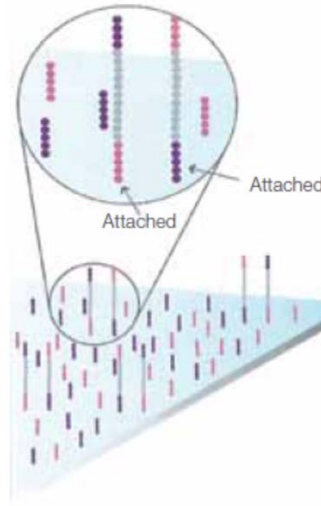
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.



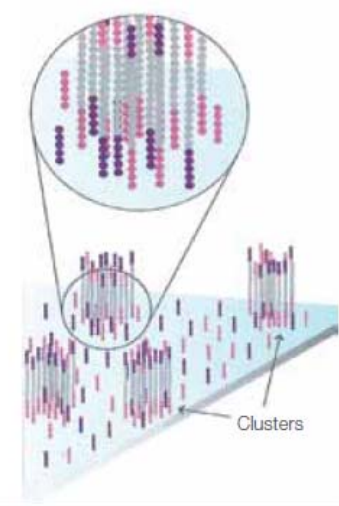
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.



Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.



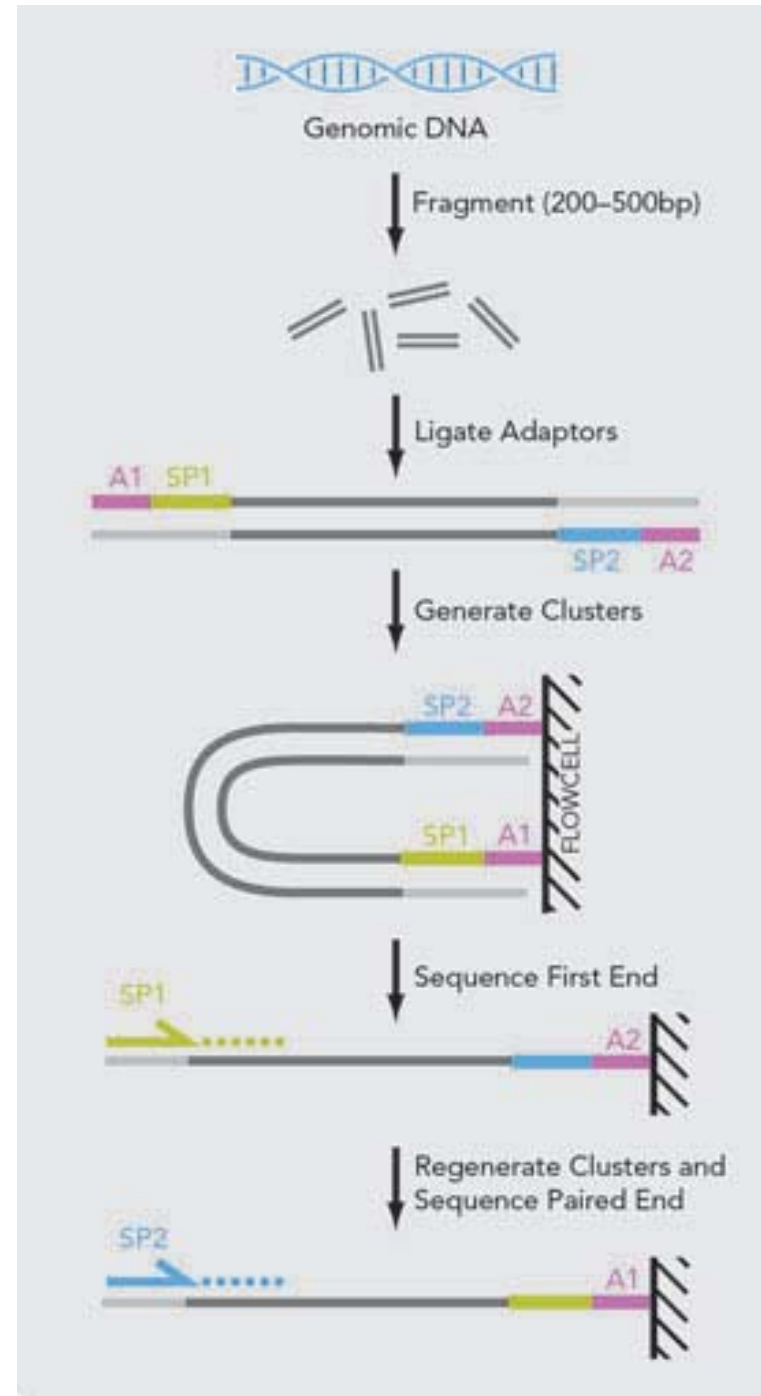
Denaturation leaves single-stranded templates anchored to the substrate.



