GENOMICS course

Genome Sequencing Strategies

Eötvös Loránd University, Faculty of Science, Department of Genetics

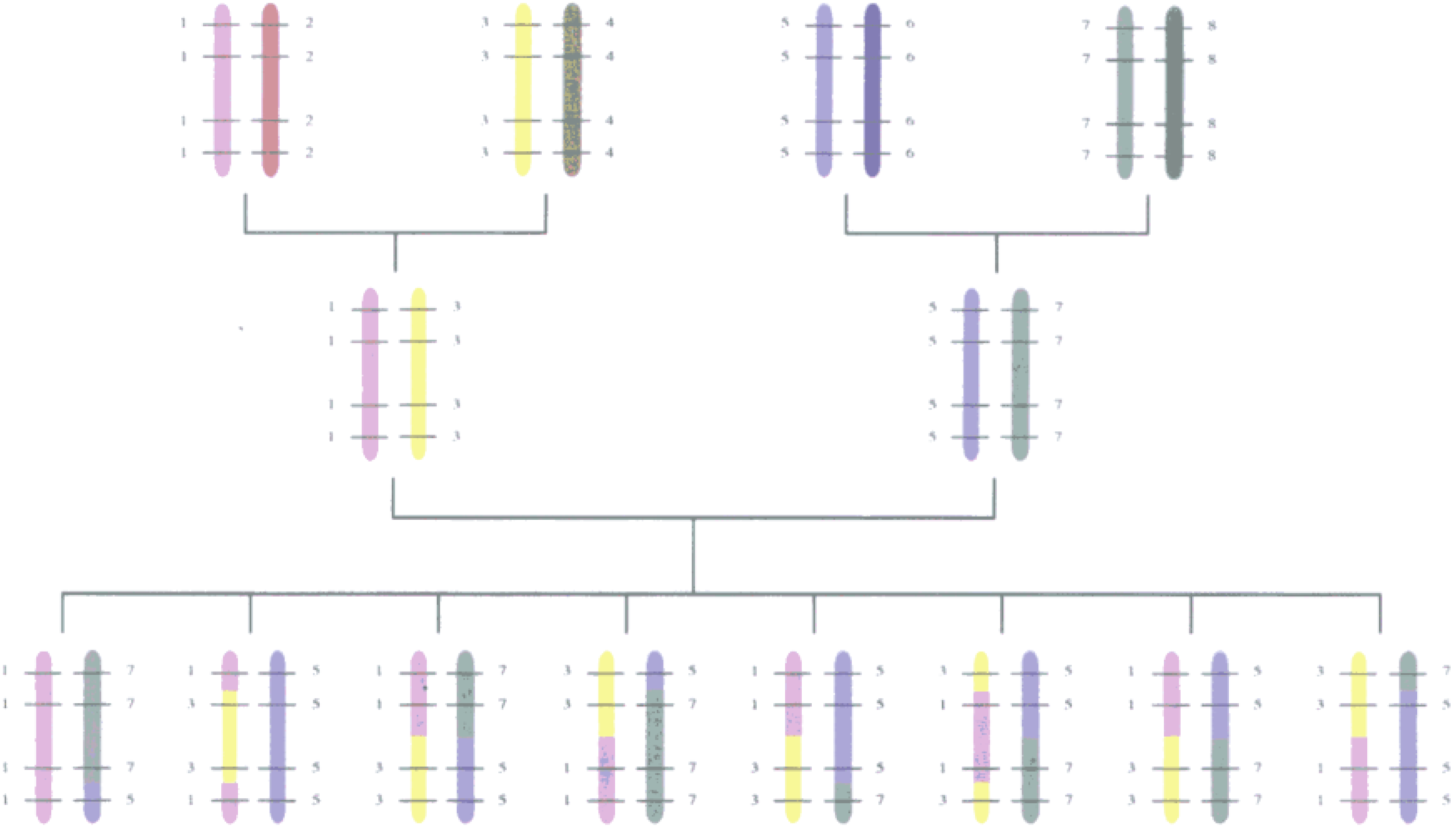
Deciphering the genetic information

- **I. phase:** cellular basis of heredity, chromosomes. (Miescher, Flemming, Mendel, Sutton, Morgan etc.)
- **II. phase:** molecular basis of heredity, DNA double helix. (Watson, Crick, Wilkins, R. Franklin, Chargaff etc.)
- **III. phase:** biological mehanism of heredity. (transcription, translation, enzymes, recombinant DNA)
- **IV. phase:** deciphering genes and genomes, **Genomics.** (genetic mapping, gene and genome sequencing, bioinformatics)
- Genome sequencing projects: **OMICS**

Human Genome Project
- backgrounds

- *First scientific initials: in the early 1980s*
- accelerate biomedical research, infrastructure investment
- *On-going genome sequencing projects*
- λ-phage, SV40 virus, human mitochondrial genome (1981)
- *Genetic and physical mapping in human genome*
- Botstein et al., 1980; Coulson et al., 1986;
- *Developement in DNA sequencing technologies*
- shotgun sequencing, ESTs, STSs etc.
- *US NRC Report 1988, US DOE and NIH.*
- parallel model organism genome projects; genetic, physical and sequence maps of human genome; bioetical issues.

Meiotic Breaks – Genetic Linkage Maps



Universal Landmark

Sequence Tagged Site (STS) 1989

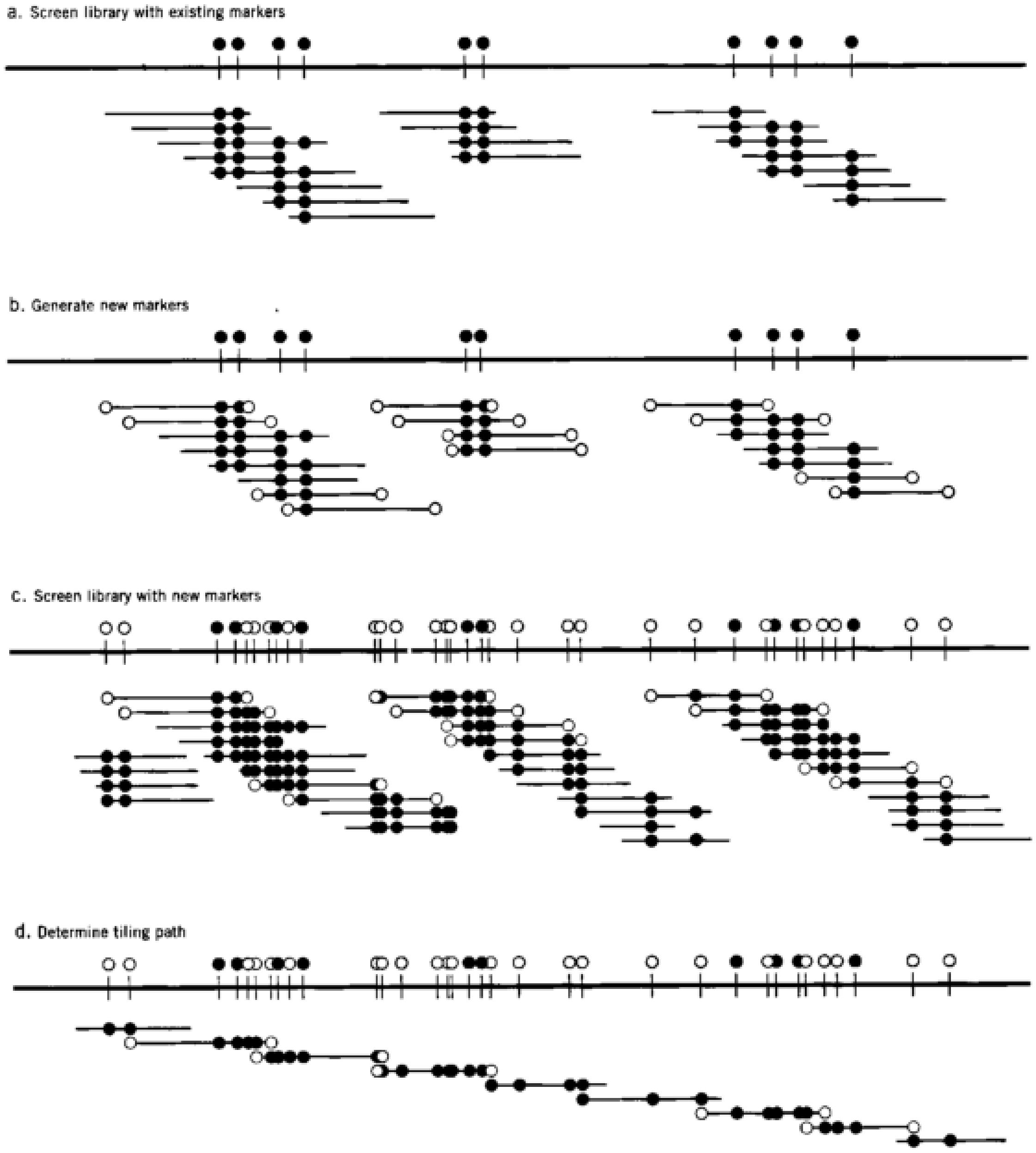
Replaces cloned DNA probe mapping landmarks with PCR assays.

Each STS is uniquely described by a pair of oligonucleotides, a product size, and PCR reaction conditions. Can be stored and distributed electronically.

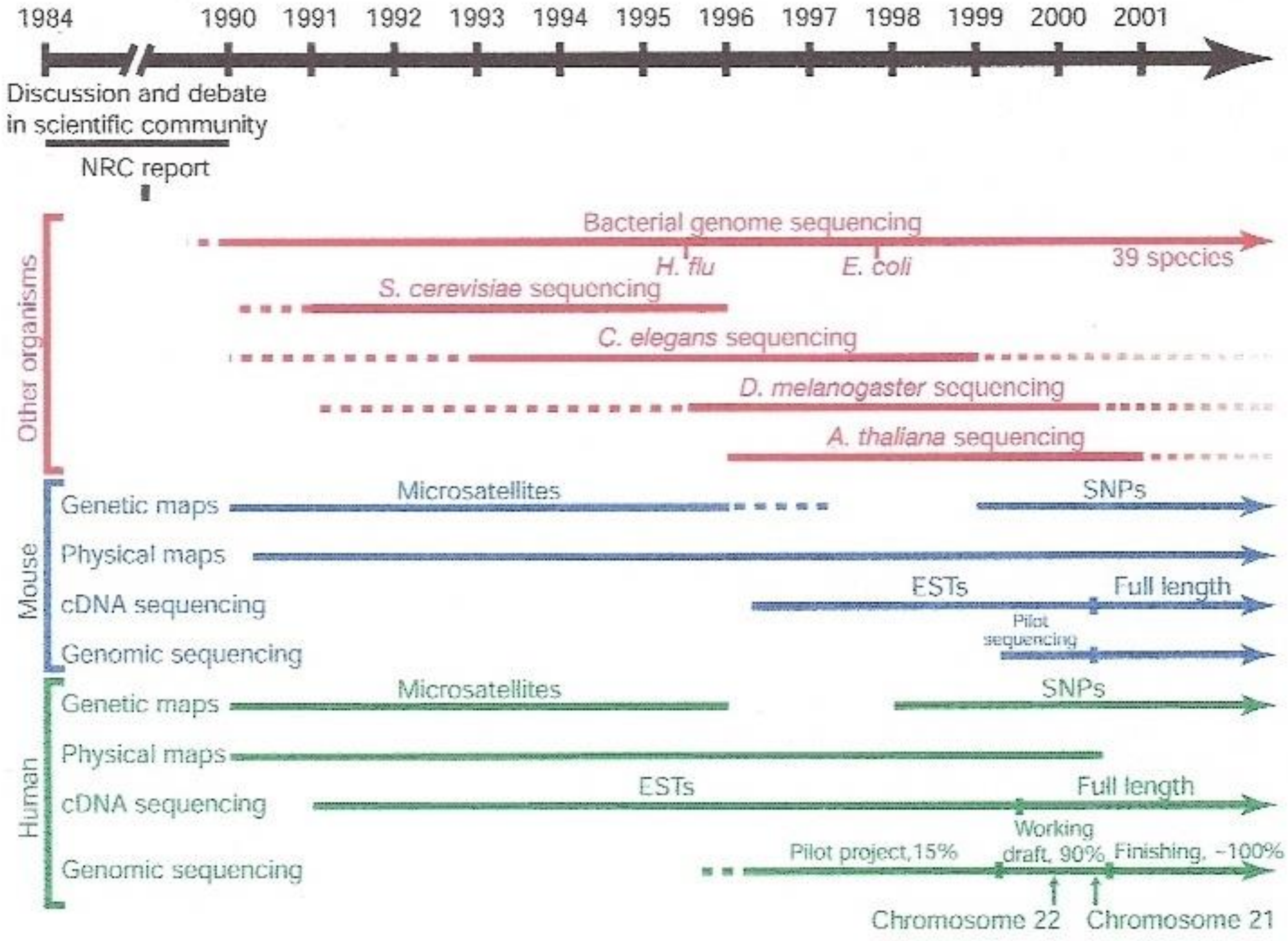
Enables merging of mapping data obtained from many labs using many different methods into a single consensus map of landmarks along a chromosome.

Eliminates the need for huge collections of cloned probe segments upon which prior maps depended.

Clone ends – Clone-based Physical Map


a. Screen library with existing markers
b. Generate new markers
c. Screen library with new markers
d. Determine tiling path

Genome projects at timescale



Goals of Human Genome Project

1. To identify all the genes in human DNA.
2. To develop a genetic linkage map of human genome.
3. To obtain a physical map of human genome.
4. To develop technology for the management of human genome information.
5. To know the function of genes.
6. Determine the sequences of the 3 billion chemical base pairs that make up human DNA.
7. Store this information in public databases.
8. Develop tools for data analysis.
9. Transfer related technologies to the private sectors.

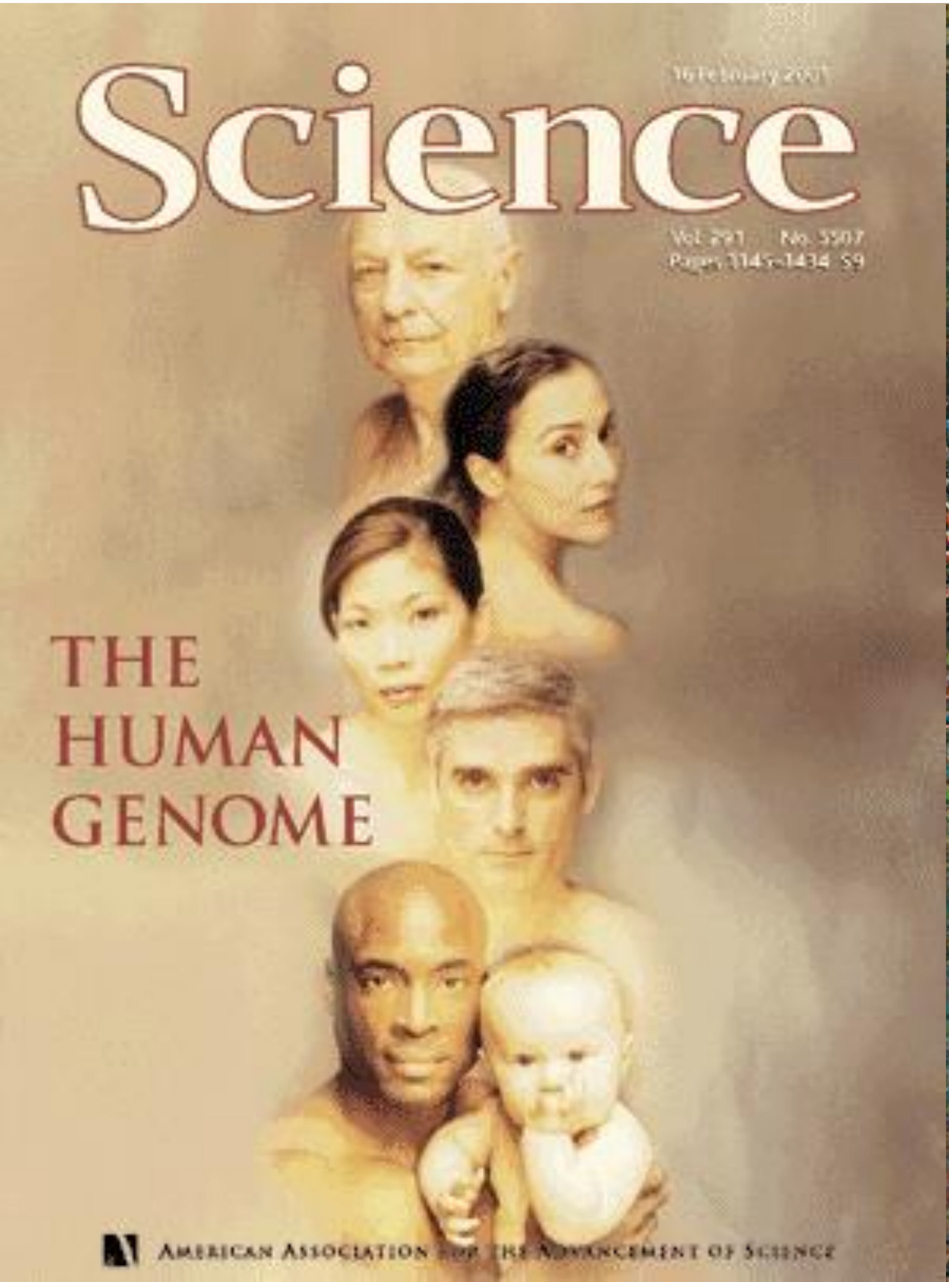
Human Genome Project - contributors and landscapes

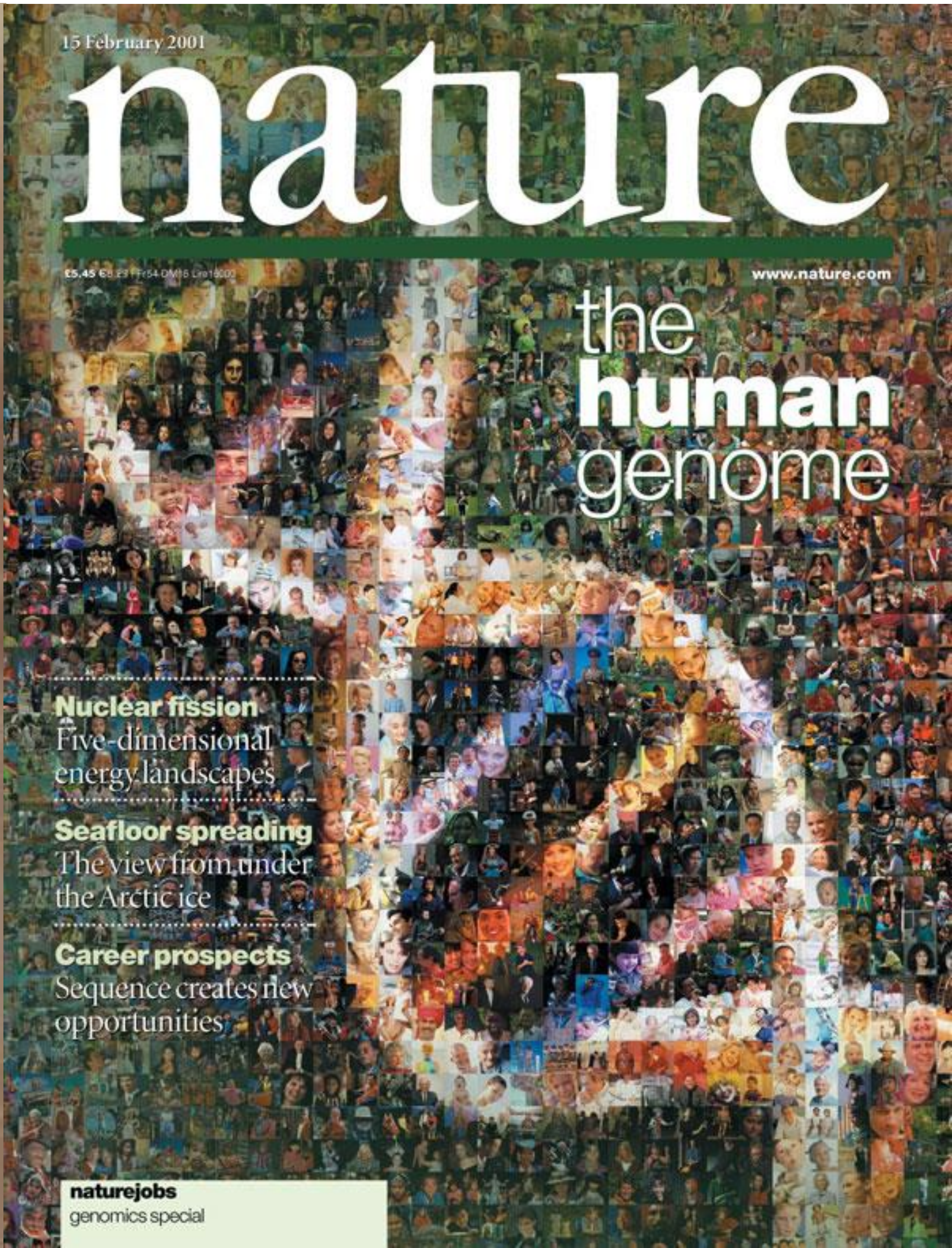
- **HUGO**: Human Genome Organization
- US DOE and NIH, UK MRC and WTSI, CEPH , FMDA, Japan, European Community (yeast genome), Germany, China.
- 1990-1995: genetic and physical mapping
- medical disorders, fixing physical loci, model organisms
- large-scale sequencing: two-phase paradigm „shotgun"
- 2001: draft genome sequence, 2003: full genome sequence

- **Celera Genomics**:
- Applied Biosystems., TIGR (C. Venter)
- 1998-2001: „whole genome shotgun"
- ABI PRISM 3700 DNA Analyzer

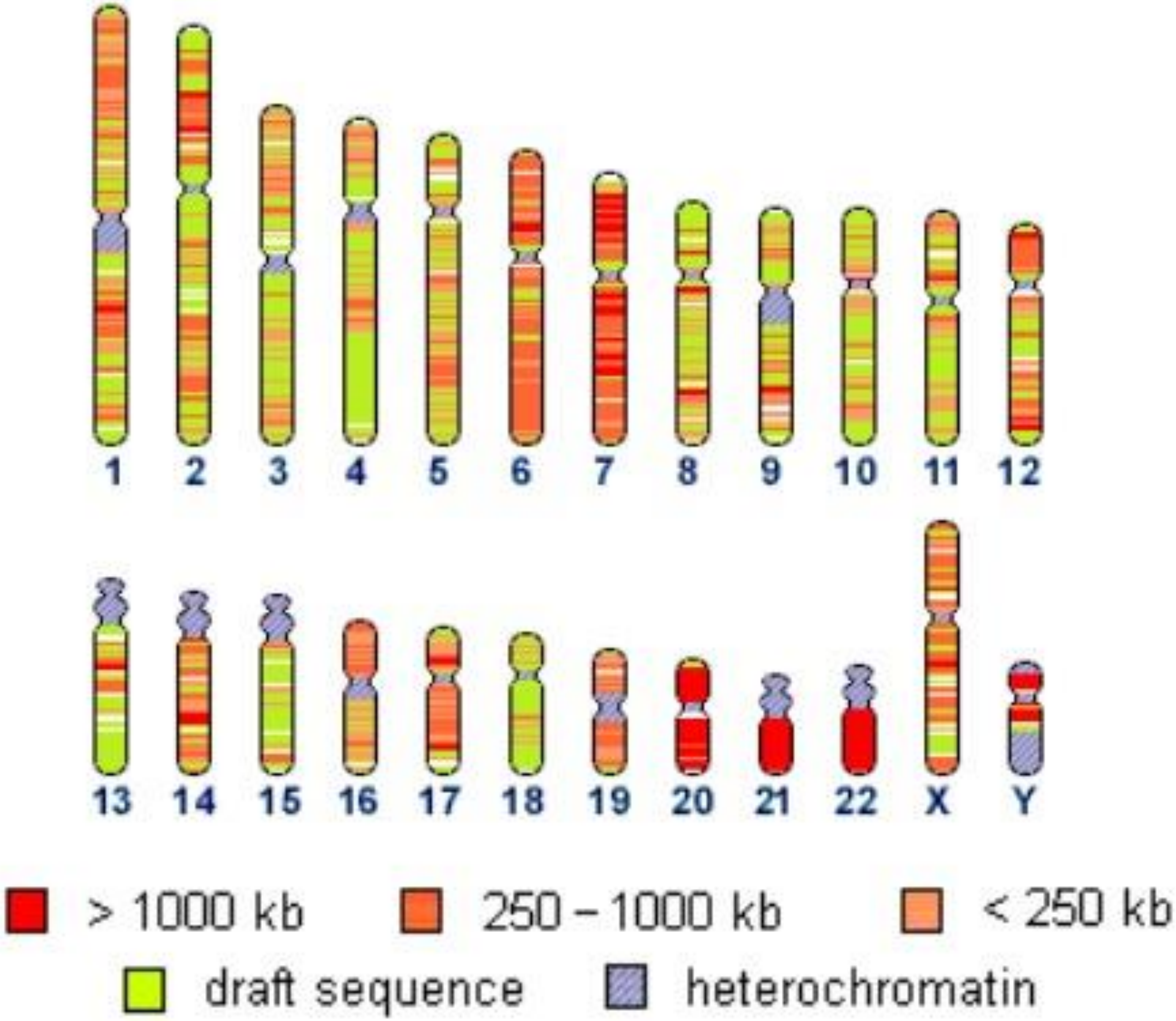
Technology speeds science. ABI sequencers at Venter Insitute, 2007.