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

“single-end” read:  
sequence one end of the  
DNAs in the fragment library

“paired-end” reads:  
sequence both ends of the  
DNAs in the fragment library





### Illumina HiSeq X Ten:

10 instruments delivering up to 3 billion paired-end reads (2 x 150 bp) per flow cell in 3 days

→ At production scale, between \$1K and \$2K per human genome.

### Illumina NovaSeq:

Up to 20 billion reads (2 x 250 bp) per flow cell in 2 days

→ At scale, cheaper & faster than HiSeq X Ten

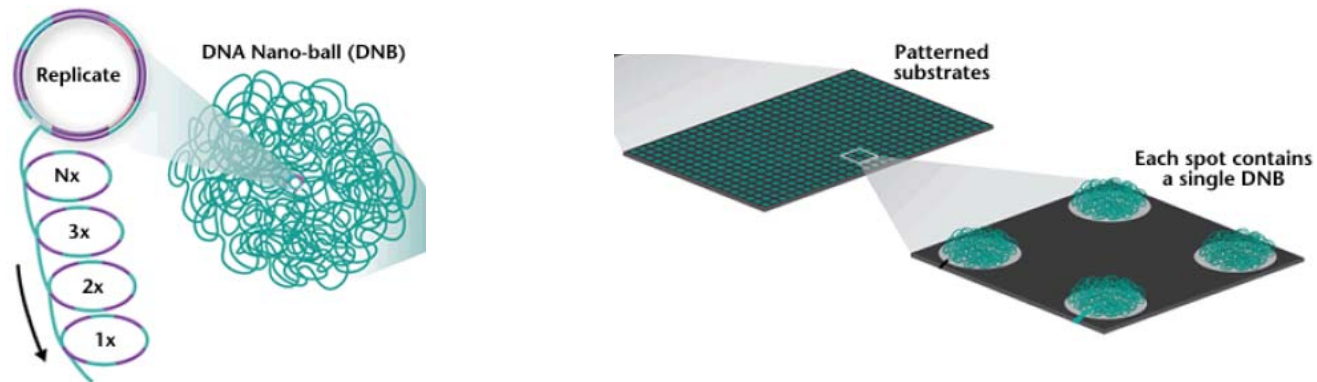
### Illumina MiSeq, NextSeq:

25M to 1B reads (2 x 300 bp MiSeq, 2 x 150 bp NextSeq) in hours to a few days

**“indexing”**: adding unique barcodes to different DNA sequencing libraries (samples), which allows for **“multiplexing”** different libraries in a single lane

# Sequencing by ligation

- Higher accuracy than sequencing by synthesis  
ex. DNA “nanoball” arrays (Complete Genomics, Inc.)
  - rolling circle amplification to create DNA “nanoballs”
  - array of DNA “nanoballs”



- sequencing by ligation of 9-mers with fluorescent dA/C/G/TTP at position  $l..n$ ; run  $n$  cycles of ligation, detection, & stripping
- 35-nt sequence from each end

# Sanger vs. 2<sup>nd</sup> Generation Sequencing Accuracy

Sanger sequencing (ABI platform): 0.001% error rate  
gold standard in sequencing

Illumina MiSeq, HiSeq: 0.1% error rate

Complete Genomics: 0.01% error rate  
*(acquired by BGI)*

(Rhoads and Au, *Genom. Proteom. Bioinf.*, 2015)

## Whole-genome sequencing:

- *Reads contain sequence errors!* Recommended to sequence to  $\geq 30x$  average depth with paired end reads, providing  $\sim 95\%$  coverage.
- For human, that requires 360 million reads & corresponds to  $>100$  Gb data (!).

## Whole-genome sequencing:

- *Reads contain sequence errors!* Recommended to sequence to  $\geq 30x$  average depth with paired end reads, providing  $\sim 95\%$  coverage.
- For human, that requires 360 million reads & corresponds to  $>100$  Gb data (!).

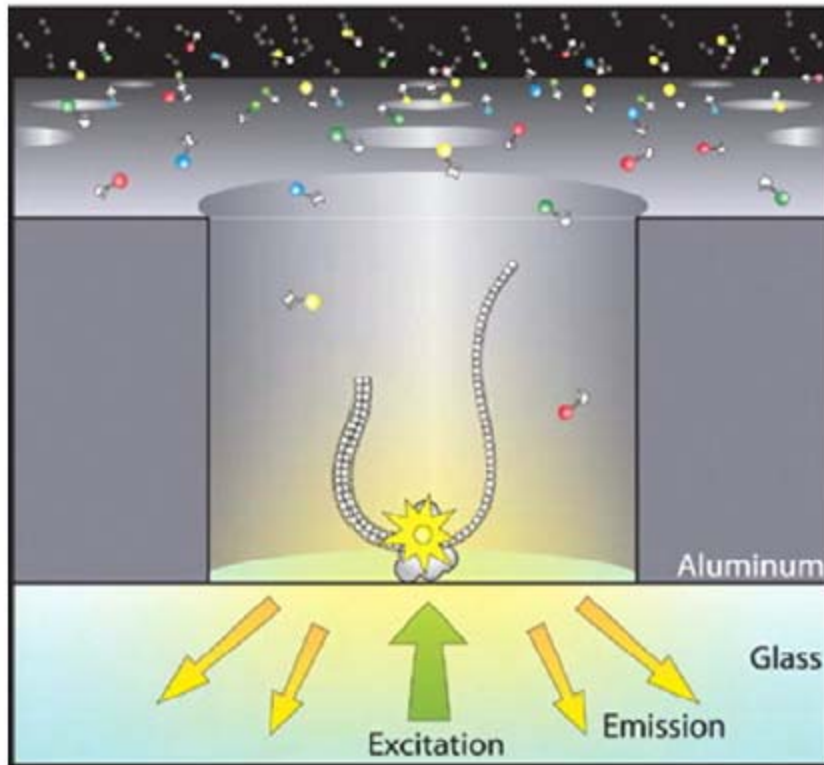
*Different sequencing technologies exhibit different sequence biases. For example, a typical Illumina genome doesn't actually cover all of the reference genome.*

## 3<sup>rd</sup> generation sequencing

- sequencing of individual DNA molecules
- very long read lengths ( $10^3$  nt to over 1 Mb)
- high error rate:  
~12-13% (or higher!) single-pass error rate, mostly indels
- sequencing accuracy improves with increasing number of times a DNA template is sequenced, enabled by use of a circular template (PacBio)

## 3<sup>rd</sup> generation sequencing

most widely used: Pacific Biosciences (“PacBio”)