Publishing the draft human genome





Human Genome Sequencing 2/11/2001
22 autosome + 2 sex chromosomes

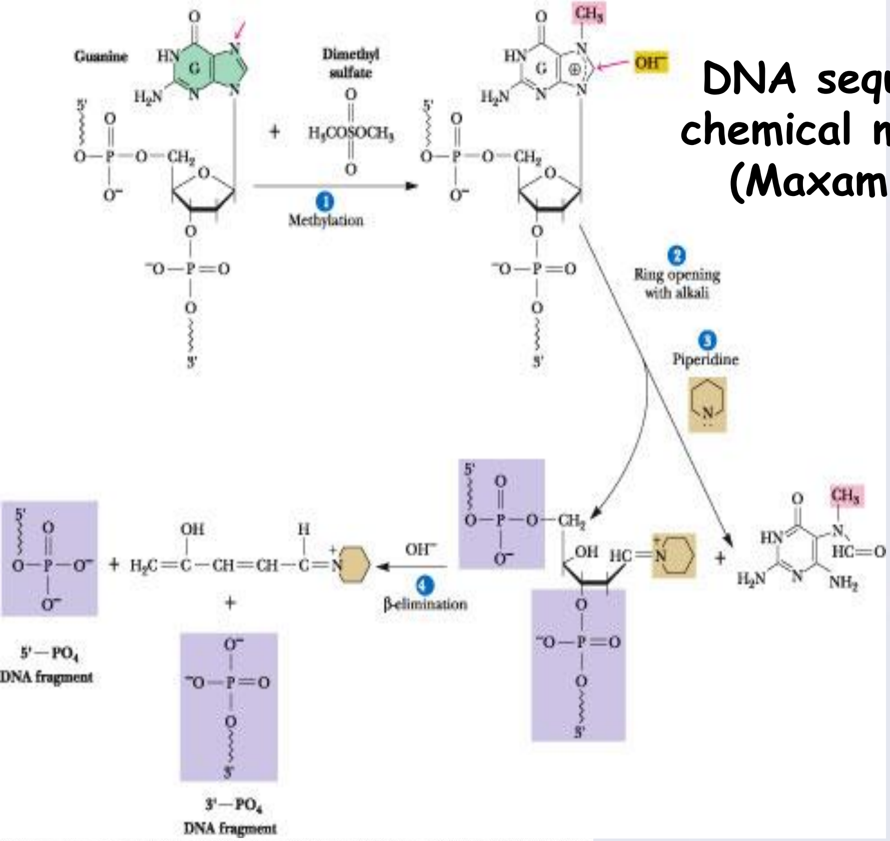


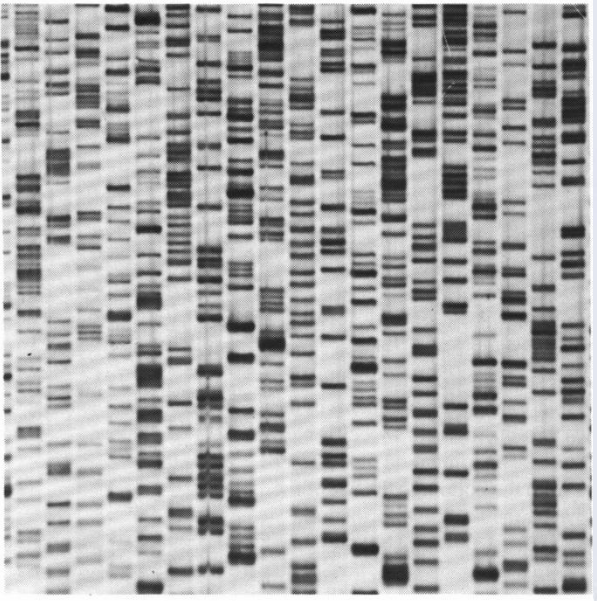
From NCBI

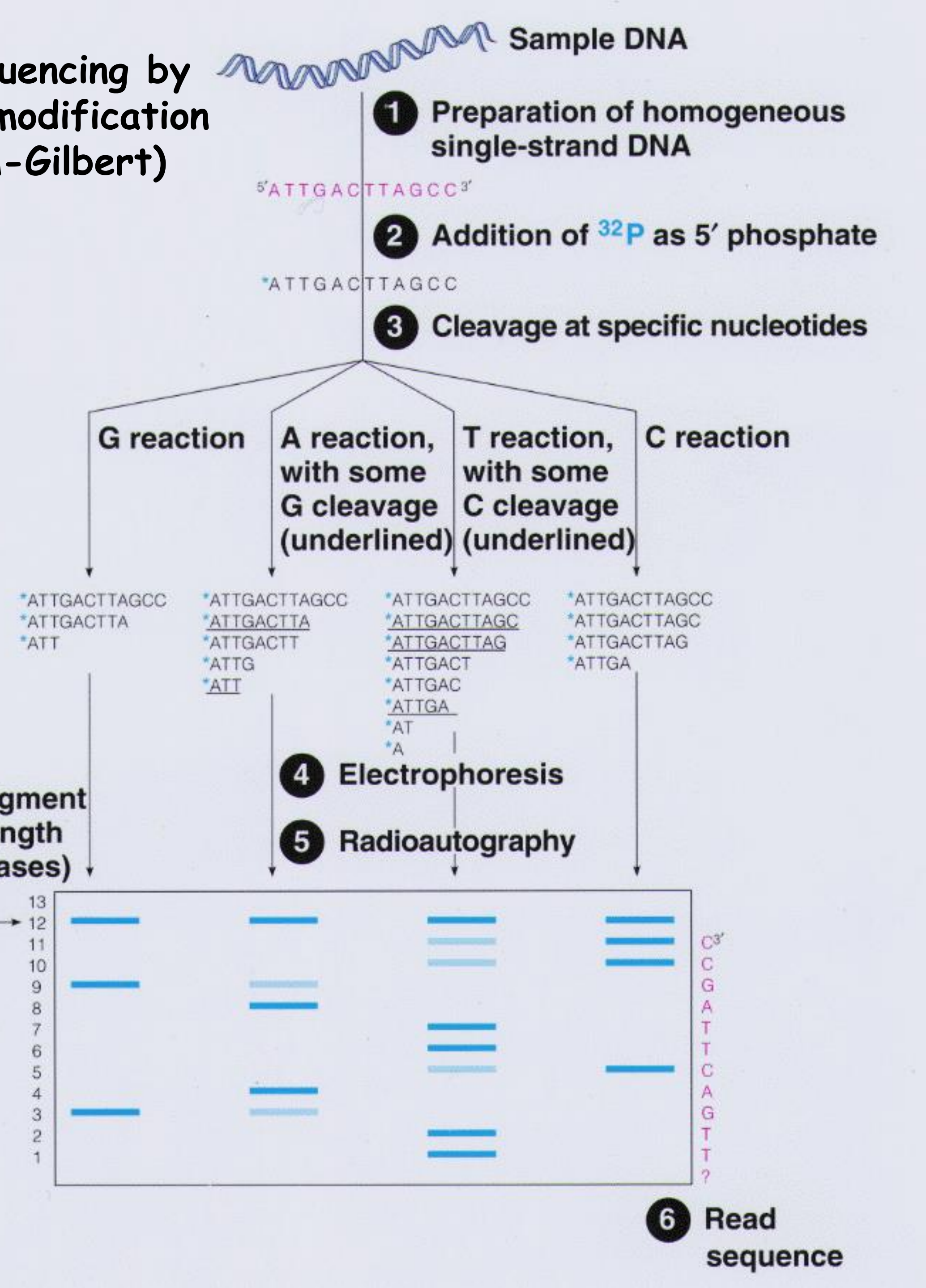
First results of the human genome draft sequence

- first Vertebrata genome, euchromatin region coverage around 96%
- considerable variability in distribution of genetic elements and features (ie. HOX clusters – „repeat poor”)
- ~ 30-40.000 genes, complexity and alternative splicing
- complex proteom, vertebrata-specific domain assambly
- horizontal gene transfer, transposable elements inactivation
- chromosome segments duplication (pericentromer, subtelomer)
- meiotic mutation rates in males and in females
- recombination rate varies between and along chromosomes
- more million of SNPs, genome-wide linkage mapping

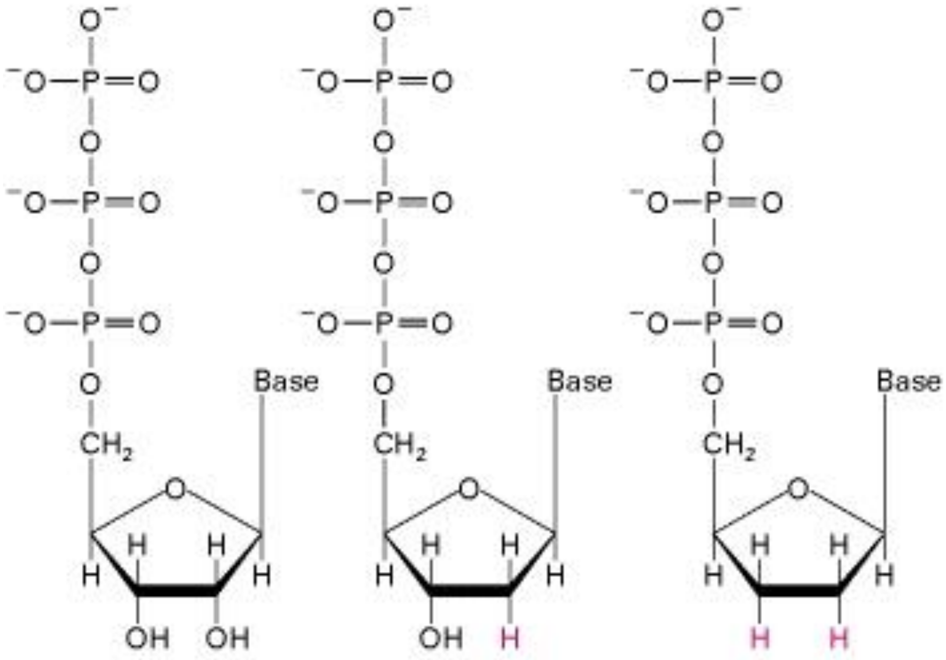
DNA sequencing by chemical modification (Maxam-Gilbert)







Sanger dideoxy sequencing

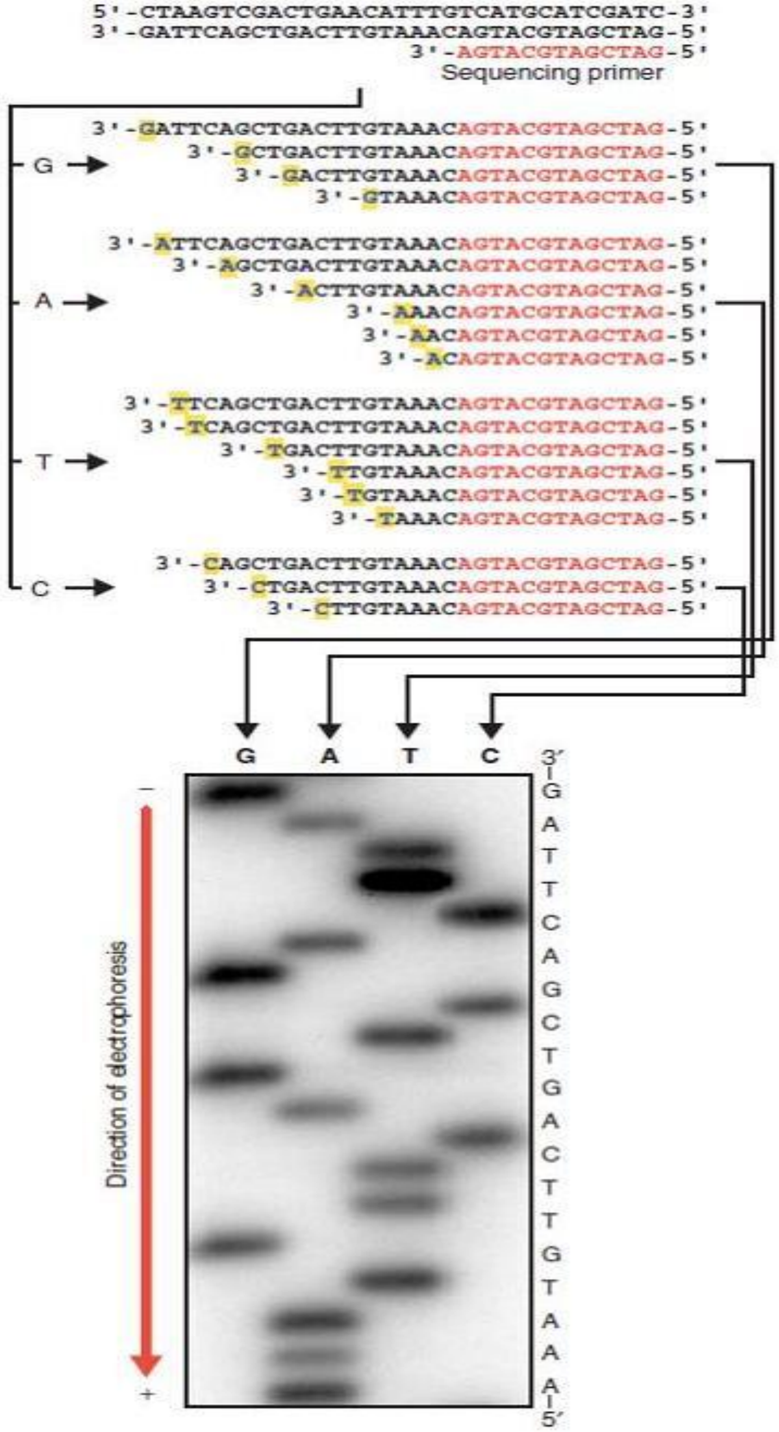


Ribonucleoside triphosphate (NTP)

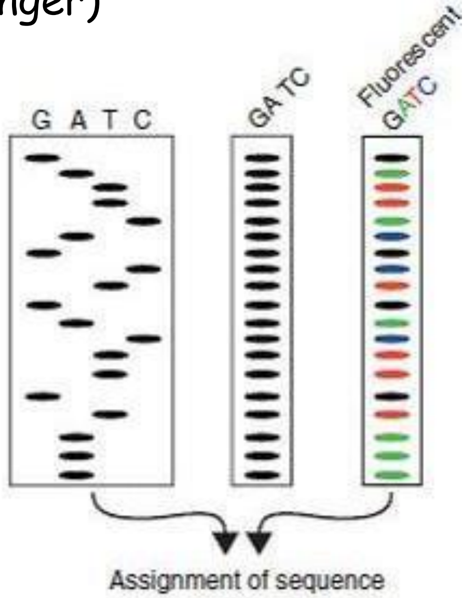
Deoxyribonucleoside triphosphate (dNTP)

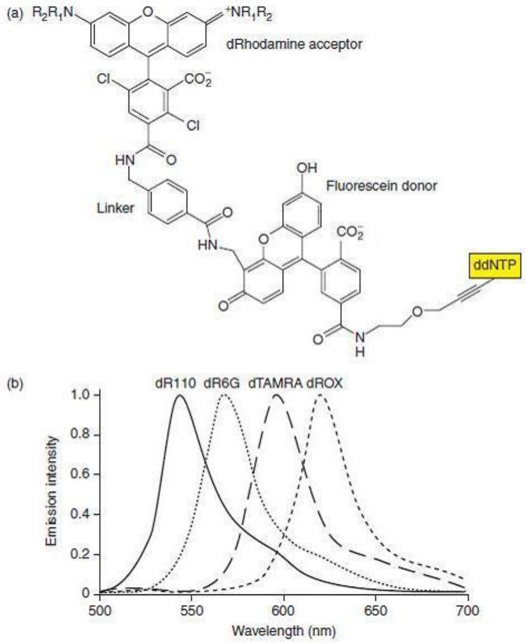
Dideoxyribonucleoside triphosphate (ddNTP)

Molecular cell biology, Lodish

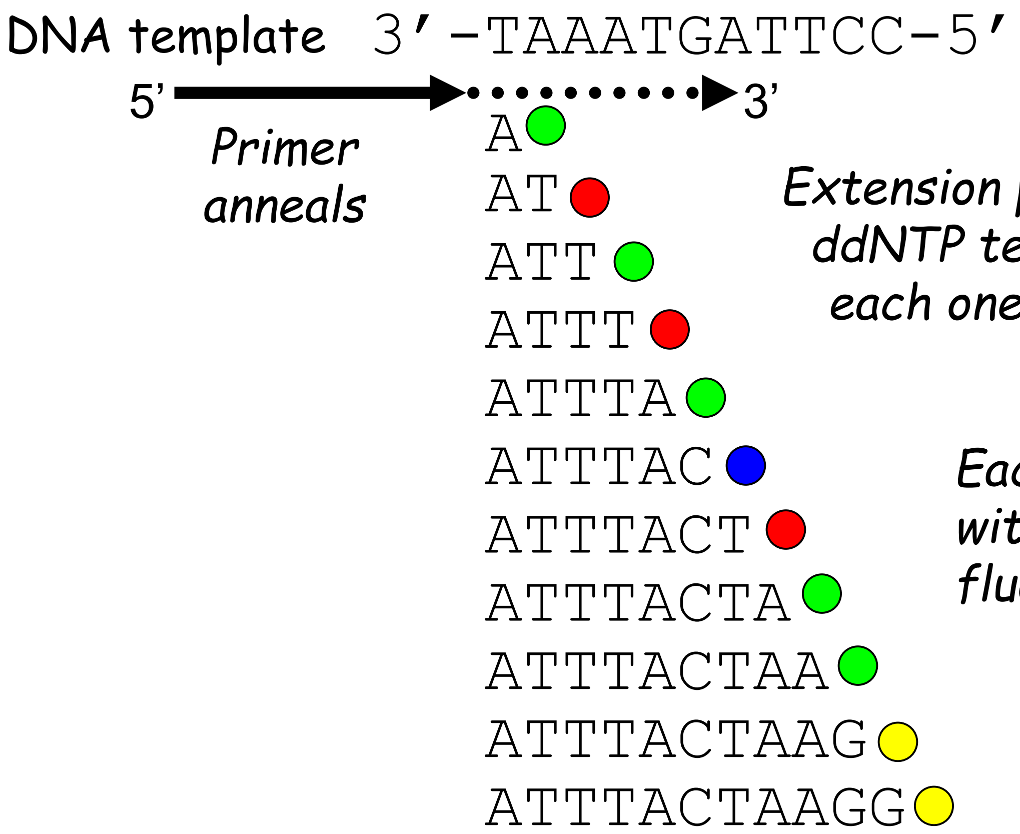

5'-CTAAGTCGACTGAACATTTGTCATGCATCGATC-3'
3'-GATTCAGCTGACTTGTAAACAGTACGTAGCTAG-5'
3'-AGTACGTAGCTAG-5'
Sequencing primer
G
3'-GATTCAGCTGACTTGTAAACAGTACGTAGCTAG-5'
3'-GCTGACTTGTAAACAGTACGTAGCTAG-5'
3'-GACTTGTAAACAGTACGTAGCTAG-5'
3'-GTAAACAGTACGTAGCTAG-5'
A
3'-ATTCAGCTGACTTGTAAACAGTACGTAGCTAG-5'
3'-AGCTGACTTGTAAACAGTACGTAGCTAG-5'
3'-ACTTGTAAACAGTACGTAGCTAG-5'
3'-AAACAGTACGTAGCTAG-5'
3'-AACAGTACGTAGCTAG-5'
3'-ACAGTACGTAGCTAG-5'
T
3'-TTCAGCTGACTTGTAAACAGTACGTAGCTAG-5'
3'-TCAGCTGACTTGTAAACAGTACGTAGCTAG-5'
3'-TGACTTGTAAACAGTACGTAGCTAG-5'
3'-TTGTAAACAGTACGTAGCTAG-5'
3'-TGTAAACAGTACGTAGCTAG-5'
3'-TAAACAGTACGTAGCTAG-5'
C
3'-CAGCTGACTTGTAAACAGTACGTAGCTAG-5'
3'-CTGACTTGTAAACAGTACGTAGCTAG-5'
3'-CTTGTAAACAGTACGTAGCTAG-5'
G
A
T
C
3'
G
A
T
T
C
A
G
C
T
G
A
C
T
T
G
T
A
A
A
5'
−
+
Direction of electrophoresis

DNA sequencing by chain termination (Sanger)



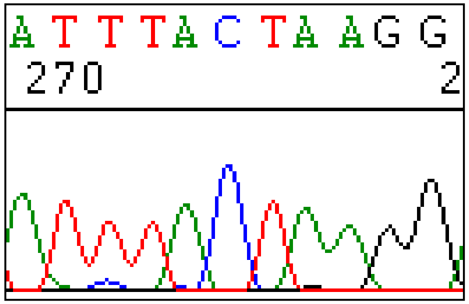


BigDye Terminator DNA Sequencing



Extension produces a series of ddNTP terminated products each one base different in length

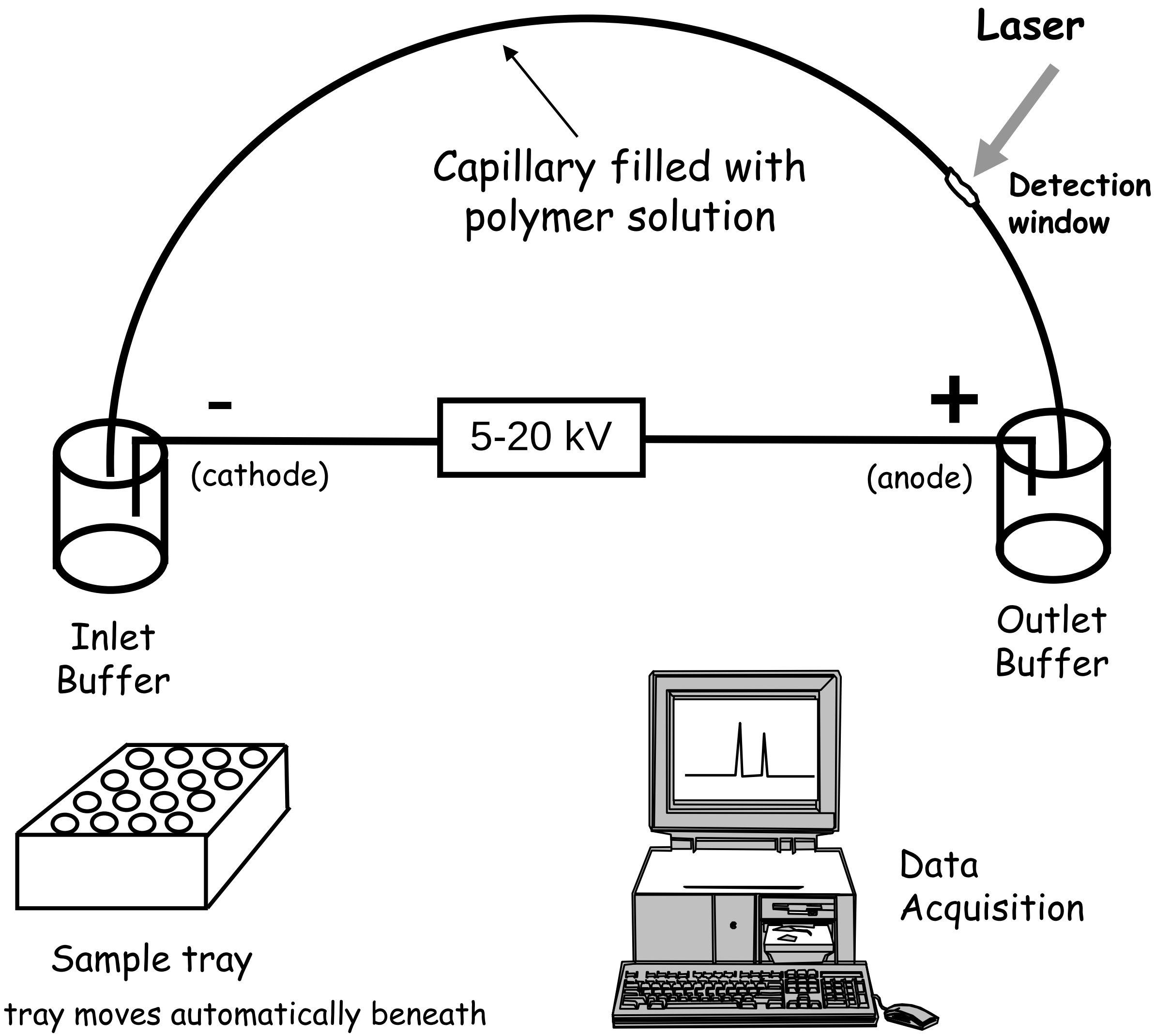
Each ddNTP is labeled with a different color fluorescent dye