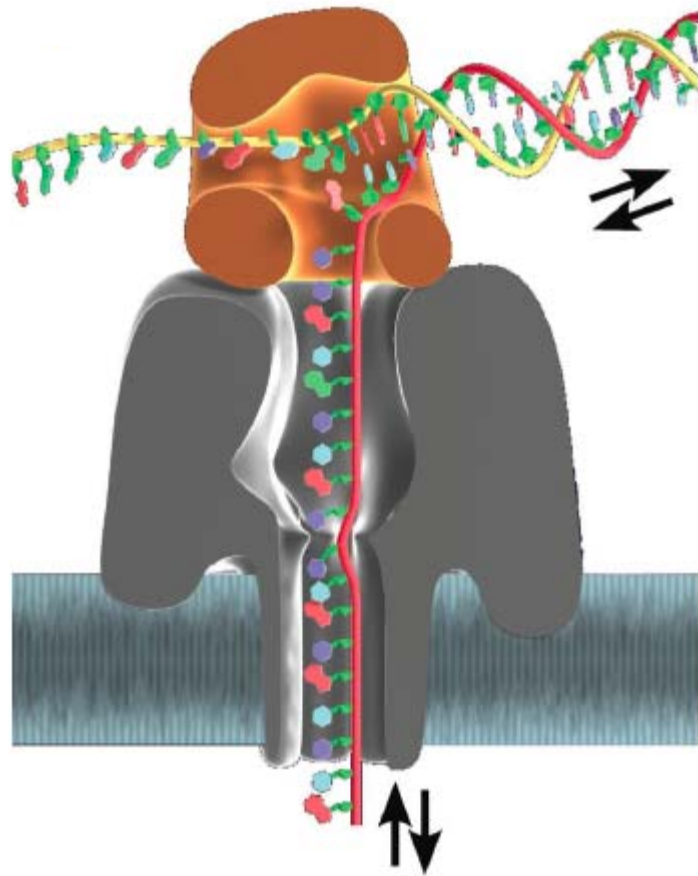
(Figure from Rhoads and Au,  
*Genom. Proteom. Bioinf.*, 2015)

Principle: Polymerase is immobilized at bottom of tiny well; as each differently fluorophore-conjugated nucleotide is incorporated by the polymerase, a camera records the emitted light and the fluorophore is cleaved prior to incorporation of next nucleotide.



# 3<sup>rd</sup> generation sequencing: Nanopore sequencing

*e.g.*, Oxford Nanopore Technologies



Principle: As DNA molecule passes through nanoscale pore, ion current passing through pore changes according to DNA sequence.

(Figure from Schneider & Dekker, *Nat. Biotechnol.*, 2012)

# Advantages of long-read sequencing over short-read sequencing

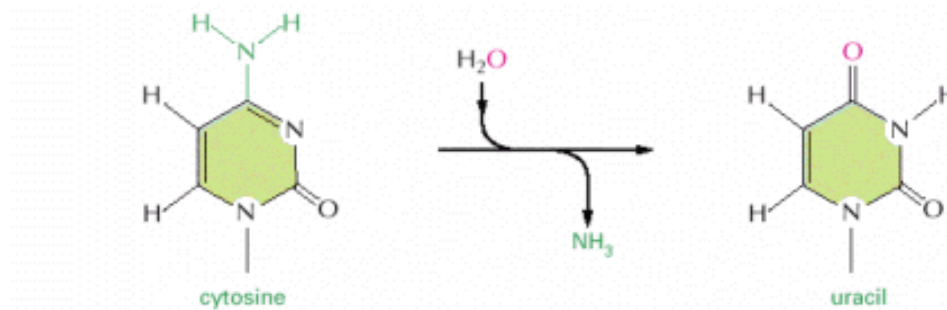
- Structural genomic variation
- Align repeat-containing reads to genome
- Span gaps in genome sequence
- Phase genomes  
Haplotype: sequence of genetic variants that co-occurs along a single chromosome
- Structure of splice isoforms

# Genome Sequencing

- brief history of Human Genome Project
- human genome sequencing
- ENCODE Project
- “next-generation” sequencing (NGS) technologies
- sequencing ancient DNA (paleogenomics)