Sequence is read by noting peak color in electropherogram (possessing single base resolution)

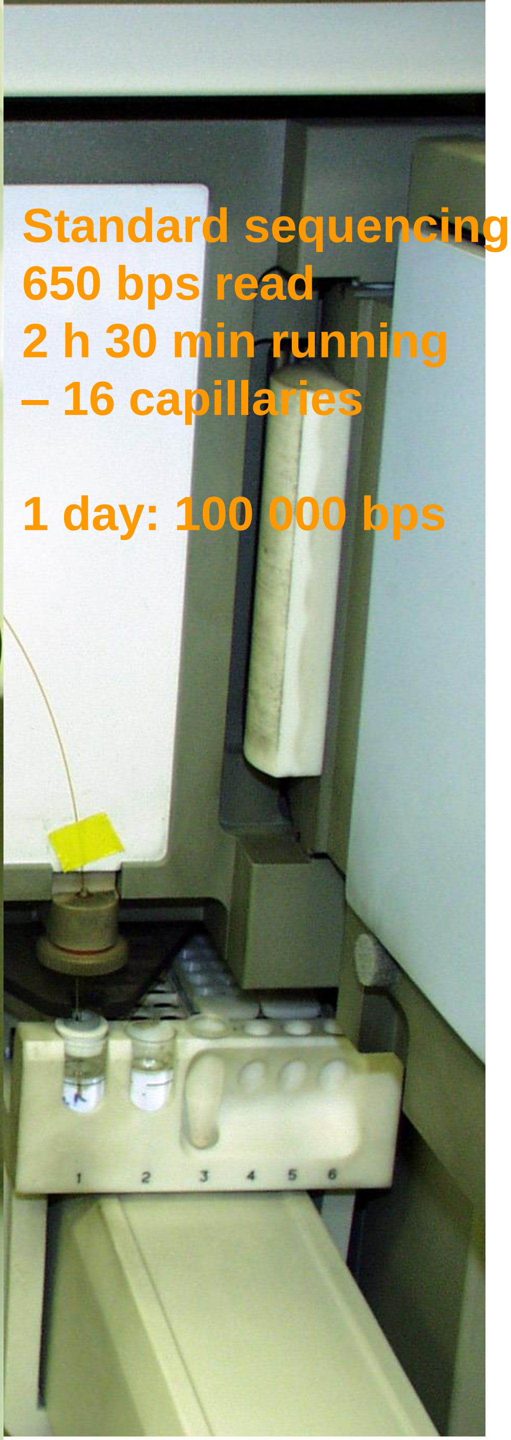
Figure 10.5, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

DNA sequencing: developement in technology and in bioinformatics



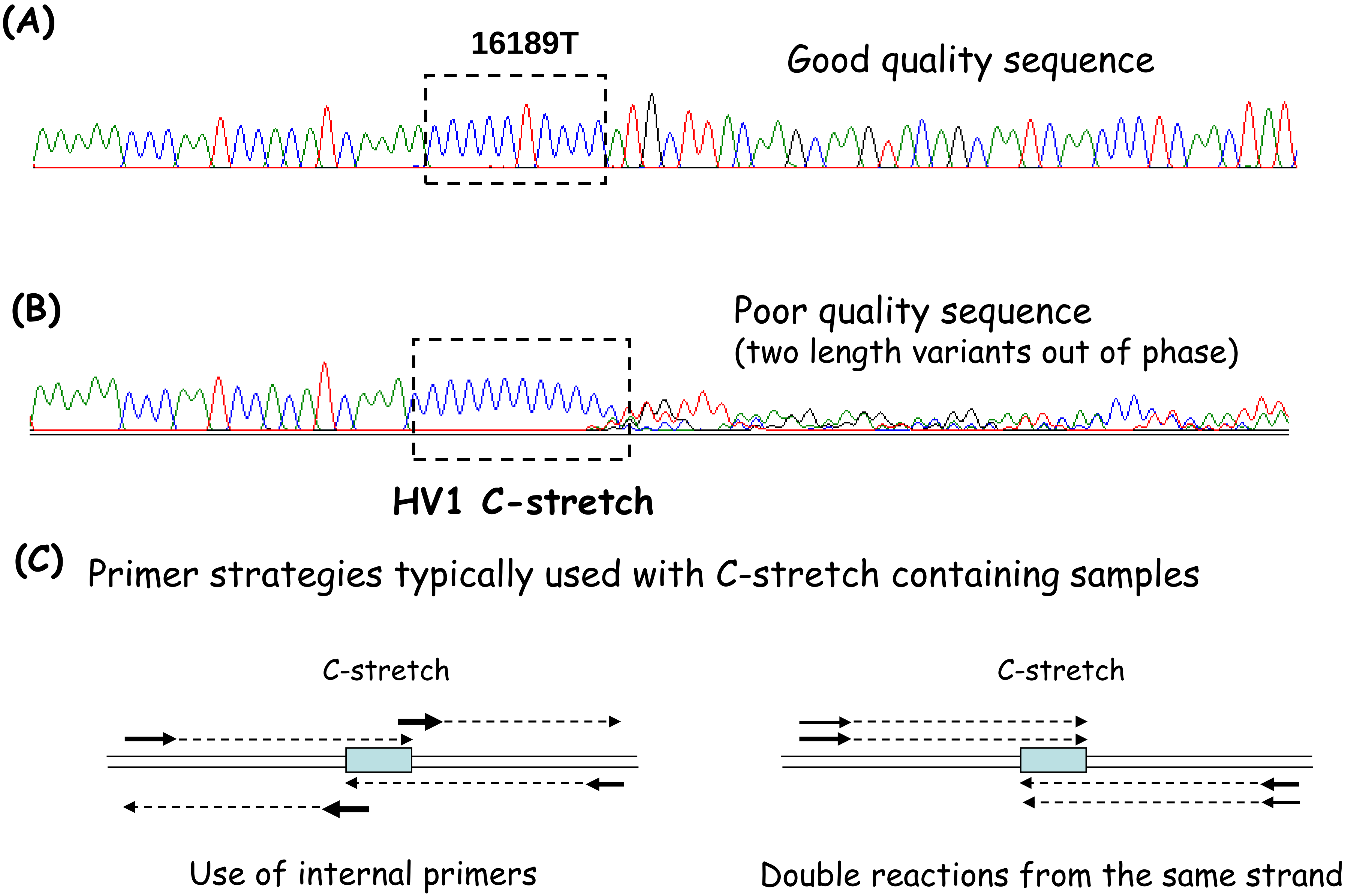
Sample tray moves automatically beneath
the cathode end of the capillary to
deliver each sample in succession

Figure 10.9, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press



Standard sequencing:
650 bps read
2 h 30 min running
– 16 capillaries

1 day: 100 000 bps



(A)

16189T

Good quality sequence

(B)

Poor quality sequence
(two length variants out of phase)

HV1 C-stretch

(C) Primer strategies typically used with C-stretch containing samples

C-stretch

Use of internal primers

C-stretch

Double reactions from the same strand

Figure 10.7, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

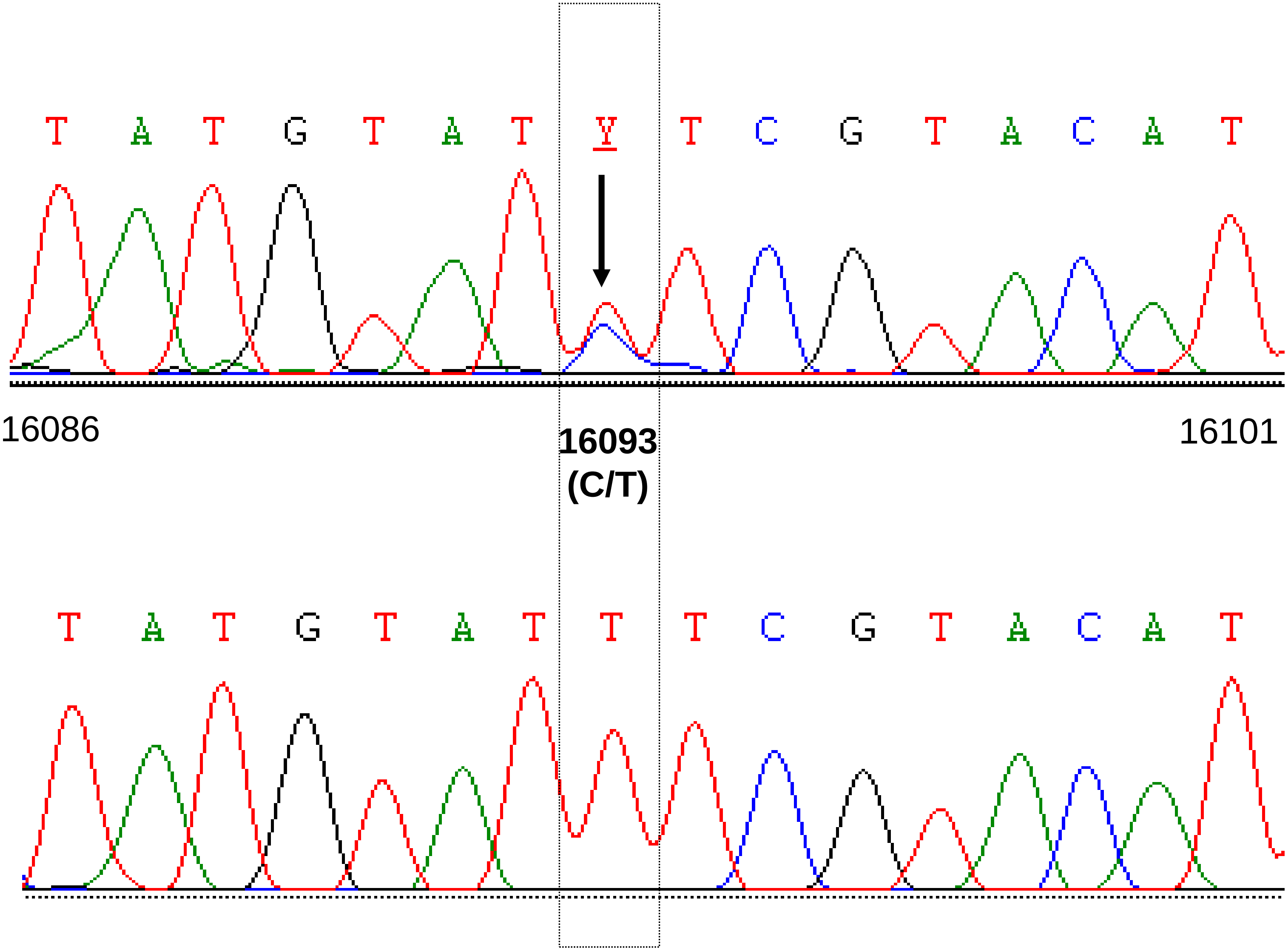


Figure 10.9, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

„shotgun" genome sequencing

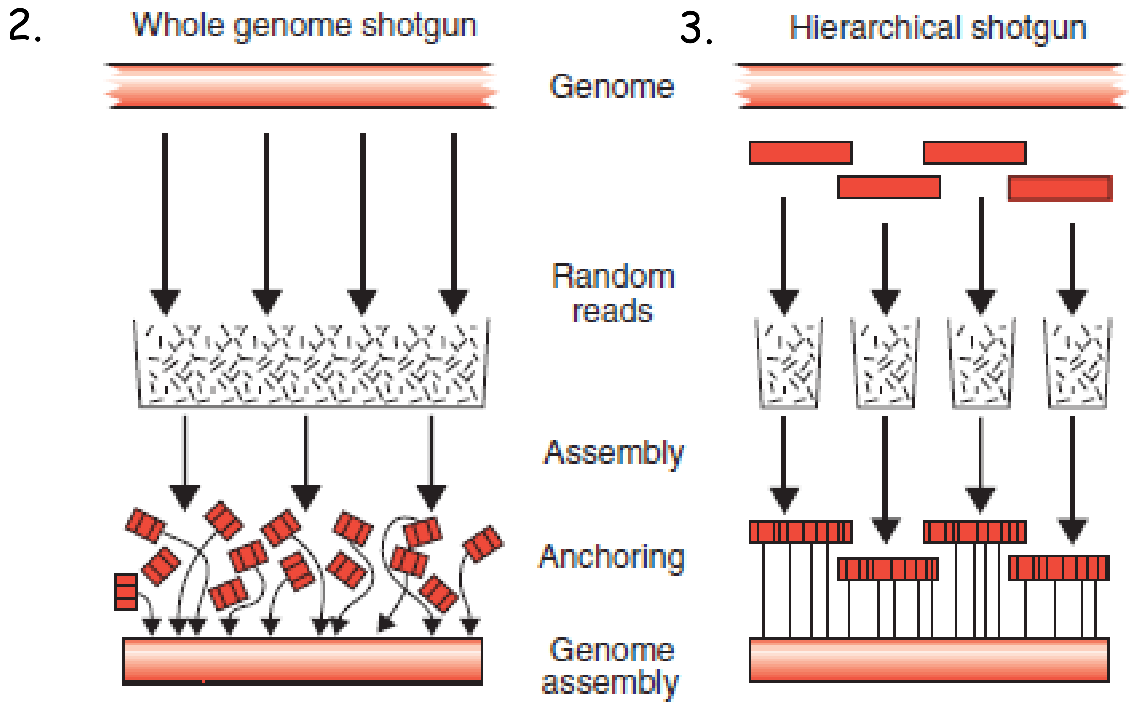


Figure 9.11. Assembling genomic data using the hierarchical and whole genome shotgun approaches. Adapted from Waterston, Lander and Sulston (2002), with permission

RJ Reece: Analysis of Genes and Genomes, 2004

Bacterial artificial chromosome (BAC)

A bacterial artificial chromosome (BAC) is an engineered DNA molecule, used to clone DNA segment in bacterial cells (*E. coli*).

It is based on a well-known natural **F plasmid** (inhabits *E. coli* cells). This plasmid allows conjugation between bacterial cells.

- Segments of an organism's DNA, ranging from **150 to about 300 kilo base** pairs, can be inserted into BACs.
- These vectors are able to maintain in **stable state *in vivo* and *in vitro***.
- Their copy number is about **two per cell**.
- Extensively **used in analysis of large genomes** but the main disadvantage of BAC vectors is some what **laborious construction** of BAC libraries.

Common gene components

Bacterial artificial chromosome is another cloning vector system in *E.coli* (**pBAC108L)**, developed by Melsimon and his colleagues in 1992, have

□ ***Hind*****III and *Bam*****HI**: the cloning sites

□ **CmR**: the **chloramphenicol resistance gene**, used as a selection tool.

□ ***oriS****:* the origin of replication

□ ***repE***: for plasmid replication and regulation of copy number.

□ ***ParA* and *ParB***: the genes governing partition of plasmids to daughter cells during division and ensures stable maintenance of the BAC.

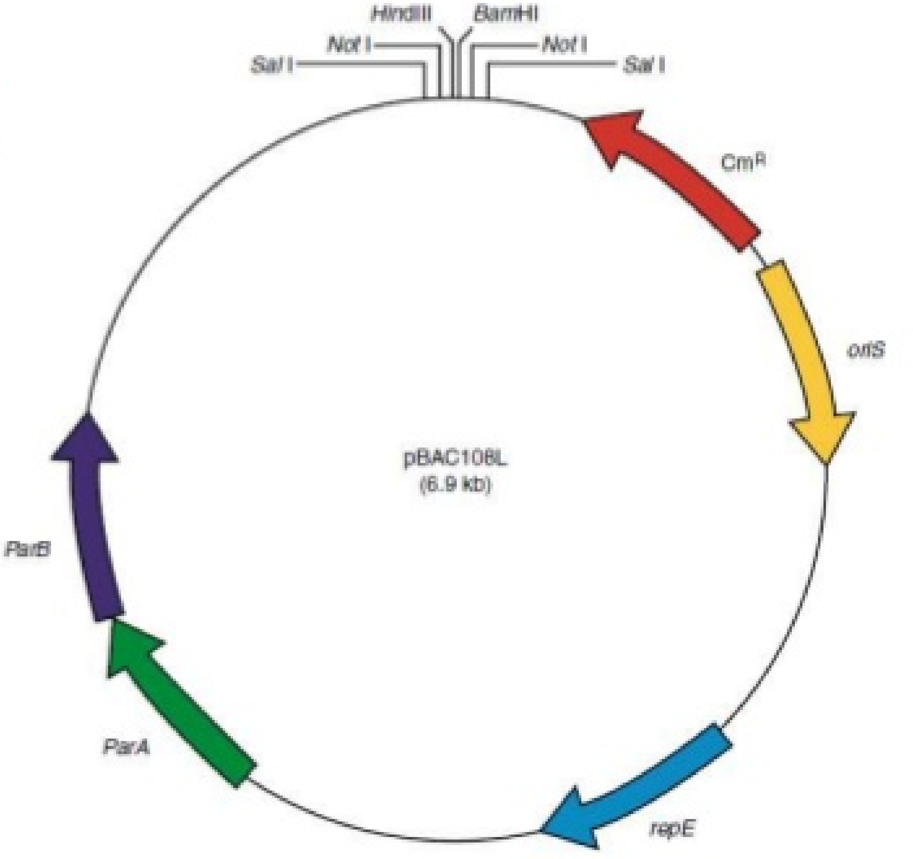


Fig: Map of the BAC vector, pBAC108L

Cloning genomic DNA into a BAC

1. Genomic DNA is isolated from a desired source and used restriction enzymes to cleave the target DNA into fragments.

2. The BAC is digested by restriction enzymes in the cloning sites ***Hin*dIII and *Bam*HI**.

3. Those two elements recombine by the DNA ligase and attach into a host bacterium.

4. As the bacterial cells grow and divide, they amplify the BAC DNA, which can then be isolated and used in sequencing DNA.

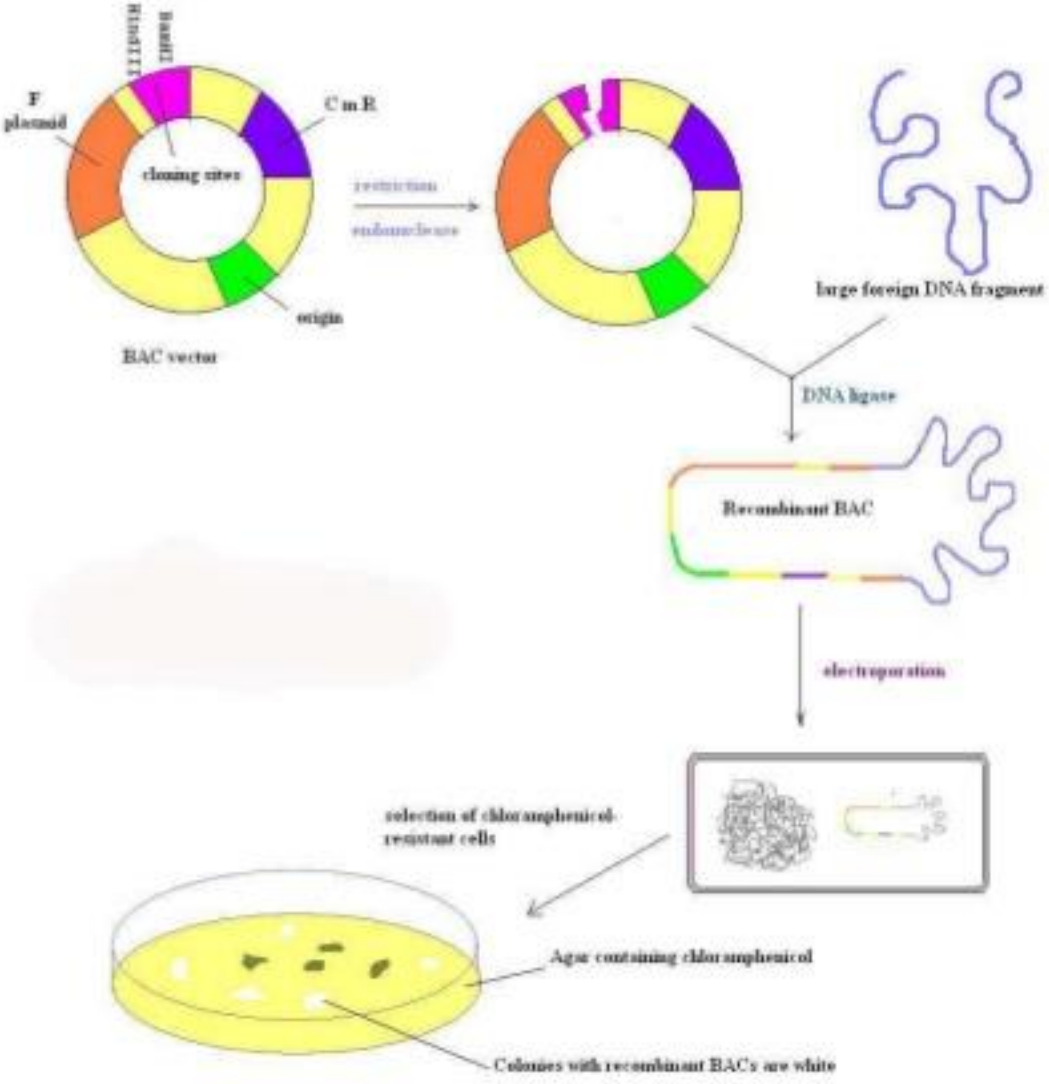
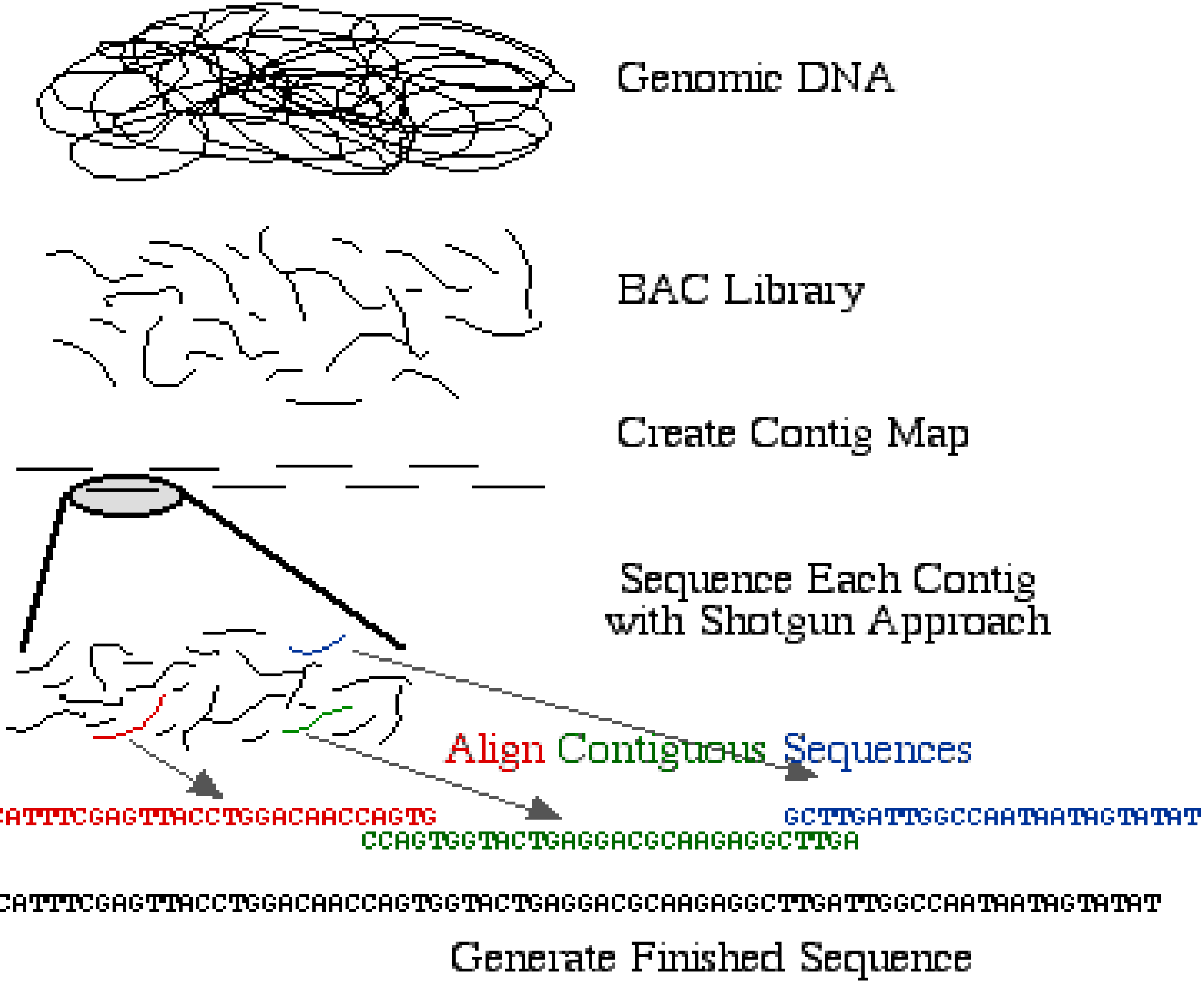

HindIII
BamHI
F plasmid
CmR
cloning sites
origin
BAC vector
restriction endonuclease
large foreign DNA fragment
DNA ligase
Recombinant BAC
electroporation
selection of chloramphenicol-resistant cells
Agar containing chloramphenicol
Colonies with recombinant BACs are white

Fig: BAC as a Cloning vector

Hierarchical Shotgun Sequencing Method



Genomic DNA

BAC Library

Create Contig Map

Sequence Each Contig
with Shotgun Approach

Align Contiguous Sequences

GCATTTCGAGTTACCTGGACAACCAGTG

CCAGTGGTACTGAGGACGCAAGAGGCTTGA

GCTTGATTGGCCAATAATAGTATAT

GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT

Generate Finished Sequence

STS genome mapping

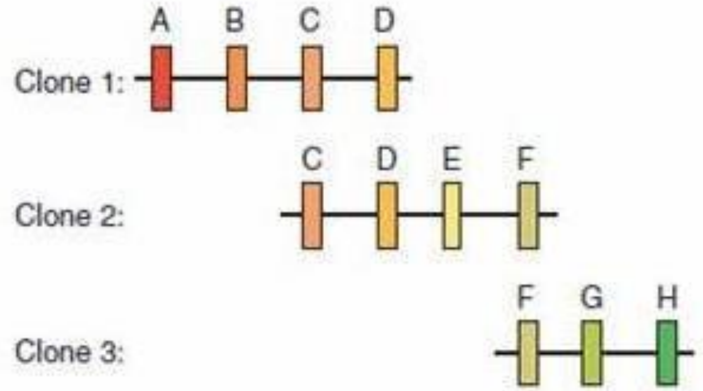


Figure 9.5. Aligning clones by STS mapping. Each clone contains several STSs. Clone 1 has four (A, B, C and D). Clone 2 also contains STSs C and D. Therefore clones 1 and 2 overlap with each other

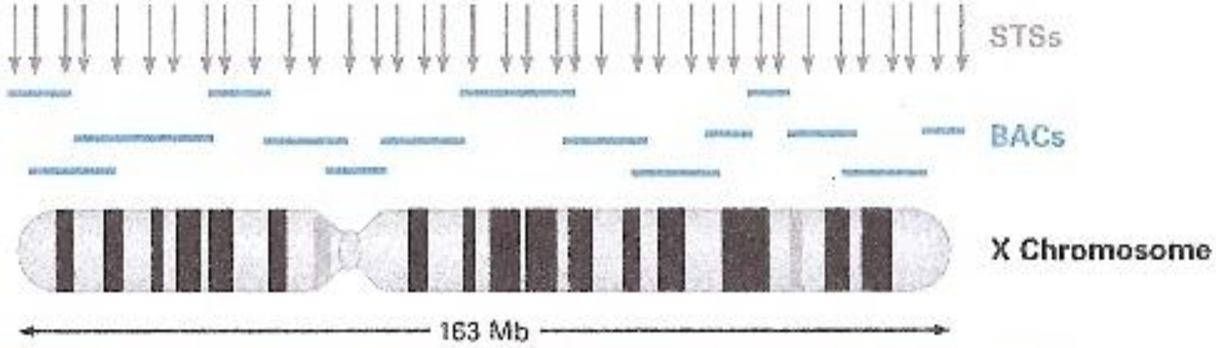


FIGURE 1.3 • Relationships of chromosomes to genome sequencing markers. The X chromosome is about 163 Mb in length. In this diagram, there are 16 overlapping BAC clones that span the entire length. In reality, 1,408 BACs were needed to span the X chromosome. Arrows (top) mark STSs scattered throughout the chromosome and on overlapping BACs.

Chromosome mapping

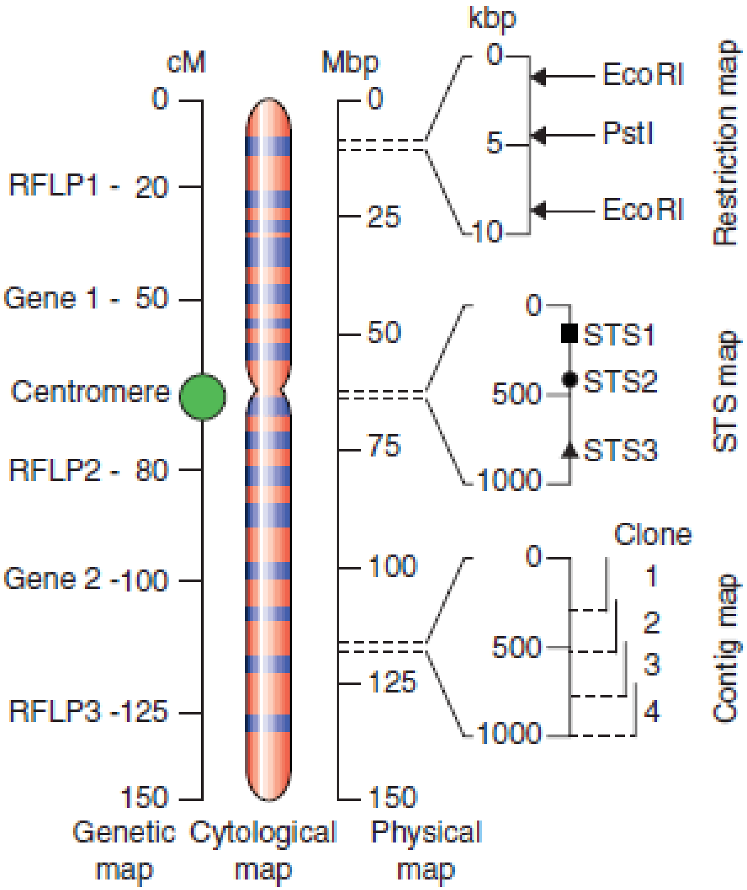
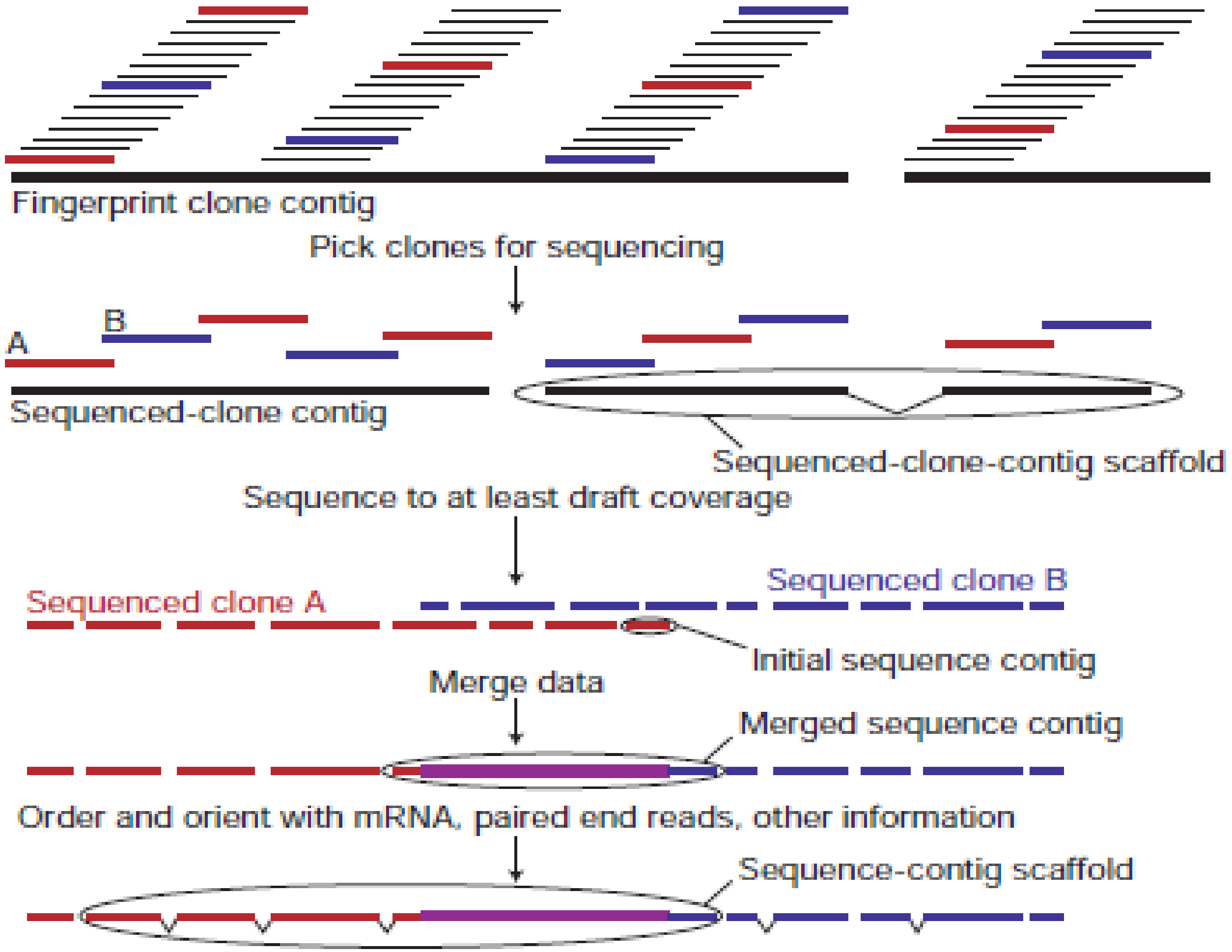


Figure 9.3. The different types of cytological, genetic and physical map of a chromosome. Genetic map distances are based on crossover frequencies and are measured in centiMorgans (cM), while physical distances are measured in megabase pairs (Mbp) or kilobase pairs (kbp)

RJ Reece: Analysis of Genes and Genomes, 2004

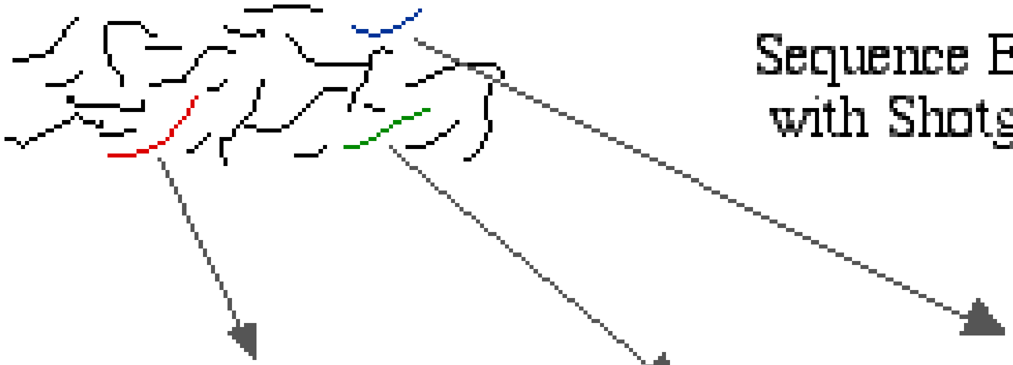
‚Fingerprint clone contig' assembly



International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome, Nature 409, 860 (2001)

Whole Genome Shotgun Sequencing Method

Genomic DNA



Sequence Each Fragment
with Shotgun Approach

GCATTTCGAGTTACCTGGACAACCAGTG

CCAGTGGTACTGAGGACGCAAGAGGCTTGA

GCTTGATTGGCCAATAATAGTATAT

Align Contiguous Sequences

GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT

Generate Finished Sequence

Whole genome sequence assembly

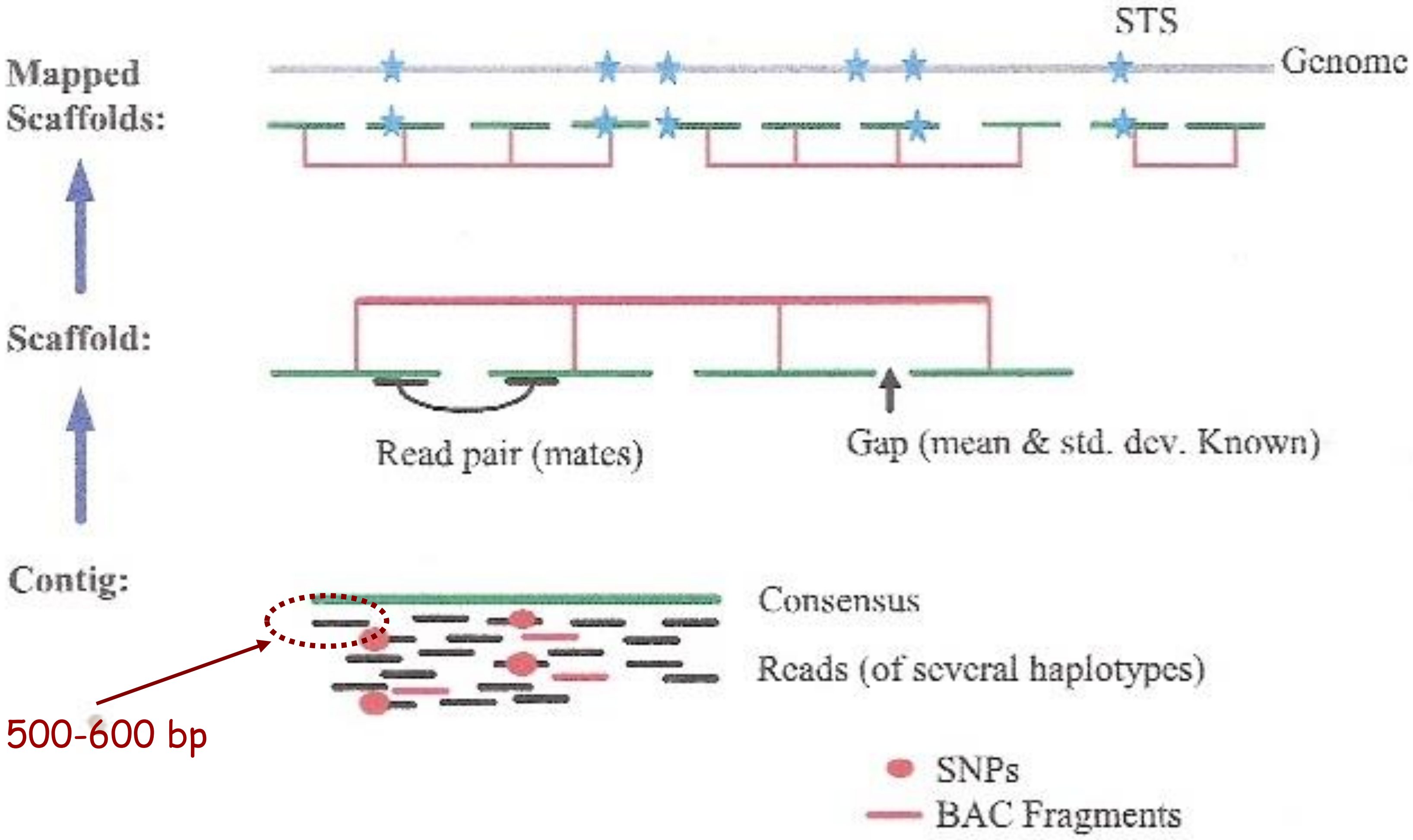


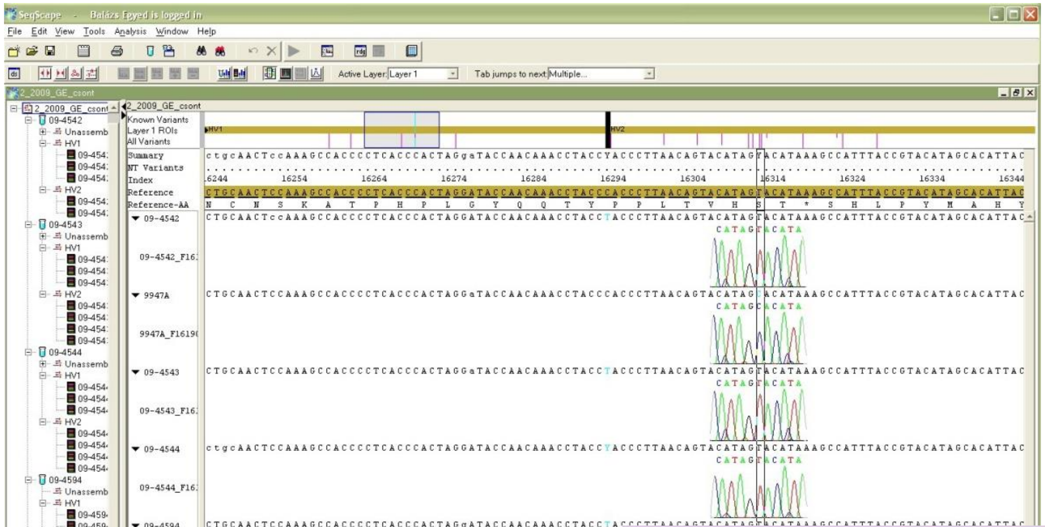
Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

JC Venter, et al.: The Sequence of the Human Genome, Science 291, 1304 (2001)

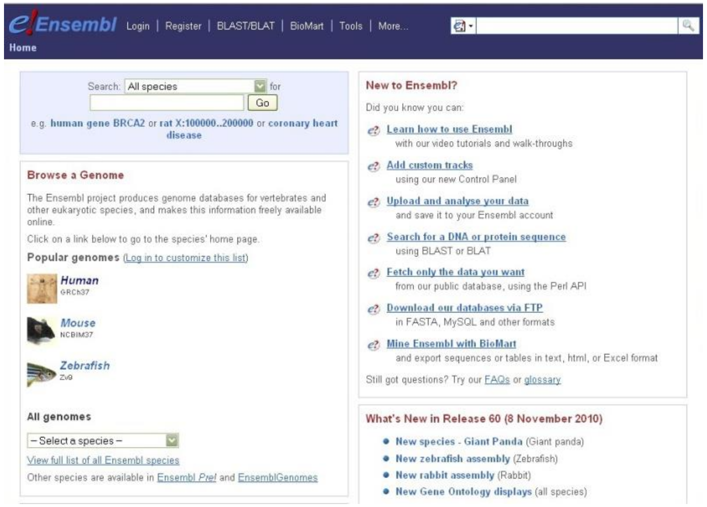
Human Genome Project
- preliminary results

- Finished in 2003 two years before planned
- 2001: draft sequence published (Science, Nature)
- DNA sequence gained from several persons' genomes
- Personal DNA and cell cultures
- Rate of failed nucleotides 1/10.000 (99,99 % accuracy)
- 4-5 X coverage, gaps closing (heterokromatin)
- Starting genome projects, annotation, data sharing:
- i.e. Ensemble, Human Genome Diversity Project, stb.

Genome sequencing: Technology and Bioinformatics



```
Summary       ctgcAACTccAAAGCCACCCCTCACCCACTAGGaTACCAACAAACCTACCYACCCTTAACAGTACATAGYACATAAAGCCATTTACCGTACATAGCACATTAC
NT Variants   .......................................................................................................
Index         16244     16254     16264     16274     16284     16294     16304     16314     16324     16334     16344
Reference     CTGCAACTCCAAAGCCACCCCTCACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTACATAAAGCCATTTACCGTACATAGCACATTAC
Reference-AA  N  C  N  S  K  A  T  P  H  P  L  G  Y  Q  Q  T  Y  P  P  L  T  V  H  S  T  *  S  H  L  P  Y  M  A  H  Y
09-4542       CTGCAACTccAAAGCCACCCCTCACCCACTAGGATACCAACAAACCTACCTACCCTTAACAGTACATAGTACATAAAGCCATTTACCGTACATAGCACATTAC
                                                                           CATAGTACATA
9947A         CTGCAACTCCAAAGCCACCCCTCACCCACTAGGaTACCAACAAACCTACCCACCCTTAACAGTACATAGCACATAAAGCCATTTACCGTACATAGCACATTAC
                                                                           CATAGCACATA
09-4543       CTGCAACTCCAAAGCCACCCCTCACCCACTAGGaTACCAACAAACCTACCTACCCTTAACAGTACATAGTACATAAAGCCATTTACCGTACATAGCACATTAC
                                                                           CATAGTACATA
09-4544       ctgcAACTCCAAAGCCACCCCTCACCCACTAGGATACCAACAAACCTACCYACCCTTAACAGTACATAGTACATAAAGCCATTTACCGTACATAGCACATTAC
                                                                           CATAGTACATA
09-4594       CTGCAACTCCAAAGCCACCCCTCACCCACTAGGgATACCAACAAACCTACCTACCCTTAACAGTACATAGTACATAAAGCCATTTACCGTACATAGCACATTAC
```



Search: All species for

e.g. human gene BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes (Log in to customize this list)

- *Human* GRCh37
- *Mouse* NCBIM37
- *Zebrafish* Zv9

All genomes

– Select a species –

View full list of all Ensembl species

Other species are available in Ensembl *Pre!* and EnsemblGenomes

New to Ensembl?

Did you know you can:

- **Learn how to use Ensembl** with our video tutorials and walk-throughs
- **Add custom tracks** using our new Control Panel
- **Upload and analyse your data** and save it to your Ensembl account
- **Search for a DNA or protein sequence** using BLAST or BLAT
- **Fetch only the data you want** from our public database, using the Perl API
- **Download our databases via FTP** in FASTA, MySQL and other formats
- **Mine Ensembl with BioMart** and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs or glossary

What's New in Release 60 (8 November 2010)

- **New species - Giant Panda** (Giant panda)
- **New zebrafish assembly** (Zebrafish)
- **New rabbit assembly** (Rabbit)
- **New Gene Ontology displays** (all species)

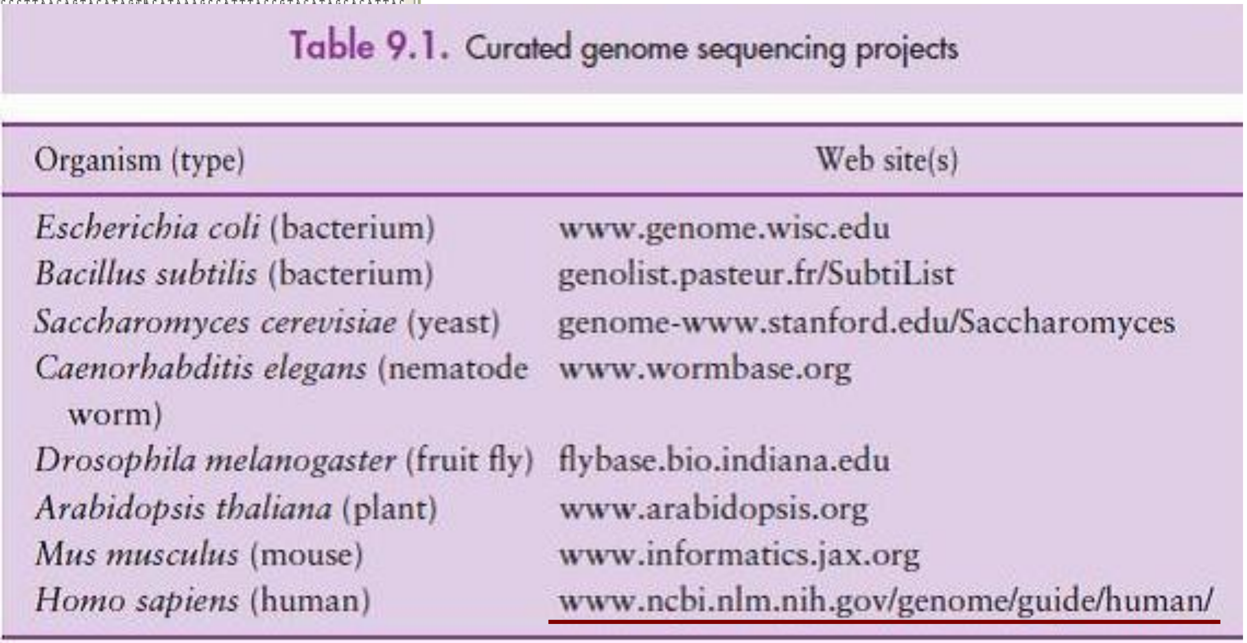
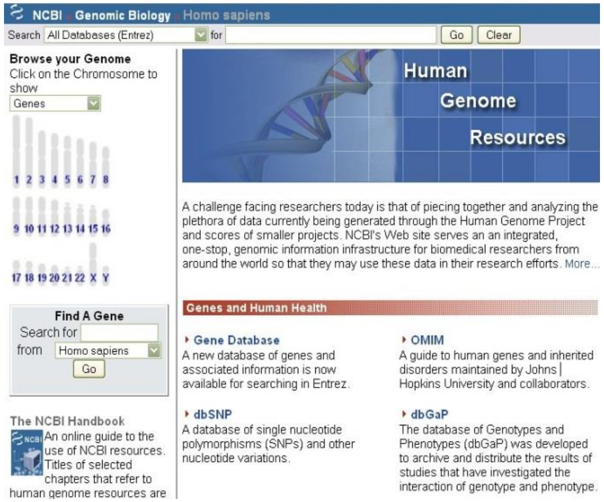


Table 9.1. Curated genome sequencing projects

Organism (type)	Web site(s)
Escherichia coli (bacterium)	www.genome.wisc.edu
Bacillus subtilis (bacterium)	genolist.pasteur.fr/SubtiList
Saccharomyces cerevisiae (yeast)	genome-www.stanford.edu/Saccharomyces
Caenorhabditis elegans (nematode worm)	www.wormbase.org
Drosophila melanogaster (fruit fly)	flybase.bio.indiana.edu
Arabidopsis thaliana (plant)	www.arabidopsis.org
Mus musculus (mouse)	www.informatics.jax.org
Homo sapiens (human)	www.ncbi.nlm.nih.gov/genome/guide/human/



Search All Databases (Entrez) for

Browse your Genome
Click on the Chromosome to show
Genes

Find A Gene
Search for
from Homo sapiens

The NCBI Handbook
An online guide to the use of NCBI resources. Titles of selected chapters that refer to human genome resources are

A challenge facing researchers today is that of piecing together and analyzing the plethora of data currently being generated through the Human Genome Project and scores of smaller projects. NCBI's Web site serves an an integrated, one-stop, genomic information infrastructure for biomedical researchers from around the world so that they may use these data in their research efforts. More...