# Challenges in Sequencing Ancient DNA

- DNA is degraded to small fragment sizes (<100 nt)
- degradation also involves deamination (C→T most commonly, but other degradation events too), which causes C to be read as T during sequencing



- most DNA in fossils is from other organisms that colonized fossil after the animal's death
- ***modern human contamination is a particular problem for sequencing of ancient hominid DNA!!!***

# Analysis of one million base pairs of Neanderthal DNA

Richard E. Green<sup>1</sup>, Johannes Krause<sup>1</sup>, Susan E. Ptak<sup>1</sup>, Adrian W. Briggs<sup>1</sup>, Michael T. Ronan<sup>2</sup>, Jan F. Simons<sup>2</sup>, Lei Du<sup>2</sup>, Michael Egholm<sup>2</sup>, Jonathan M. Rothberg<sup>2</sup>, Maja Paunovic<sup>3,†</sup> & Svante Pääbo<sup>1</sup>

Vol 444 | 16 November 2006 | doi:10.1038/nature05336 330

→ DNA from ~7 mg of ~38,000 year old fossilized femur bone

## Sequencing and Analysis of Neanderthal Genomic DNA

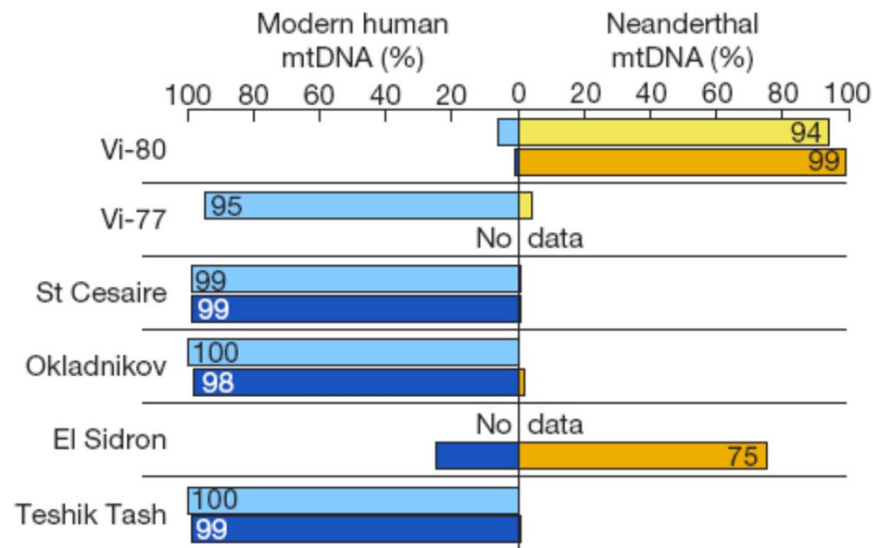
James P. Noonan,<sup>1,2</sup> Graham Coop,<sup>3</sup> Sridhar Kudaravalli,<sup>3</sup> Doug Smith,<sup>1</sup> Johannes Krause,<sup>4</sup> Joe Alessi,<sup>1</sup> Feng Chen,<sup>1</sup> Darren Platt,<sup>1</sup> Svante Pääbo,<sup>4</sup> Jonathan K. Pritchard,<sup>3</sup> Edward M. Rubin<sup>1,2\*</sup>

SCIENCE VOL 314 17 NOVEMBER 2006 1113



Vindija Cave, Croatia

PCR to amplify hypervariable region of mitochondrial DNA (mtDNA), where can distinguish Neanderthal vs. modern human DNA sequences

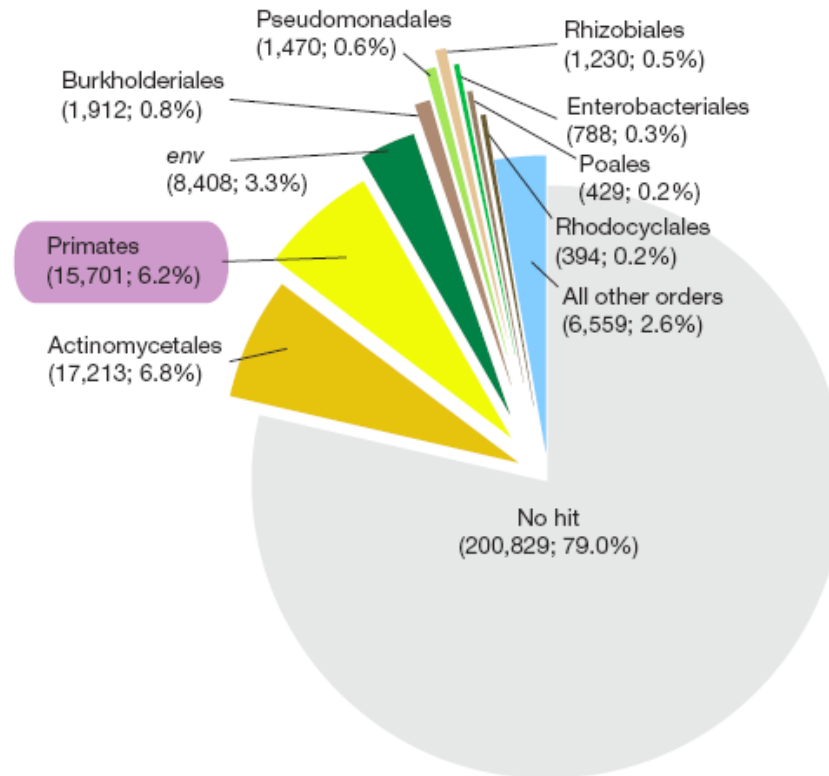


**Figure 1 | Ratio of Neanderthal to modern human mtDNA in six hominid fossils.** For each fossil, primer pairs that amplify a long (119 base pairs; upper lighter bars) and short (63 base pairs; lower darker bars) product were used to amplify segments of the mtDNA hypervariable region. The products were sequenced and determined to be either of Neanderthal (yellow) or modern human (blue) type.

(Green *et al.*, *Nature*, 2006)

More recently, enrich for human mitochondrial DNA and screen bone fragments to identify those with high content of Neanderthal DNA and low content of present day human mtDNA (Prüfer *et al.*, *Science*, 2017)