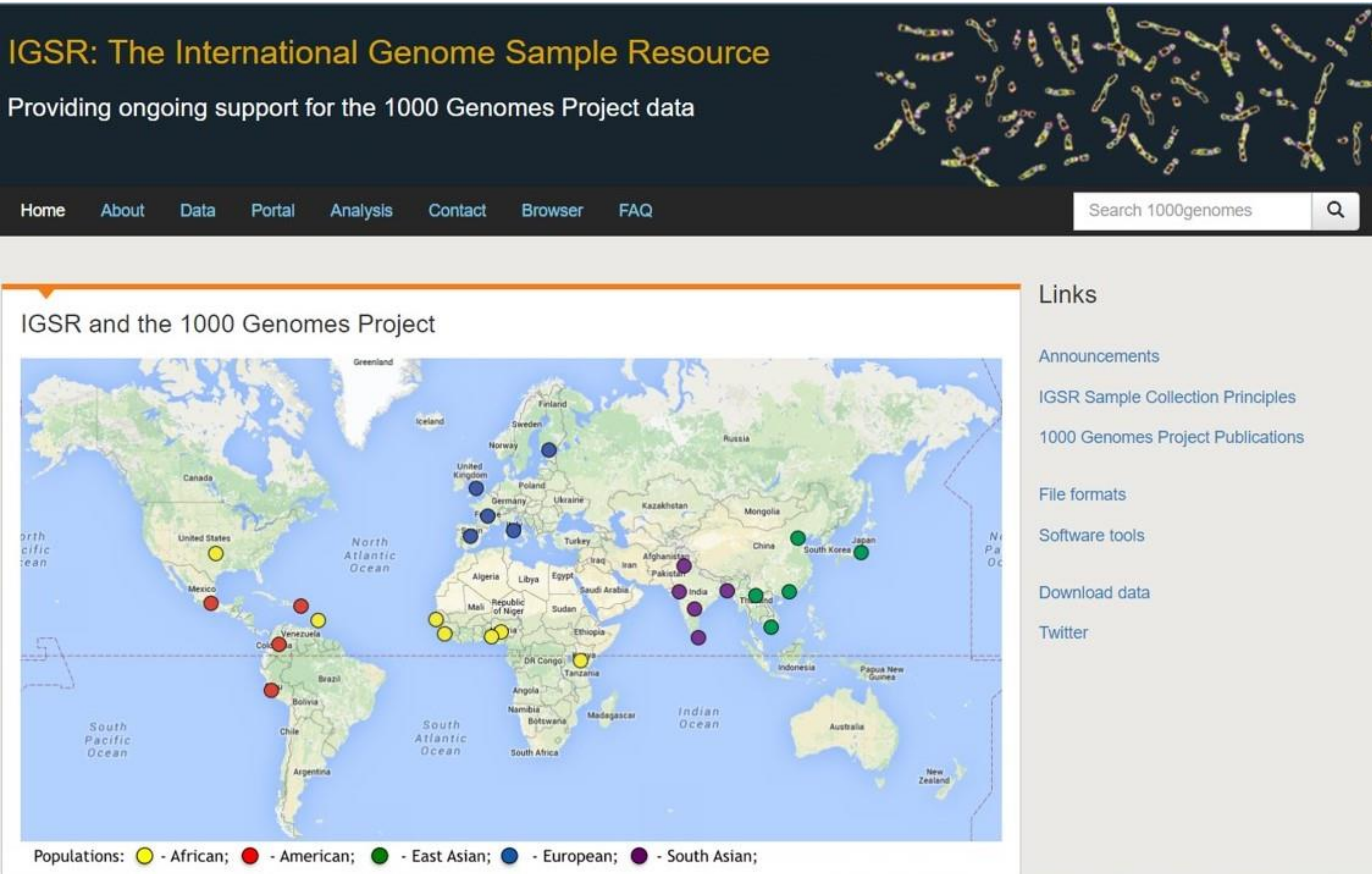
Genes and Human Health

Gene Database
A new database of genes and associated information is now available for searching in Entrez.

OMIM
A guide to human genes and inherited disorders maintained by Johns Hopkins University and collaborators.

dbSNP
A database of single nucleotide polymorphisms (SNPs) and other nucleotide variations.

dbGaP
The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype.



IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

IGSR and the 1000 Genomes Project

Populations: 🟡 - African; 🔴 - American; 🟢 - East Asian; 🔵 - European; 🟣 - South Asian;

Links

Announcements

IGSR Sample Collection Principles

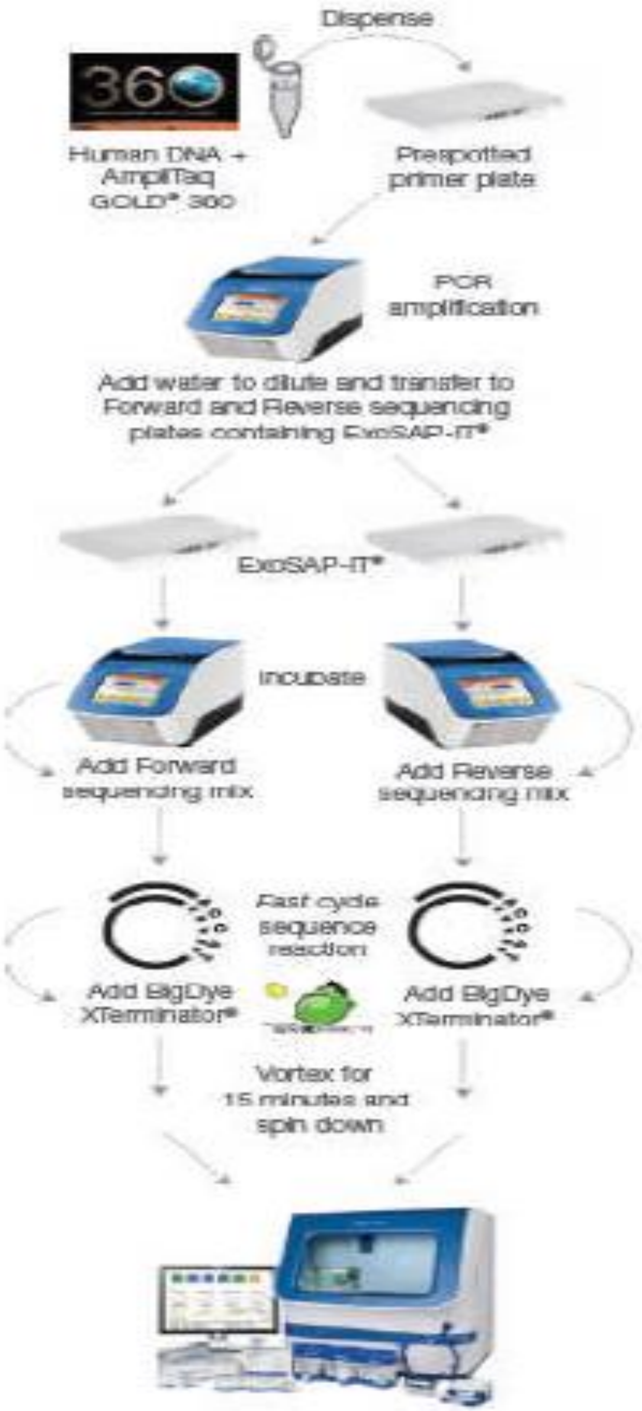
1000 Genomes Project Publications

File formats

Software tools

Download data

Twitter



Dispense

Human DNA + AmpliTaq GOLD® 360

Prespotted primer plate

PCR amplification

Add water to dilute and transfer to Forward and Reverse sequencing plates containing ExoSAP-IT®

ExoSAP-IT®

Incubate

Add Forward sequencing mix

Add Reverse sequencing mix

Fast cycle sequence reaction

Add BigDye XTerminator®

Add BigDye XTerminator®

Vortex for 15 minutes and spin down

BRCA1 / BRCA2 genes resequencing

- Molecular diagnostics of mutations

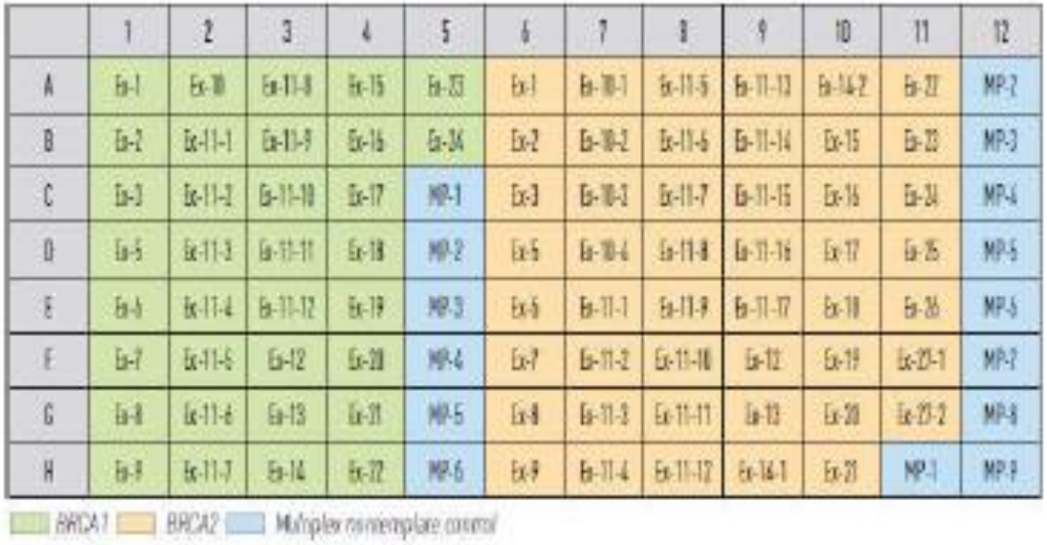
BRCA1 / BRCA2: 23 /27 exons (80Kb)

No prior screening: ~~SSCP, DGGE, dHPLC~~ etc.

One sample – one assay concept

Quick, accurate, full coverage

BRCA1 / BRCA2: 34 / 47 amplicons respectively

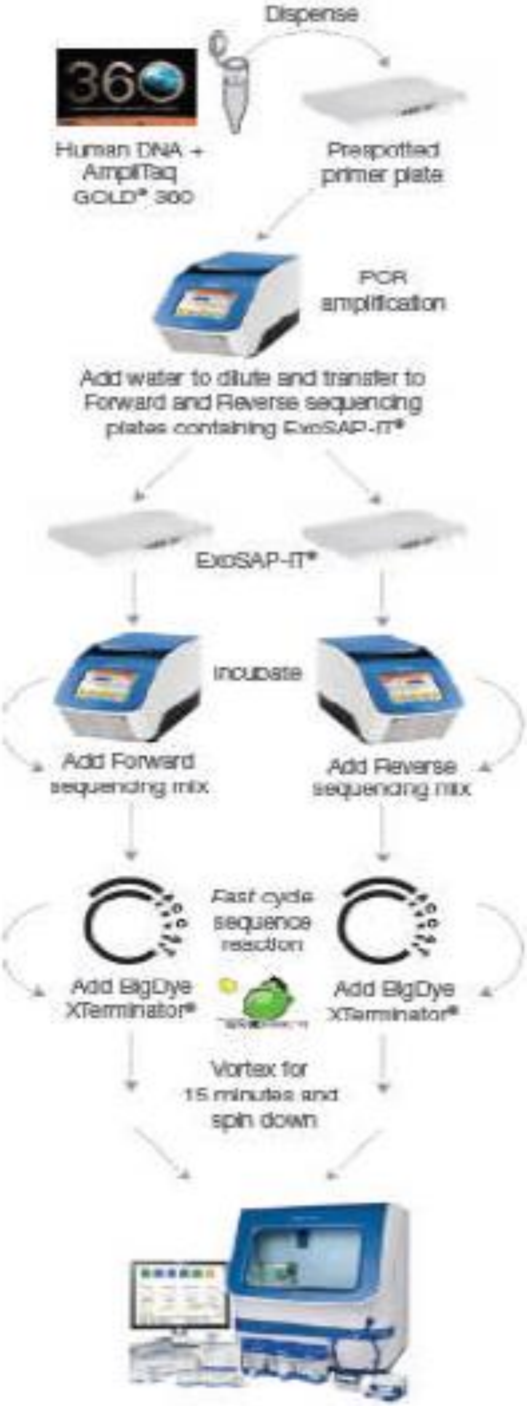


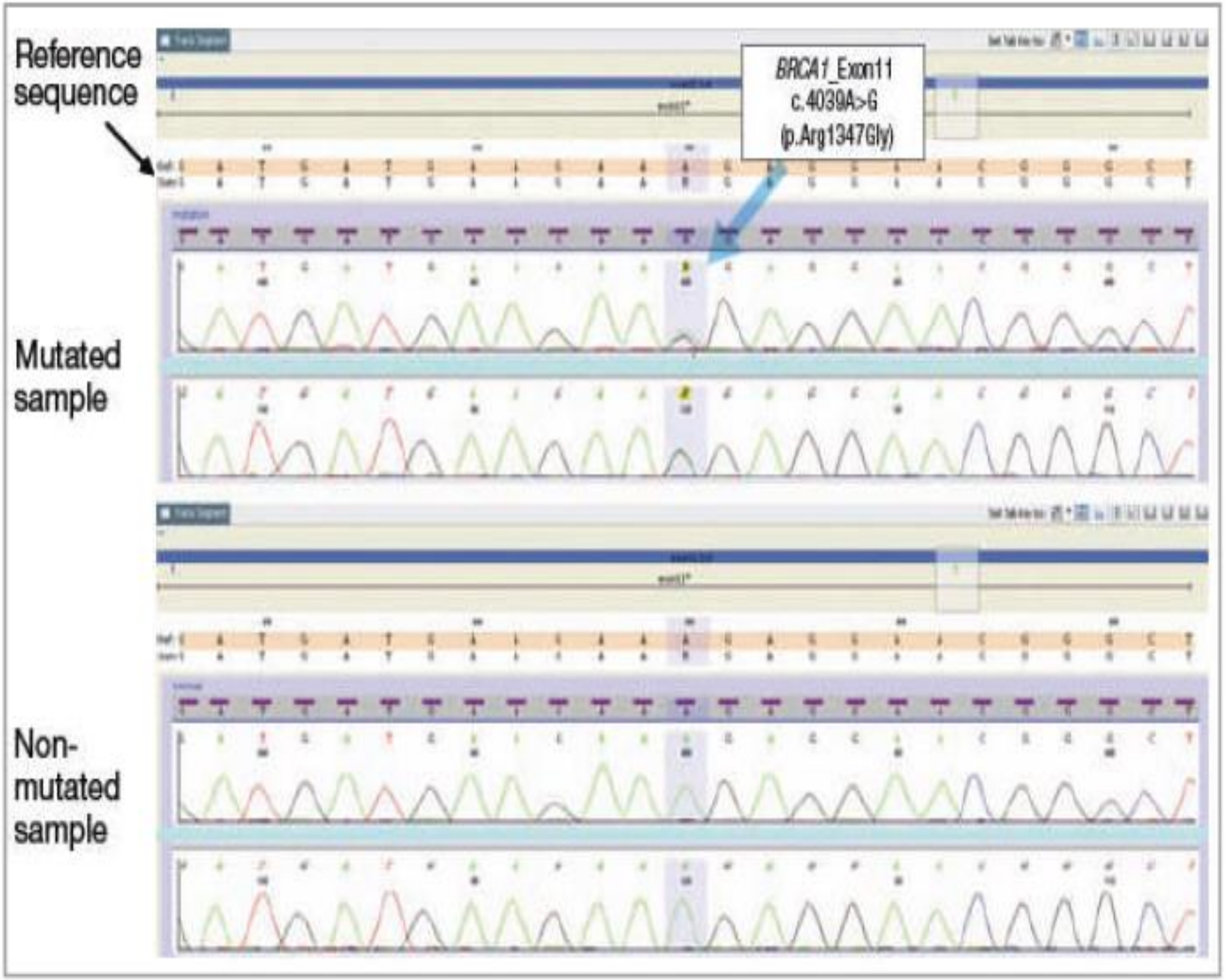
	1	2	3	4	5	6	7	8	9	10	11	12
A	Ex-1	Ex-10	Ex-11-8	Ex-15	Ex-23	Ex-1	Ex-10-1	Ex-11-5	Ex-11-13	Ex-14-2	Ex-22	MP-2
B	Ex-2	Ex-11-1	Ex-11-9	Ex-16	Ex-24	Ex-2	Ex-10-2	Ex-11-6	Ex-11-14	Ex-15	Ex-23	MP-3
C	Ex-3	Ex-11-2	Ex-11-10	Ex-17	MP-1	Ex-3	Ex-10-3	Ex-11-7	Ex-11-15	Ex-16	Ex-24	MP-4
D	Ex-5	Ex-11-3	Ex-11-11	Ex-18	MP-2	Ex-5	Ex-10-4	Ex-11-8	Ex-11-16	Ex-17	Ex-25	MP-5
E	Ex-6	Ex-11-4	Ex-11-12	Ex-19	MP-3	Ex-6	Ex-11-1	Ex-11-9	Ex-11-17	Ex-18	Ex-26	MP-6
F	Ex-7	Ex-11-5	Ex-12	Ex-20	MP-4	Ex-7	Ex-11-2	Ex-11-10	Ex-12	Ex-19	Ex-27-1	MP-7
G	Ex-8	Ex-11-6	Ex-13	Ex-21	MP-5	Ex-8	Ex-11-3	Ex-11-11	Ex-13	Ex-20	Ex-27-2	MP-8
H	Ex-9	Ex-11-7	Ex-14	Ex-22	MP-6	Ex-9	Ex-11-4	Ex-11-12	Ex-14-1	Ex-21	MP-1	MP-9

🟩 BRCA1 🟧 BRCA2 🟦 Multiplex nontemplate control

BRCA1 / BRCA2 gene resequencing

- Molecular diagnostics of mutations


Dispense
Human DNA + AmpliTaq GOLD® 360
Prespotted primer plate
PCR amplification
Add water to dilute and transfer to Forward and Reverse sequencing plates containing ExoSAP-IT®
ExoSAP-IT®
Incubate
Add Forward sequencing mix
Add Reverse sequencing mix
Fast cycle sequence reaction
Add BigDye XTerminator®
Add BigDye XTerminator®
Vortex for 15 minutes and spin down



Next Generation Sequencing –

Massively Parallel Sequencing of clonally amplified (or single) DNA molecules

-Process millions of sequence reads in parallel

-Library preparation

-Specific adaptor oligos

-Little volume DNA template

-Produce shorter read lengths (35-400 bp)

-100 Mb to several Gb nucleotid sequence determination

Pyrosequencing
chemiluminescent detection of pyrophosphate

Enzymes:
Klenow fragment
ATP sulfurylase
Luciferase
Apyrase

Reagents:
Adenozin-phosphosulphate (APS)
D-luciferin
DNA template
Primers
dNTPs one by one

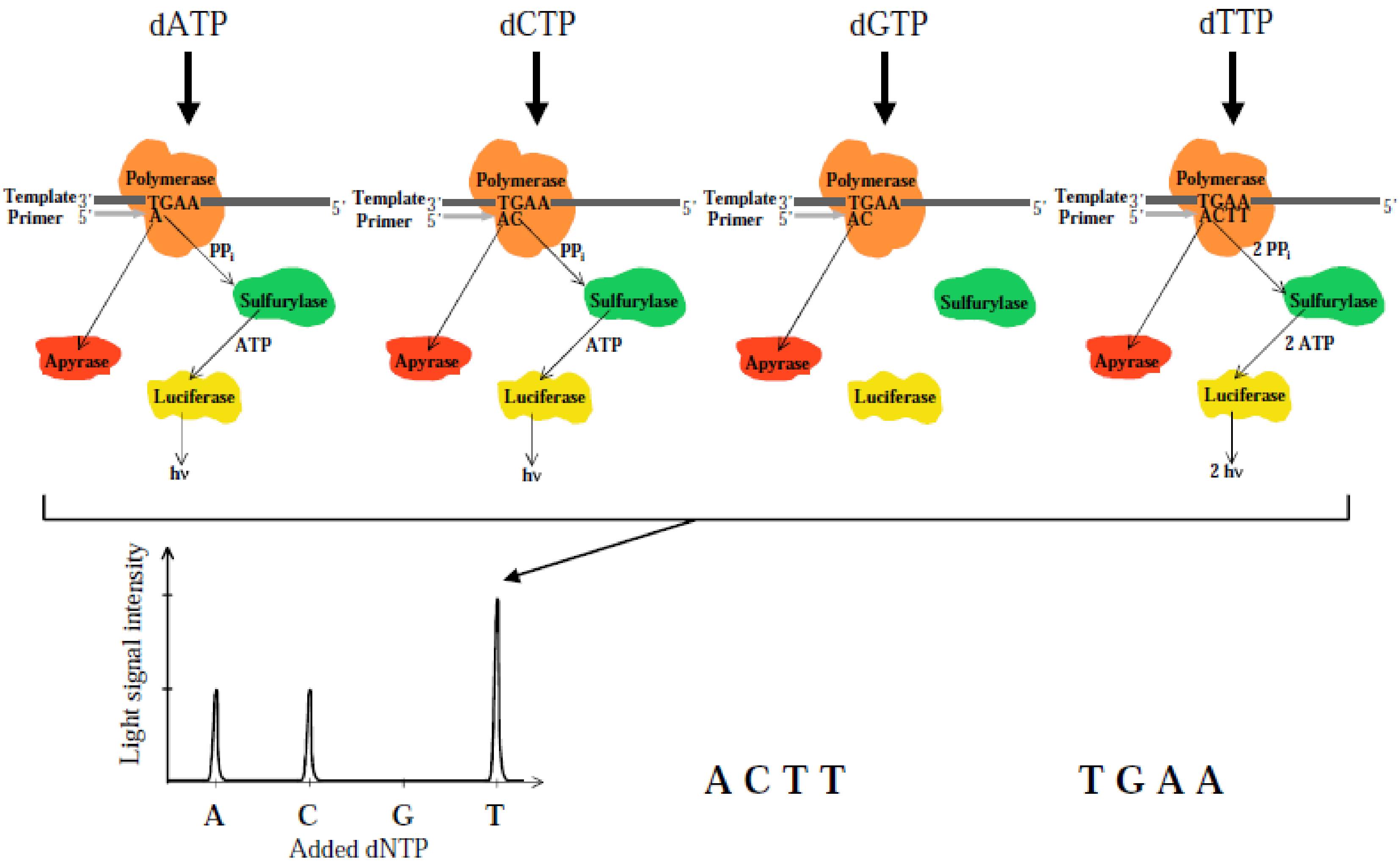
$$(\mathrm{NA})_n + \mathrm{Nucleotide} \xrightarrow{\text{Polymerase}} (\mathrm{NA})_{n+1} + \mathrm{PPi}$$

$$\mathrm{PPi} + \mathrm{APS} \xrightarrow{\text{ATP sulfurylase}} \mathrm{ATP} + \mathrm{SO_4^{2-}}$$

$$\mathrm{ATP} + \mathrm{Luciferin} + \mathrm{O_2} \xrightarrow{\text{Luciferase}} \mathrm{AMP} + \mathrm{PPi} + \mathrm{Oxyluciferin} + \mathrm{CO_2} + \mathrm{Light}$$

Ronaghi M (2001) Genom Res

Pyrosequencing

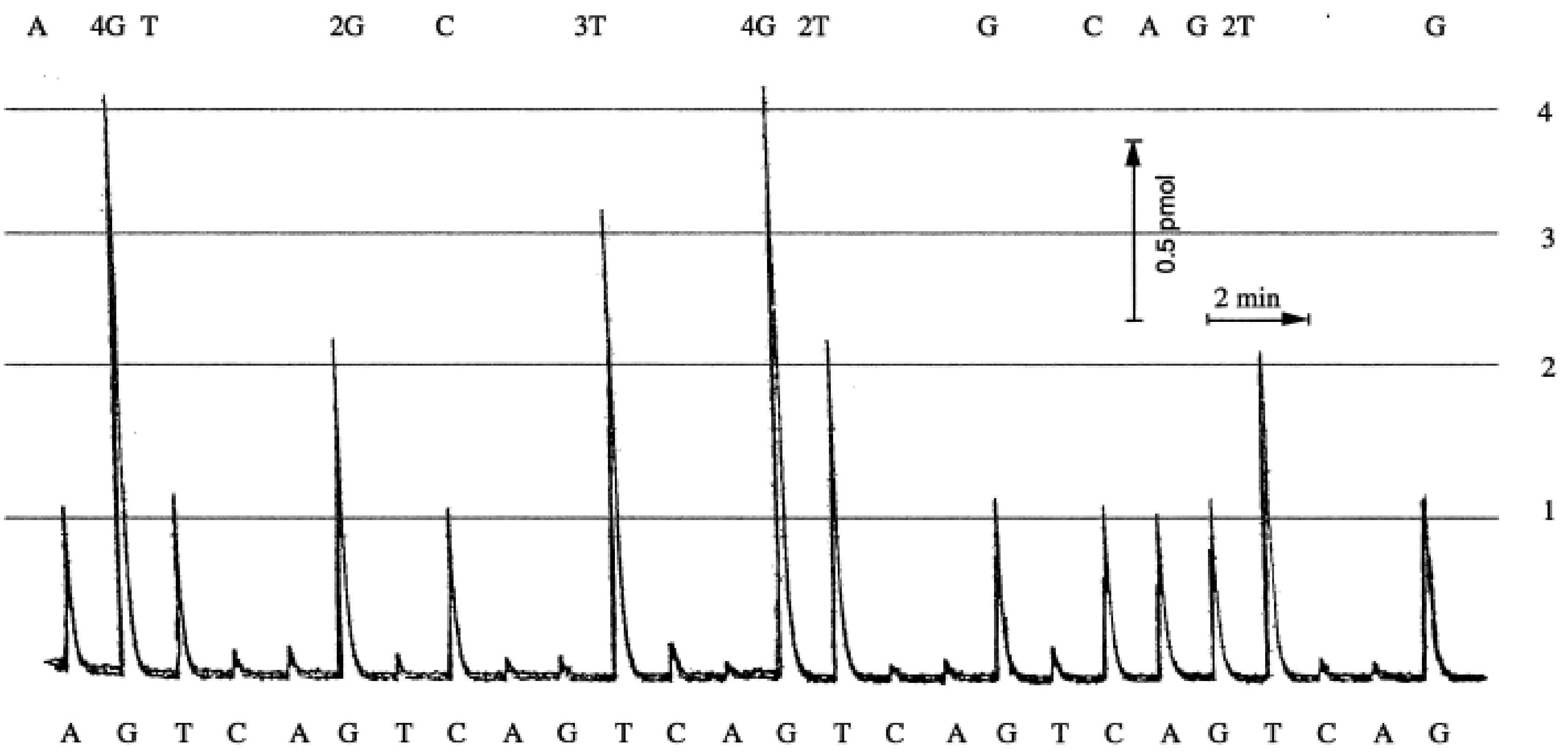


A C T T

T G A A

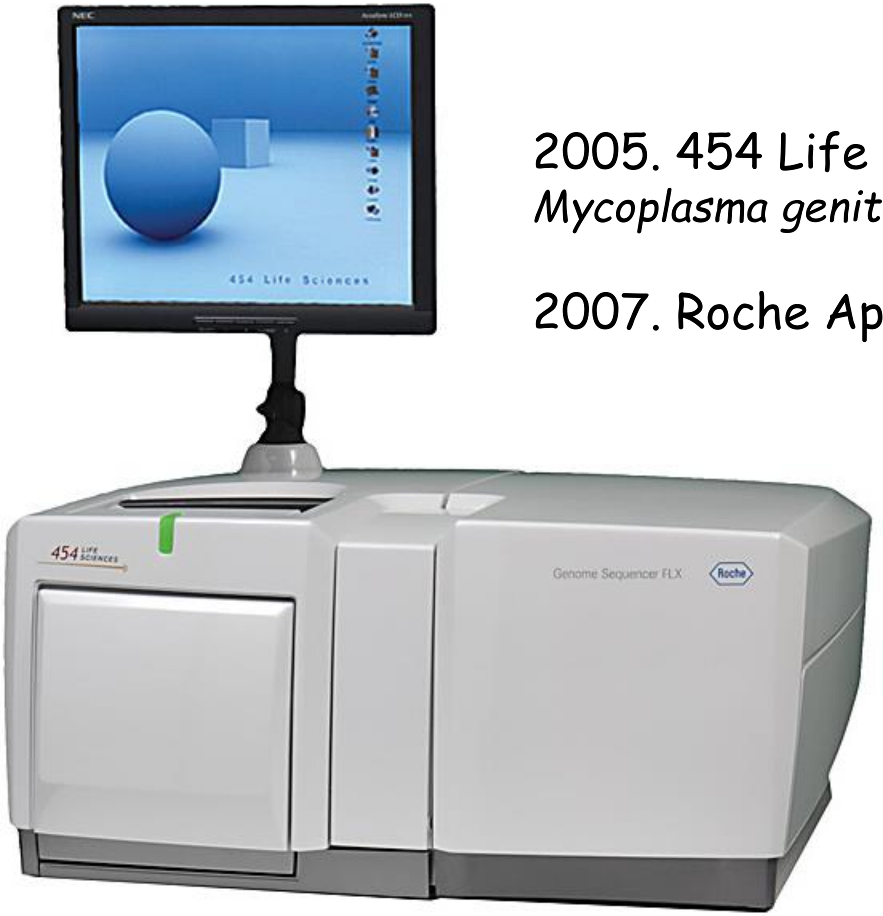
Pyrogram → Sequence of synthesised DNA → Sequence of template DNA

Ahmadian A (2006) Clin Chim Acta



Ronaghi M (2001) Genom Res

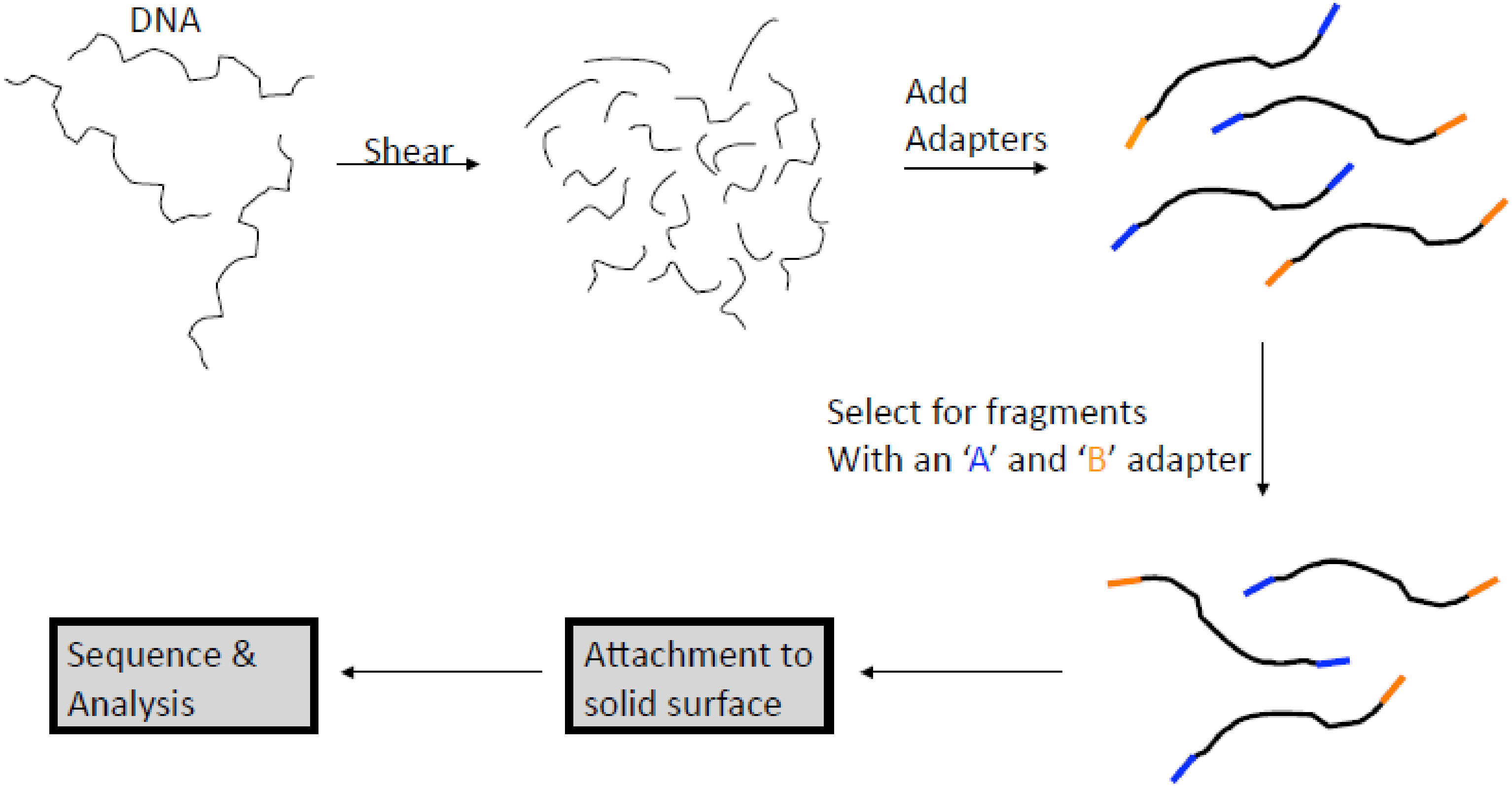
Roche/454 sequencing technology



2005. 454 Life Sciences developed (GS 20)
Mycoplasma genitalia 580 kb genome, 99.96% accuracy

2007. Roche Applied Science (GS FLX series)

DNA preparation



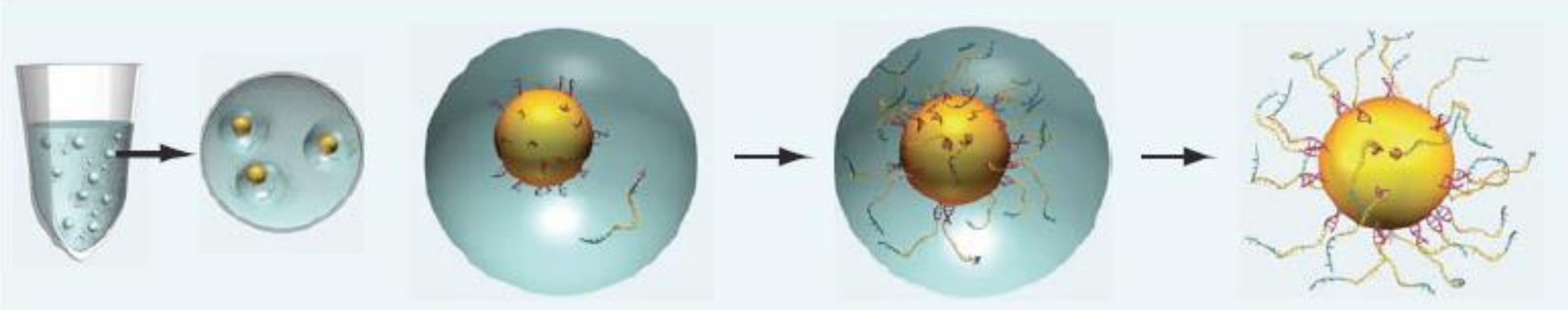
Shearing DNA (some several 100 bps long)
End-repair
Adapter adding

Roche/454 sequencing technology
Clonal amplification

Emulsion PCR

Microreactors
Water in Oil emulsion

Several million copies of a fragment



Anneal sstDNA to an excess of DNA Capture Beads

Emulsify beads and PCR reagents in water-in-oil microreactors

Clonal amplification occurs inside microreactors

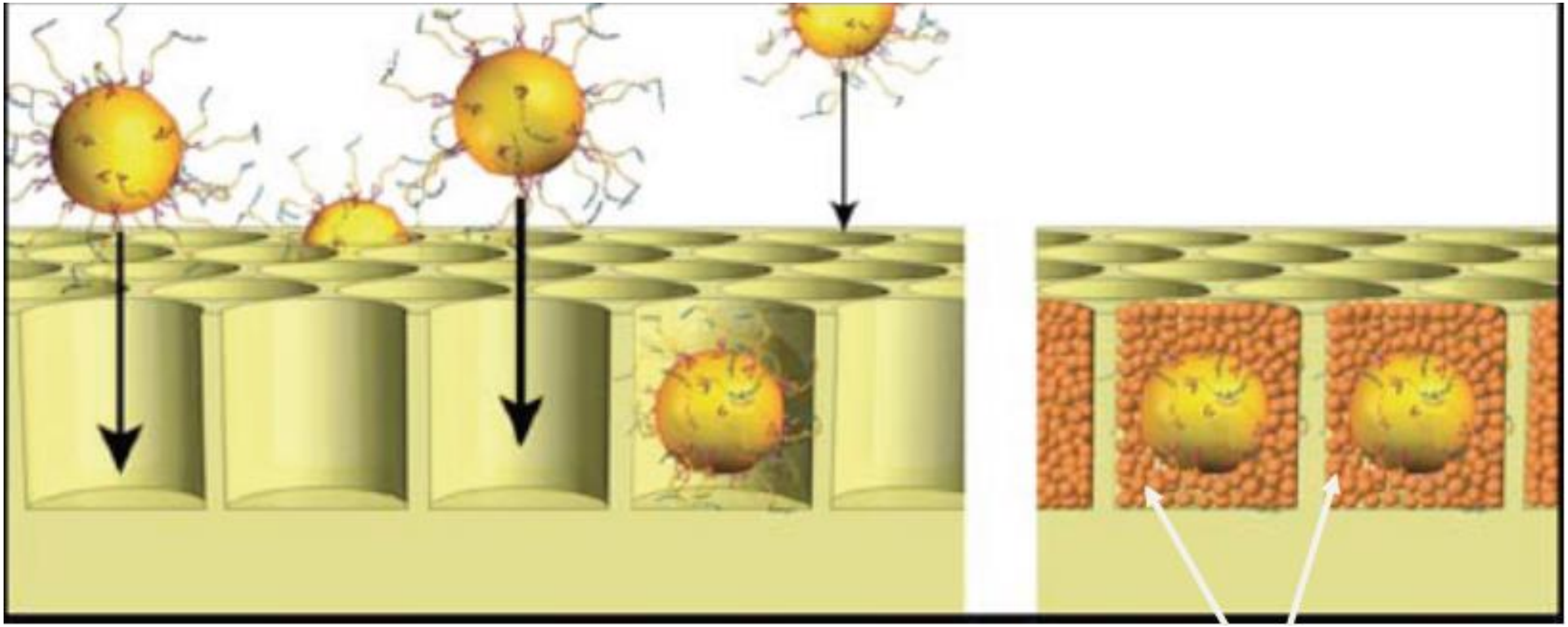
Break microreactors, enrich for DNA-positive beads

Each bubble in the emulsion will potentially contain a different fragment.

Roche/454 sequencing technology

Picotiter well plate mounting

3,4*10^6 wells
Sequencing reaction in picoliter volumes

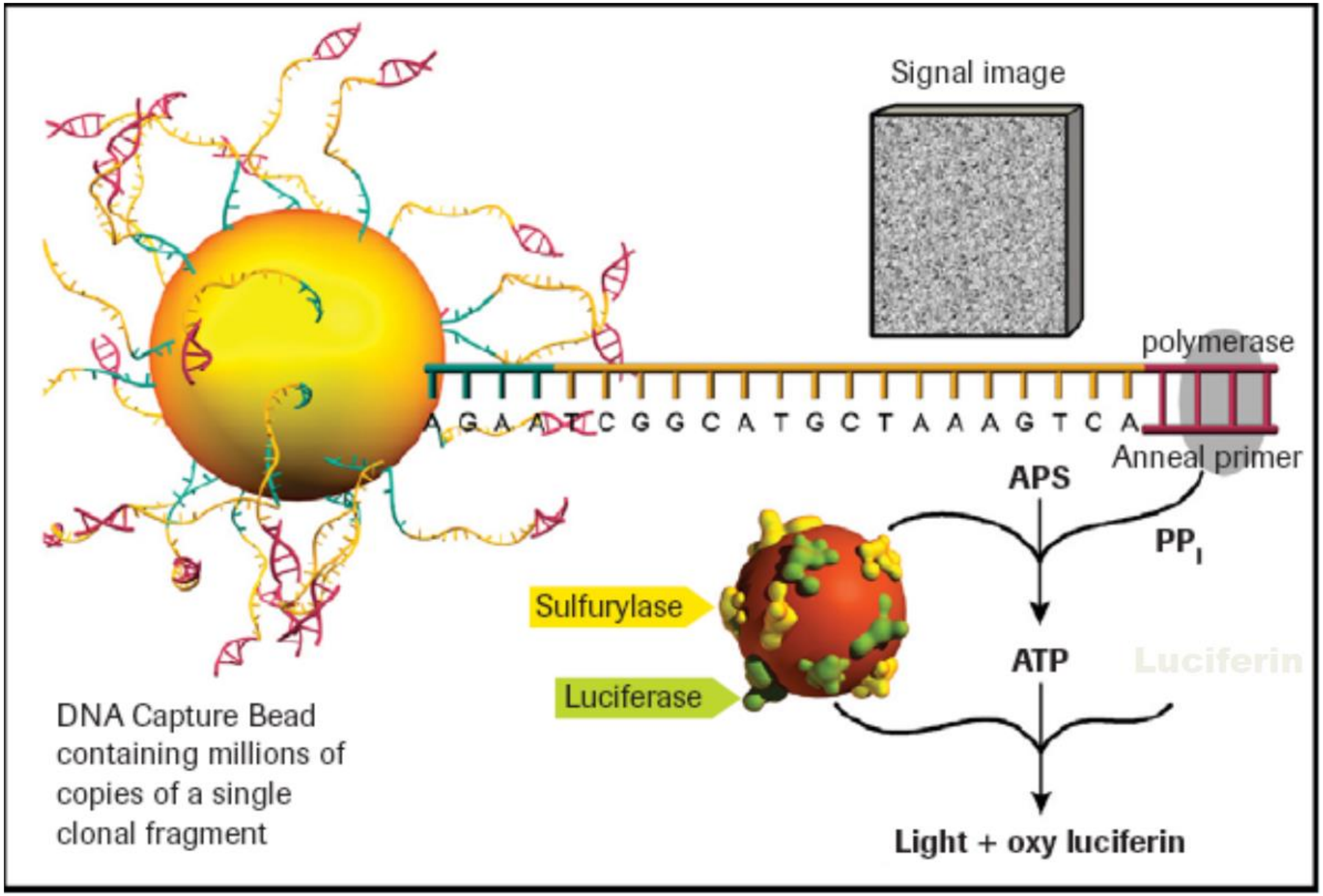


Instead of 96 reads/run, there are hundreds of thousands.
Packing beads and enzyme beads

Ansorge (2009) New biotechnology

Roche/454 sequencing technology

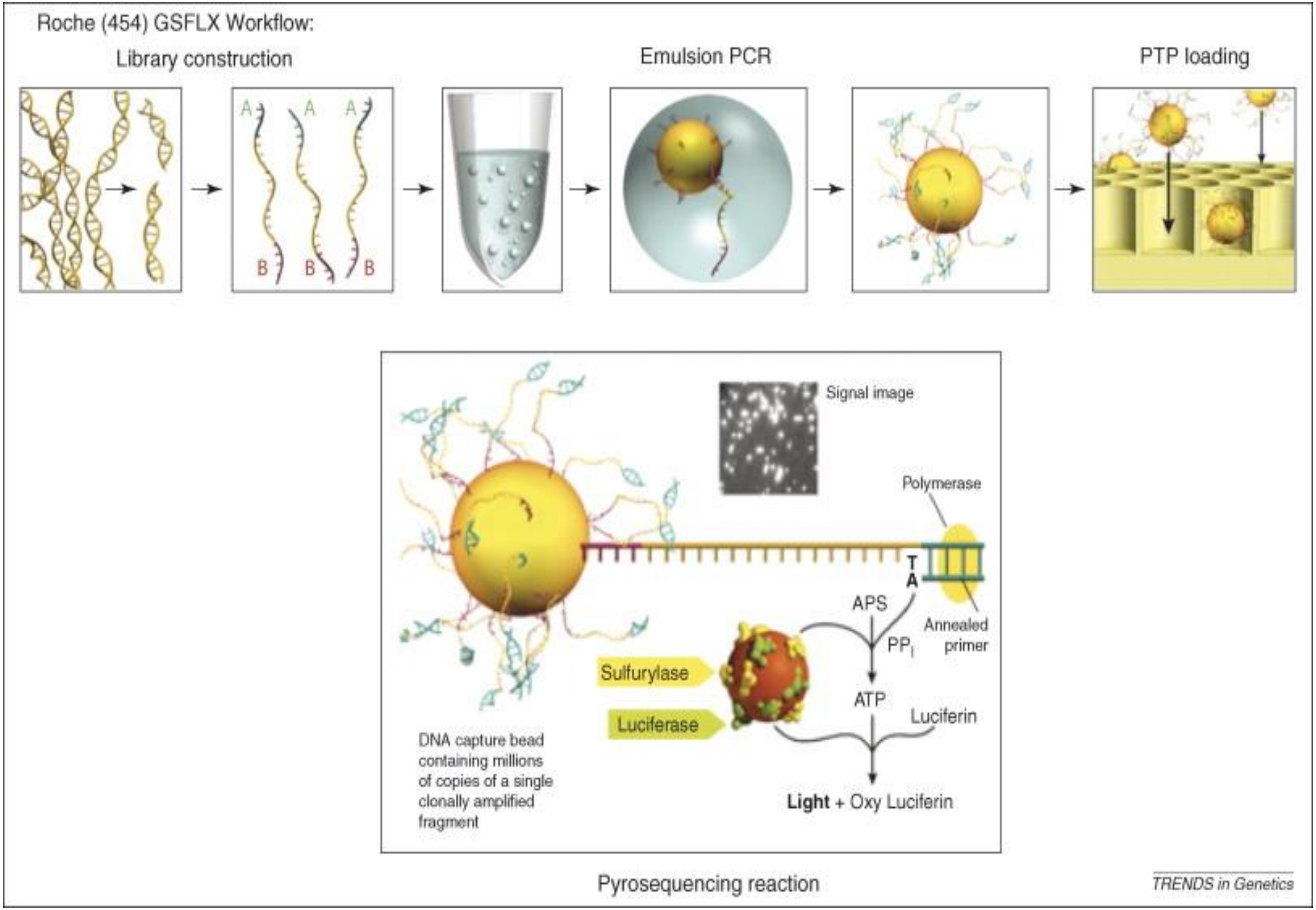
Sequencing by pyrosequencing



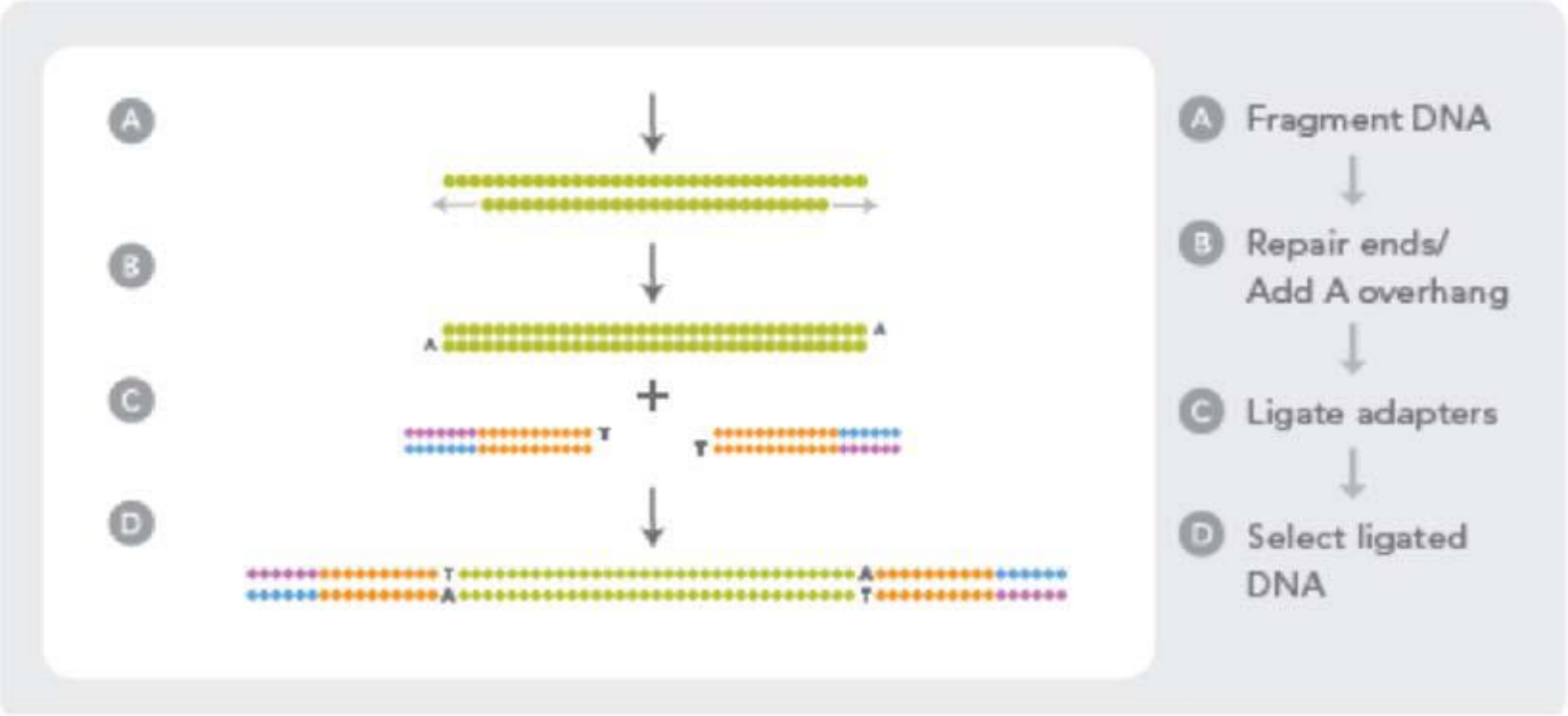
Signal image
polymerase
A G A A T C G G C A T G C T A A A G T C A
Anneal primer
APS
PP$_i$
Sulfurylase
Luciferase
ATP
Luciferin
Light + oxy luciferin
DNA Capture Bead containing millions of copies of a single clonal fragment