# Examine similarity of sequence reads to all known DNA sequences using BLASTN



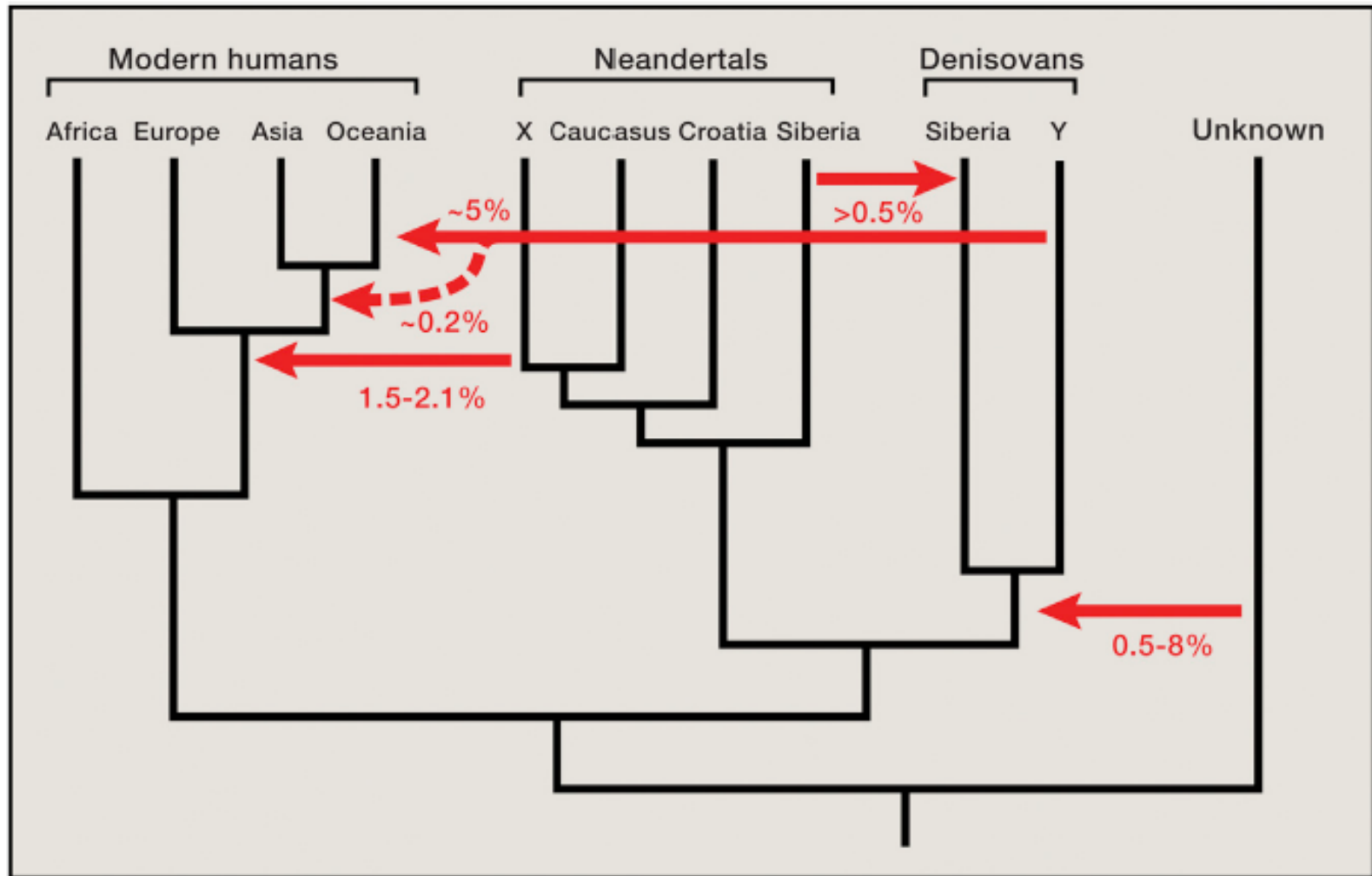
**Figure 2 | Taxonomic distribution of DNA sequences from the Vi-80 extract.** The taxonomic order of the database sequence giving the best alignment for each unique sequence read was determined. The most populous taxonomic orders are shown.

(Green *et al.*, *Nature*, 2006)

- genome sequence of Neanderthal at ~30x coverage from ~52,000 years ago from Vindija cave (Prüfer *et al.*, *Science*, 2017) or from Altai mountains in Siberia (Prüfer *et al.*, *Nature*, 2014)
- modern human / Neanderthal populations split ~550,000-765,000 years ago (assuming human/chimp divergence was 6.5-13 Mya) (Prüfer *et al.*, *Nature*, 2014)
- all or almost all of the gene flow detected was from Neanderthals into modern humans (Prüfer *et al.*, *Nature*, 2014)
- on average, 1.8-2.6% of the genomes of Eurasians are derived from Neanderthals, while essentially none of the genomes of sub-Saharan Africans derive from Neanderthal (Prüfer *et al.*, *Science*, 2017)
- Neanderthal variants associated with human phenotypes



# Gene flow between archaic and present-day humans



(Paabo, *Cell*, 2014)

## Background & optional readings for today:

- human genome sequencing papers
- ENCODE Project overview & commentaries
- next generation sequencing review article

## Reading for this Wednesday:

- 2 required papers for discussion in class

→ *All articles are posted on course website.*