Ansorge (2009) New biotechnology

Next Generation Sequencing – Roche 454 platform



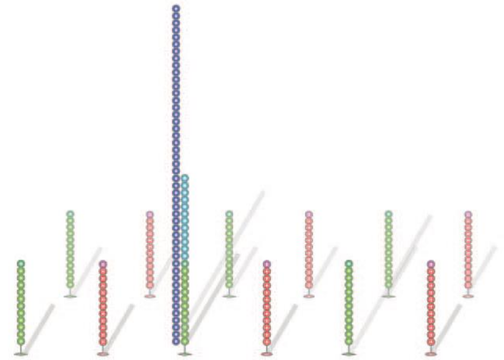
Illumina/Solexa sequencing DNA preparation



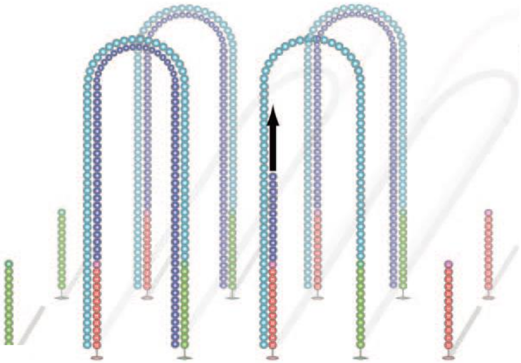
A) DNA shearing to fragments (some 100 bps long)
B) End-repair, Add A overhang
C) Adapters ligating (T overhang)

Ansorge (2009) New biotechnology

Illumina/Solexa sequencing
Clonal amplification

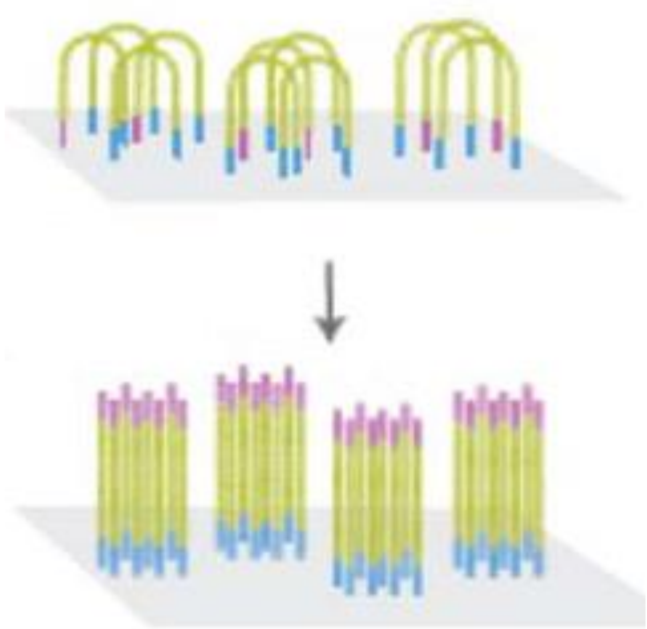


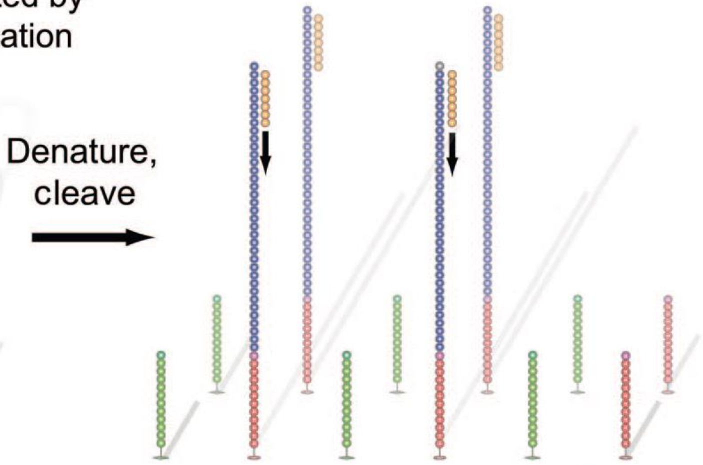
Adaptor modified DNA strand hybridized to oligonucleotide anchor



Cluster generated by bridge amplification

PCR with anchored primers
Bridge amplification





Sequencing of forward strands

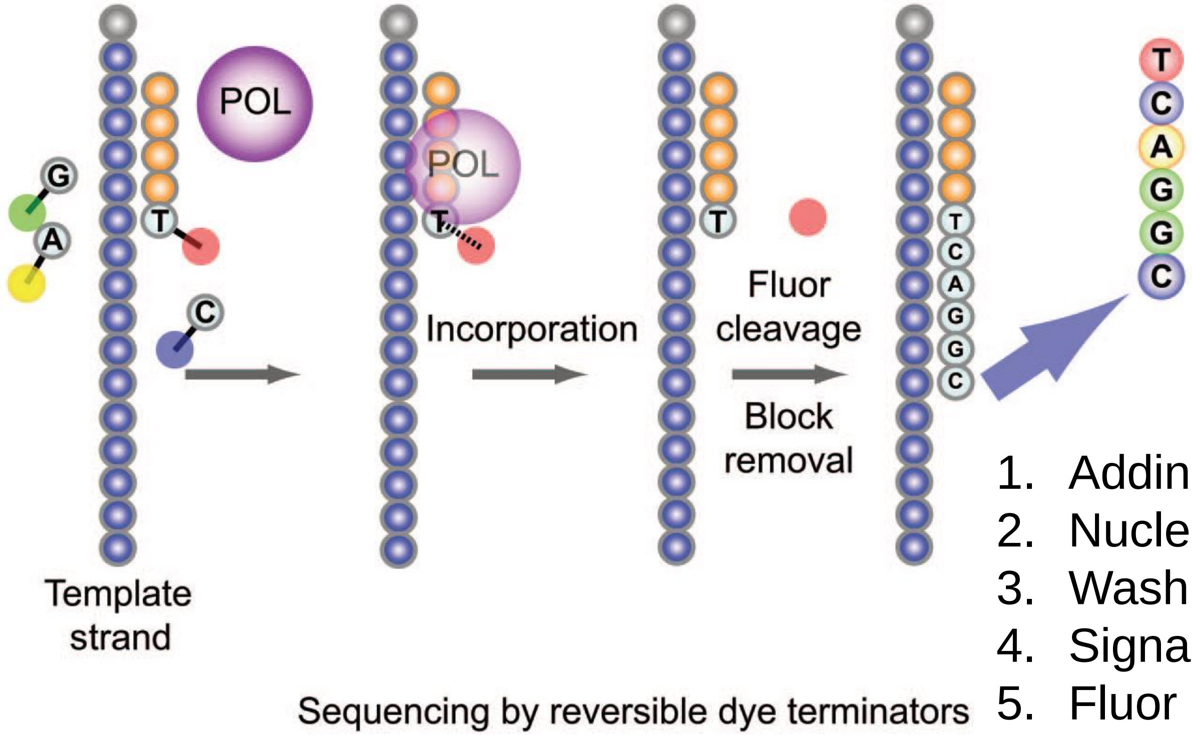
Clone clusters
1000 copies min.

50×10^6 clusters

Voelkerding (2009) Clin Chem

Illumina/Solexa sequencing

Sequencing by DNA synthesis



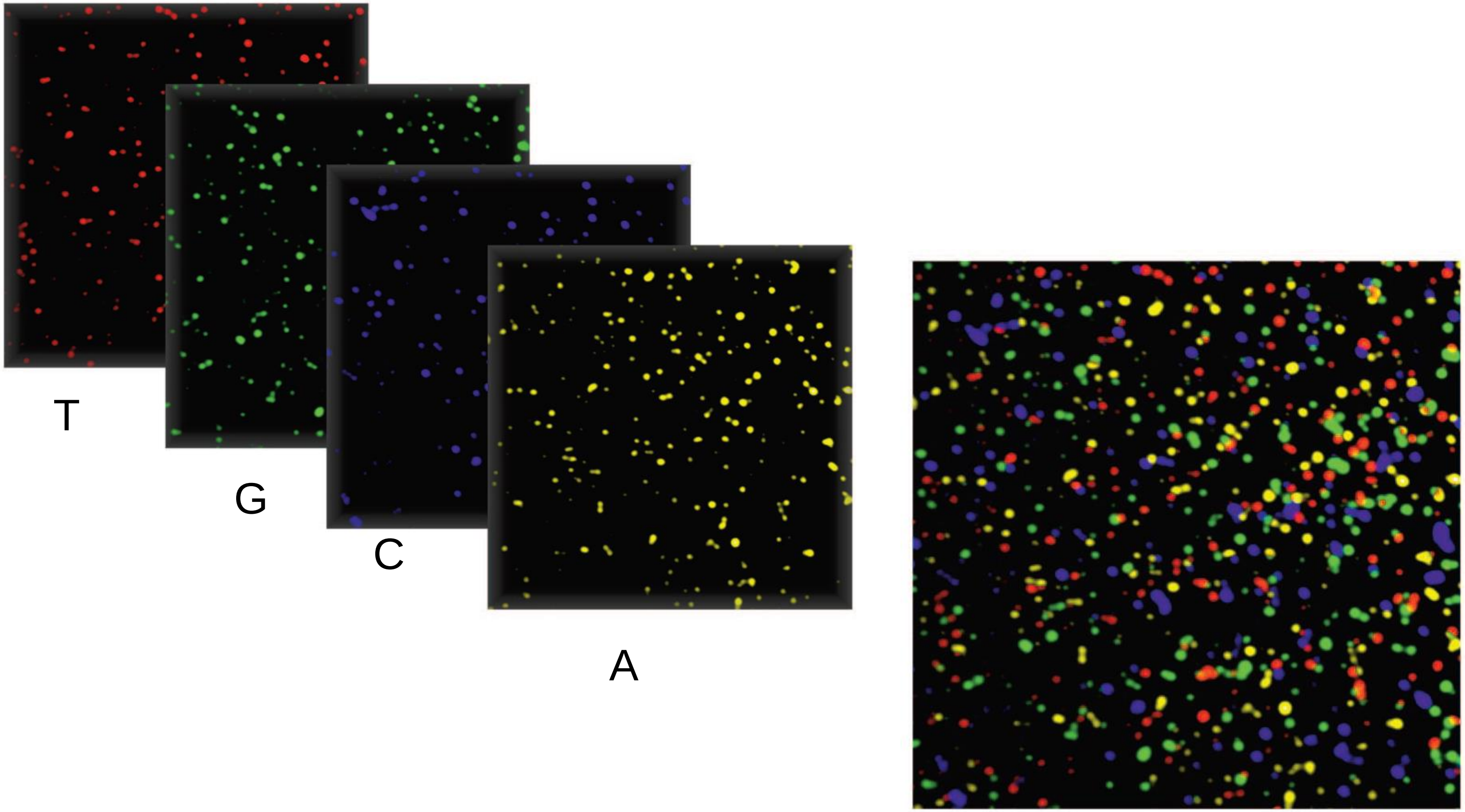
1. Adding reagents
2. Nucleotide incorporation
3. Washing
4. Signal detection
5. Fluor cleavage and block removal

Fluorescently labled reversible chain terminators
Each 4 nucleotides into the reaction

Voelkerding (2009) Clin Chem

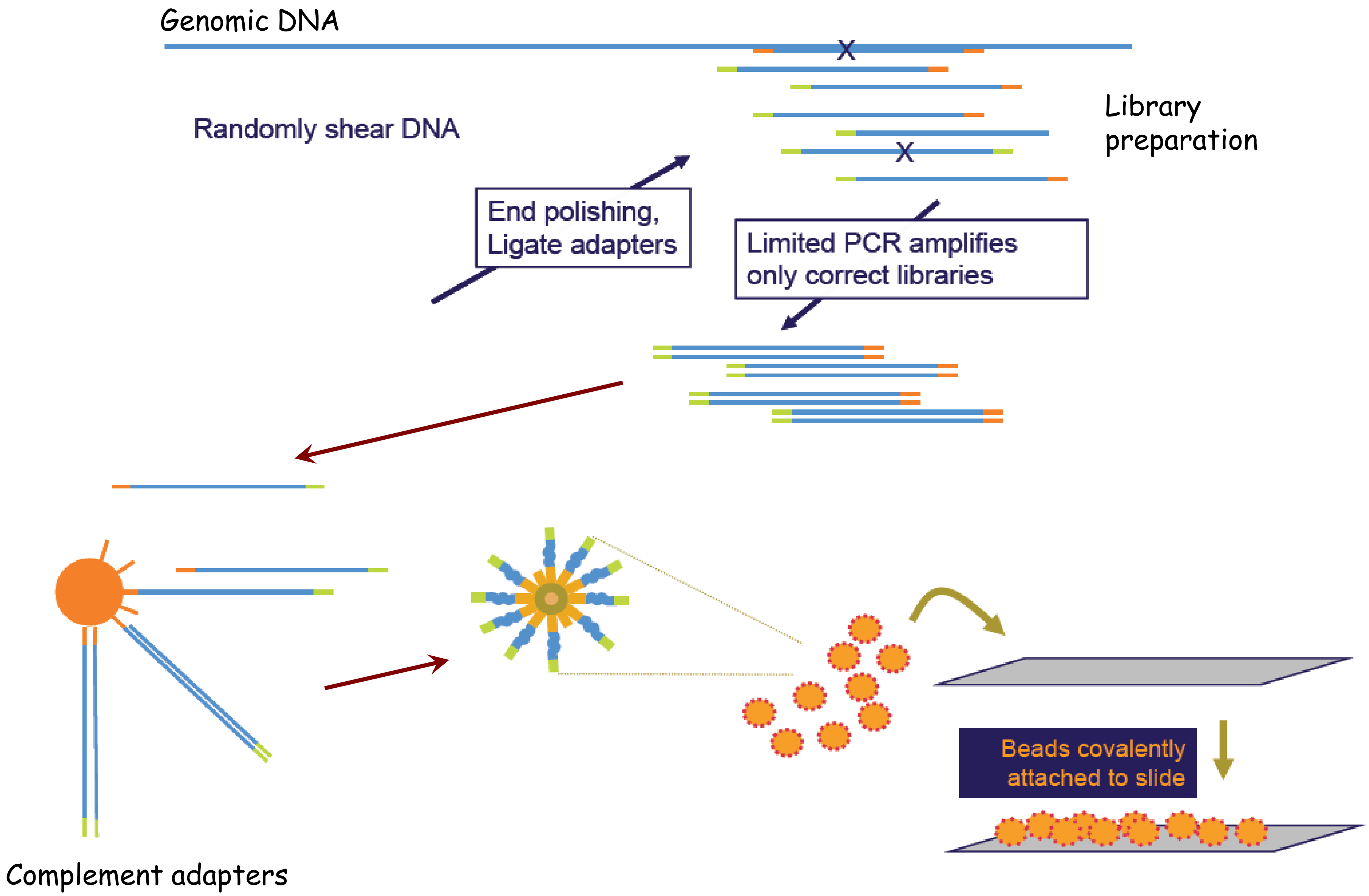
Illumina/Solexa sequencing

Fluorescent signal detection



Voelkerding (2009) Clin Chem

SOLID: Sequencing by Oligo Ligation and Detection



Genomic DNA

Randomly shear DNA

End polishing, Ligate adapters

Limited PCR amplifies only correct libraries

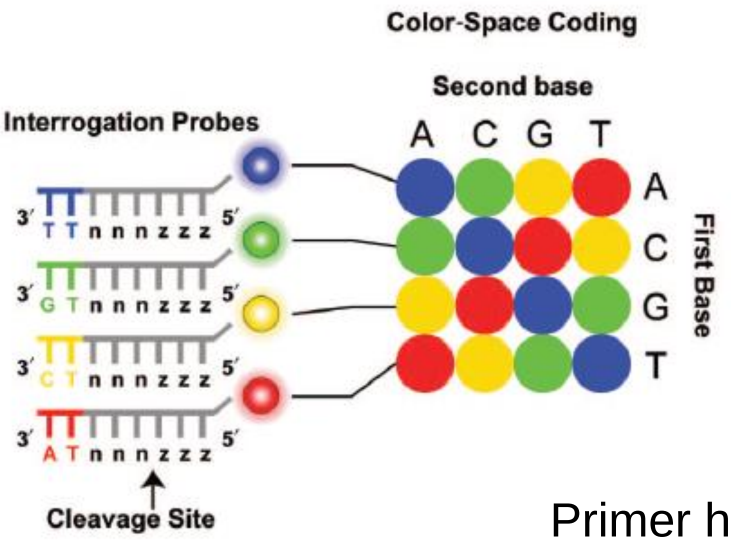
Library preparation

Beads covalently attached to slide

Complement adapters

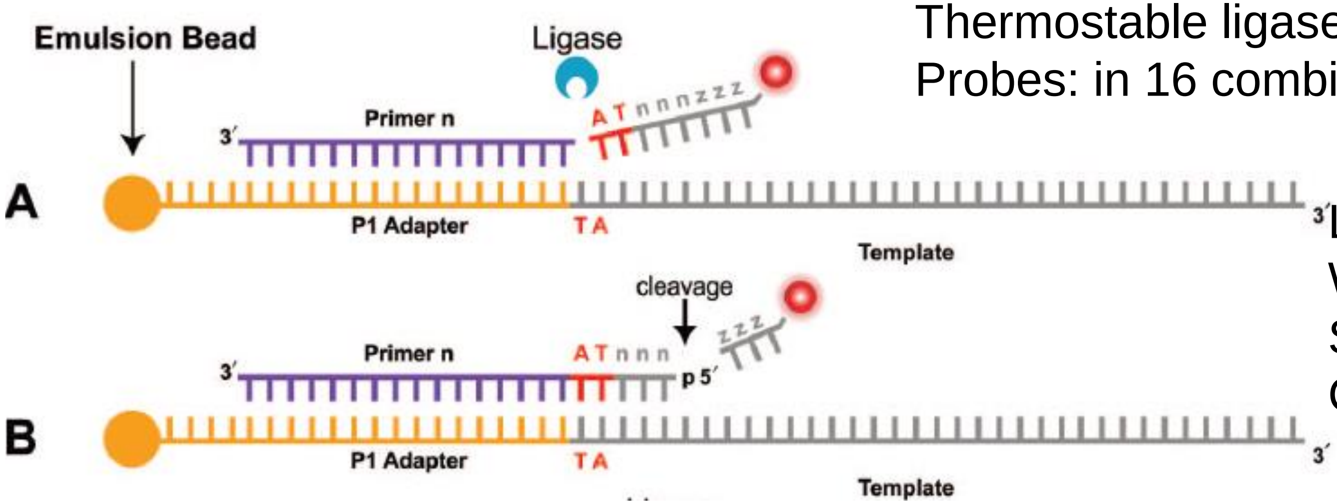
Applied Biosystems - SOLiD

Sequencing by probe ligation



Probes
Octamer
2 probe specific bases
3 degenerated bases
3 universal
Fluorescent marker

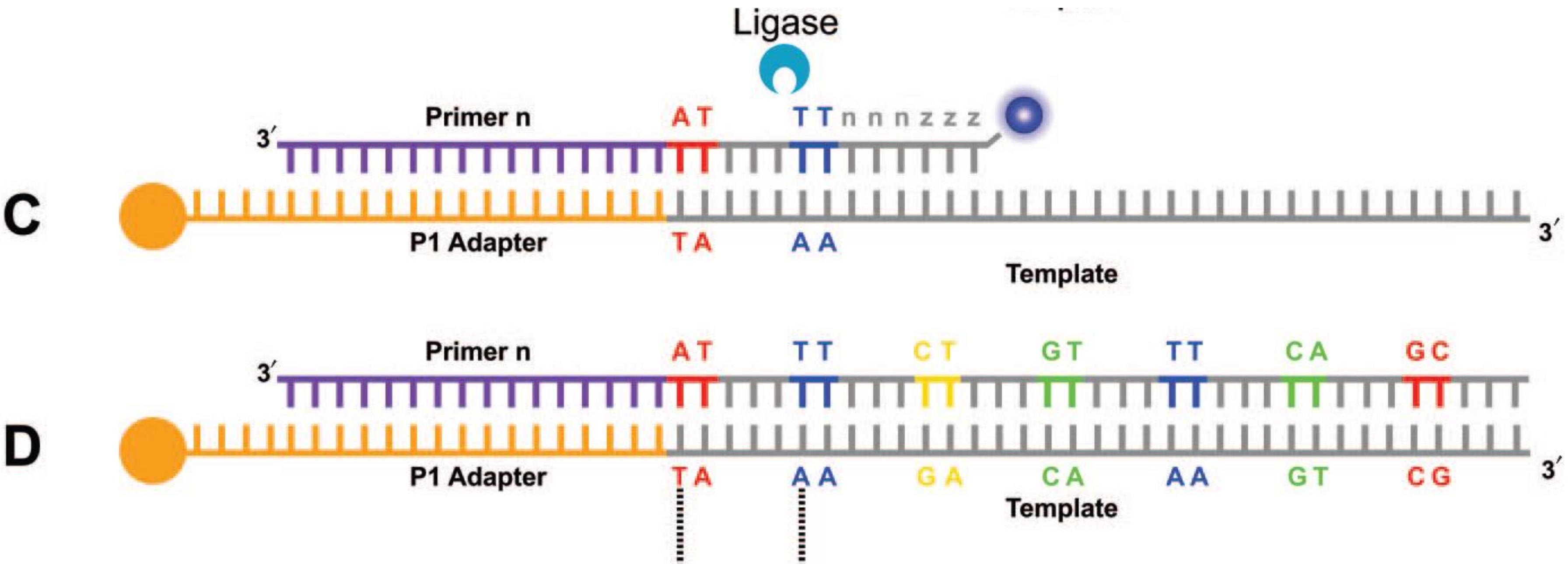
Primer hybridisation to adapter sequence
Thermostable ligase
Probes: in 16 combination



Ligation
Washing
Signal detection
Cleavage – 3 nukleotid

Voelkerding (2009) Clin Chem

Applied Biosystems - SOLiD

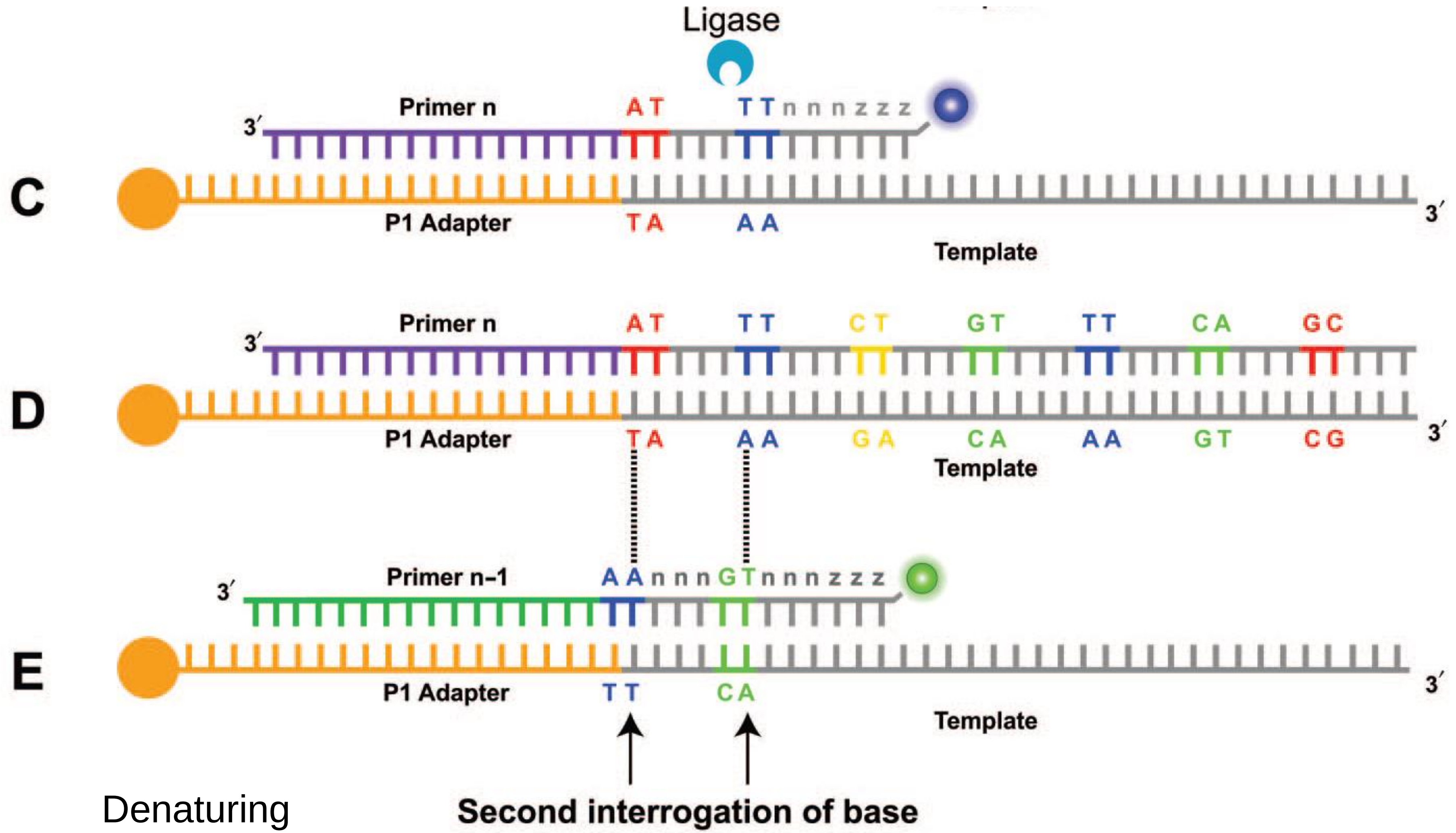


Another probe ligation

Cycle performs 7 times

Voelkerding (2009) Clin Chem

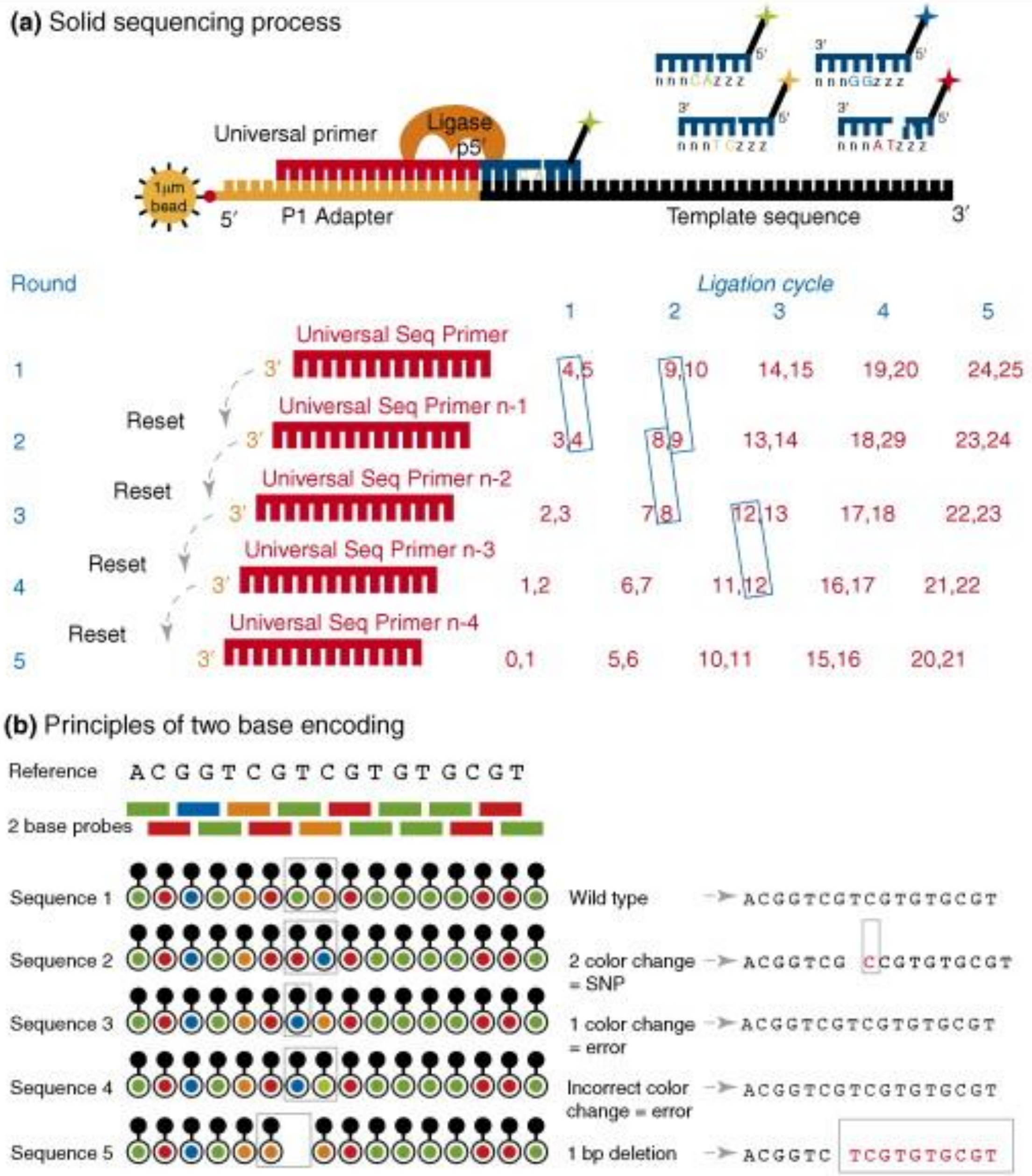
Applied Biosystems - SOLiD



Denaturing
New round starts with n-1 adapter primer
Altogether 5 rounds

Each nucleotide are queried 2*

Voelkerding (2009) Clin Chem

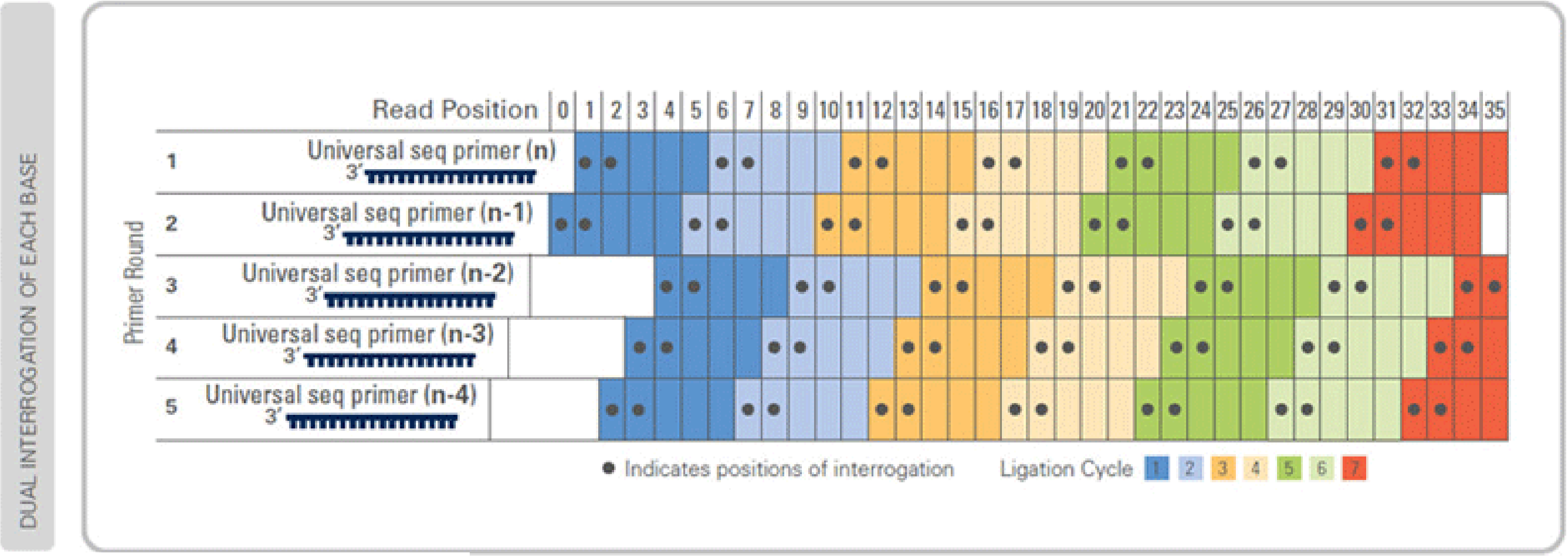


TRENDS in Genetics

Next Generation DNA Sequencing: SOLID

- Kémiai hasítás, amplifikálás és ligálás

Accuracy: 99.99 %


DUAL INTERROGATION OF EACH BASE
Read Position
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
Primer Round
1 Universal seq primer (n) 3'
2 Universal seq primer (n-1) 3'
3 Universal seq primer (n-2) 3'
4 Universal seq primer (n-3) 3'
5 Universal seq primer (n-4) 3'
● Indicates positions of interrogation
Ligation Cycle 1 2 3 4 5 6 7

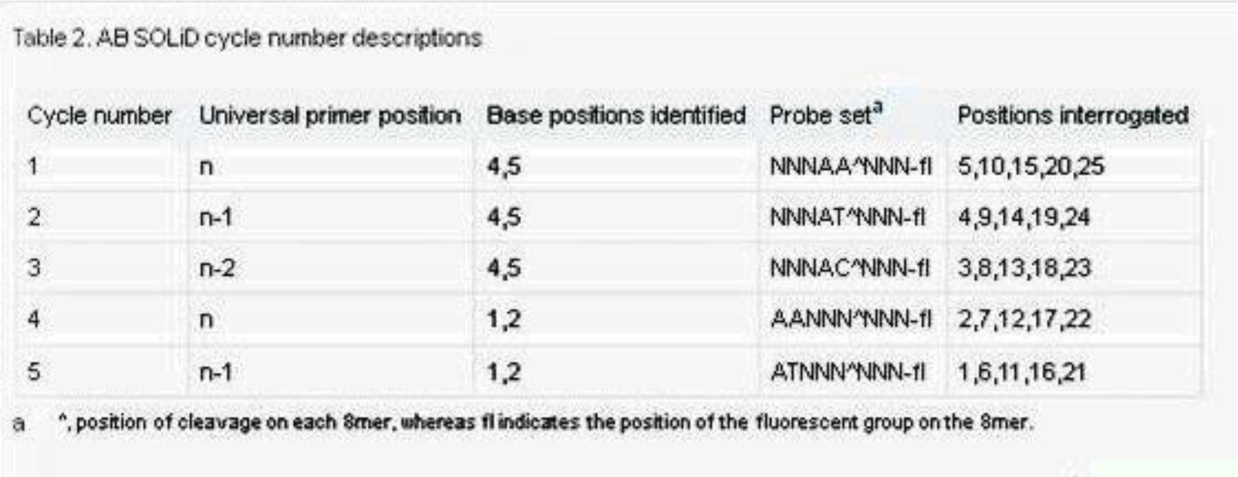


Table 2. AB SOLiD cycle number descriptions

Cycle number	Universal primer position	Base positions identified	Probe set[a]	Positions interrogated
1	n	4,5	NNNAA^NNN-fl	5,10,15,20,25
2	n-1	4,5	NNNAT^NNN-fl	4,9,14,19,24
3	n-2	4,5	NNNAC^NNN-fl	3,8,13,18,23
4	n	1,2	AANNN^NNN-fl	2,7,12,17,22
5	n-1	1,2	ATNNN^NNN-fl	1,6,11,16,21

[a] ^, position of cleavage on each 8mer, whereas fl indicates the position of the fluorescent group on the 8mer.

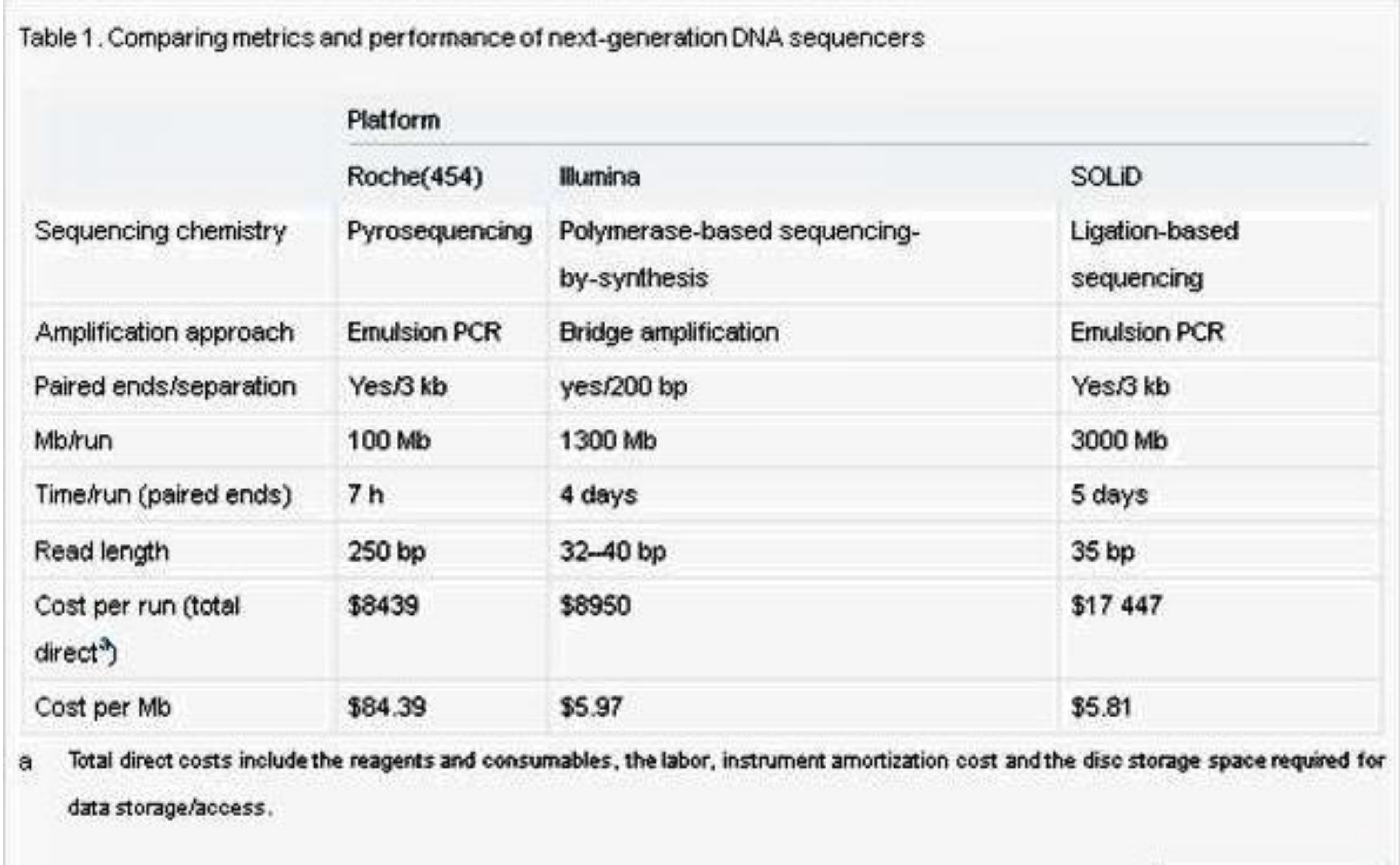
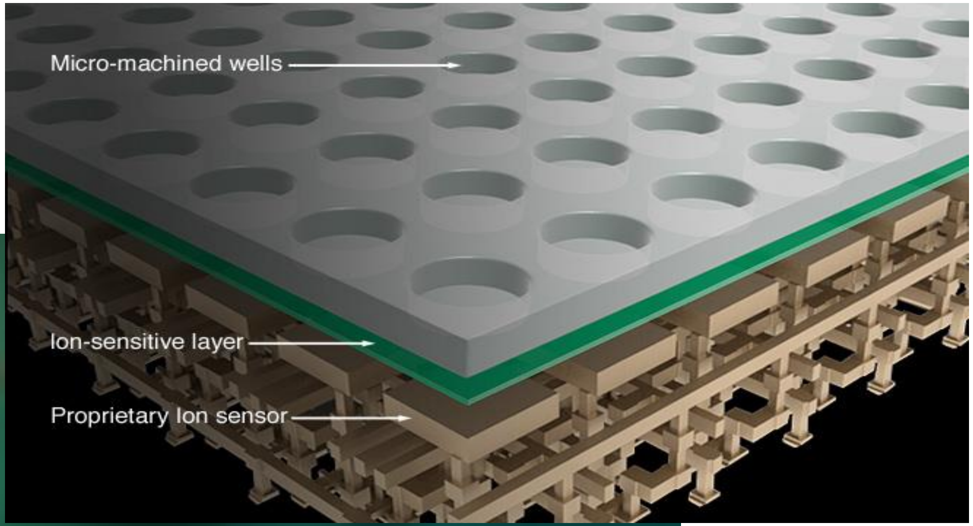


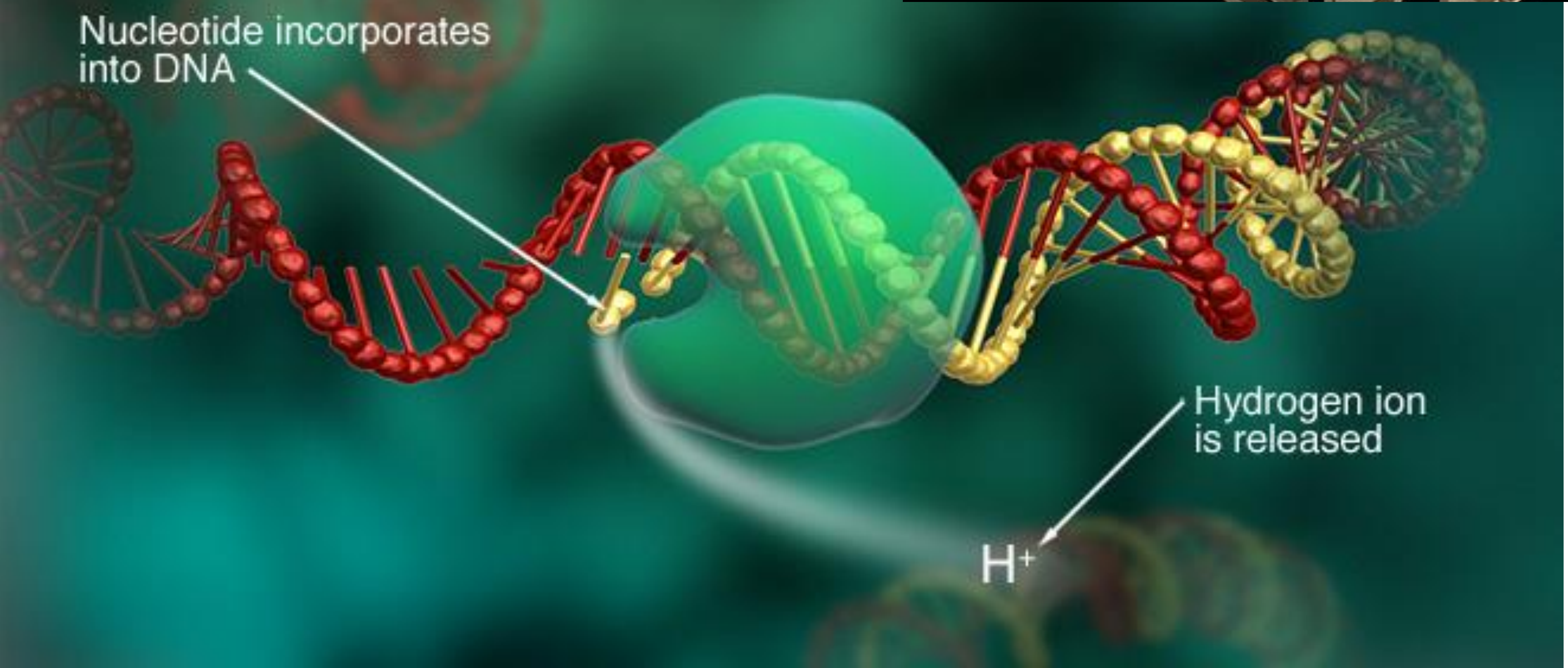
Table 1 . Comparing metrics and performance of next-generation DNA sequencers

	Platform		
	Roche(454)	Illumina	SOLiD
Sequencing chemistry	Pyrosequencing	Polymerase-based sequencing-by-synthesis	Ligation-based sequencing
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR
Paired ends/separation	Yes/3 kb	yes/200 bp	Yes/3 kb
Mb/run	100 Mb	1300 Mb	3000 Mb
Time/run (paired ends)	7 h	4 days	5 days
Read length	250 bp	32–40 bp	35 bp
Cost per run (total direct[a])	$8439	$8950	$17 447
Cost per Mb	$84.39	$5.97	$5.81

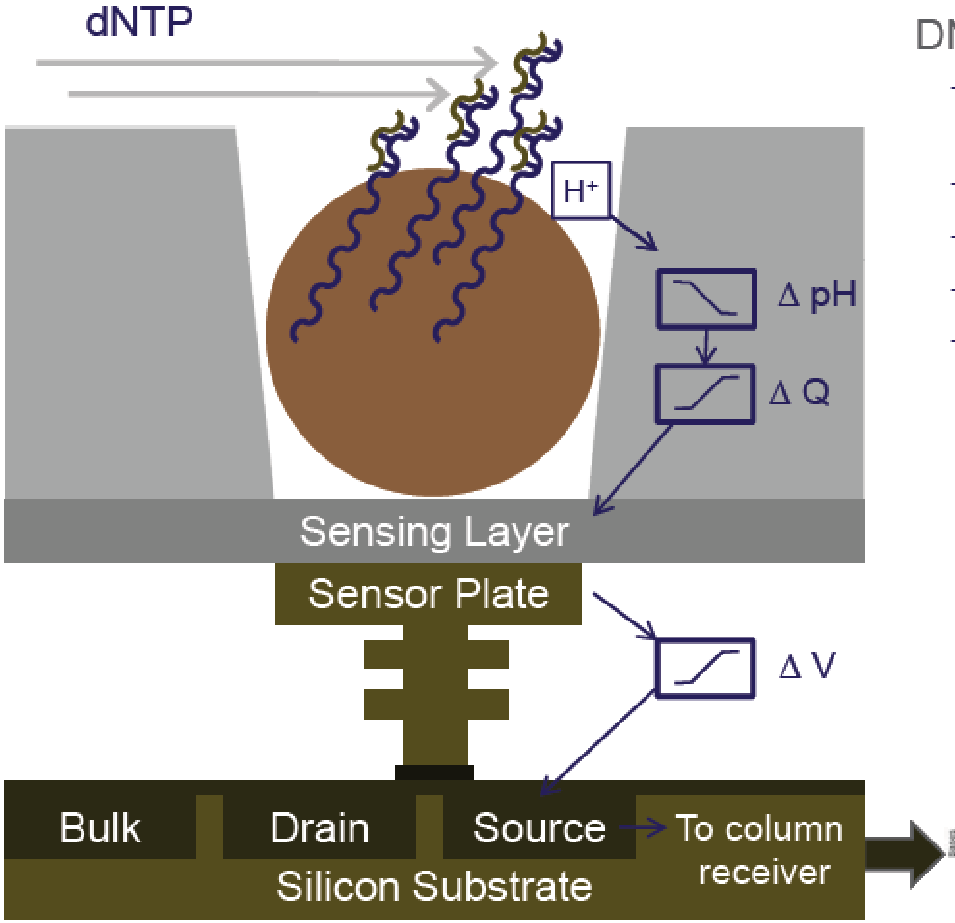
[a] Total direct costs include the reagents and consumables, the labor, instrument amortization cost and the disc storage space required for data storage/access.

Ion semiconductor DNA sequencing





Ion semiconductor DNA sequencing: Personal Genome Machine

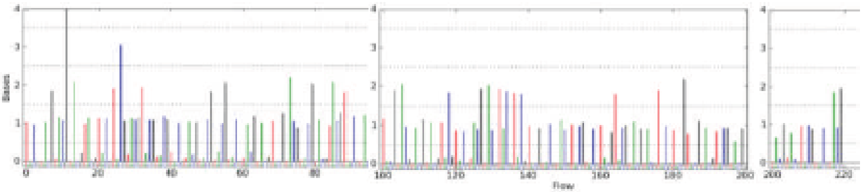


DNA → Ions → Sequence

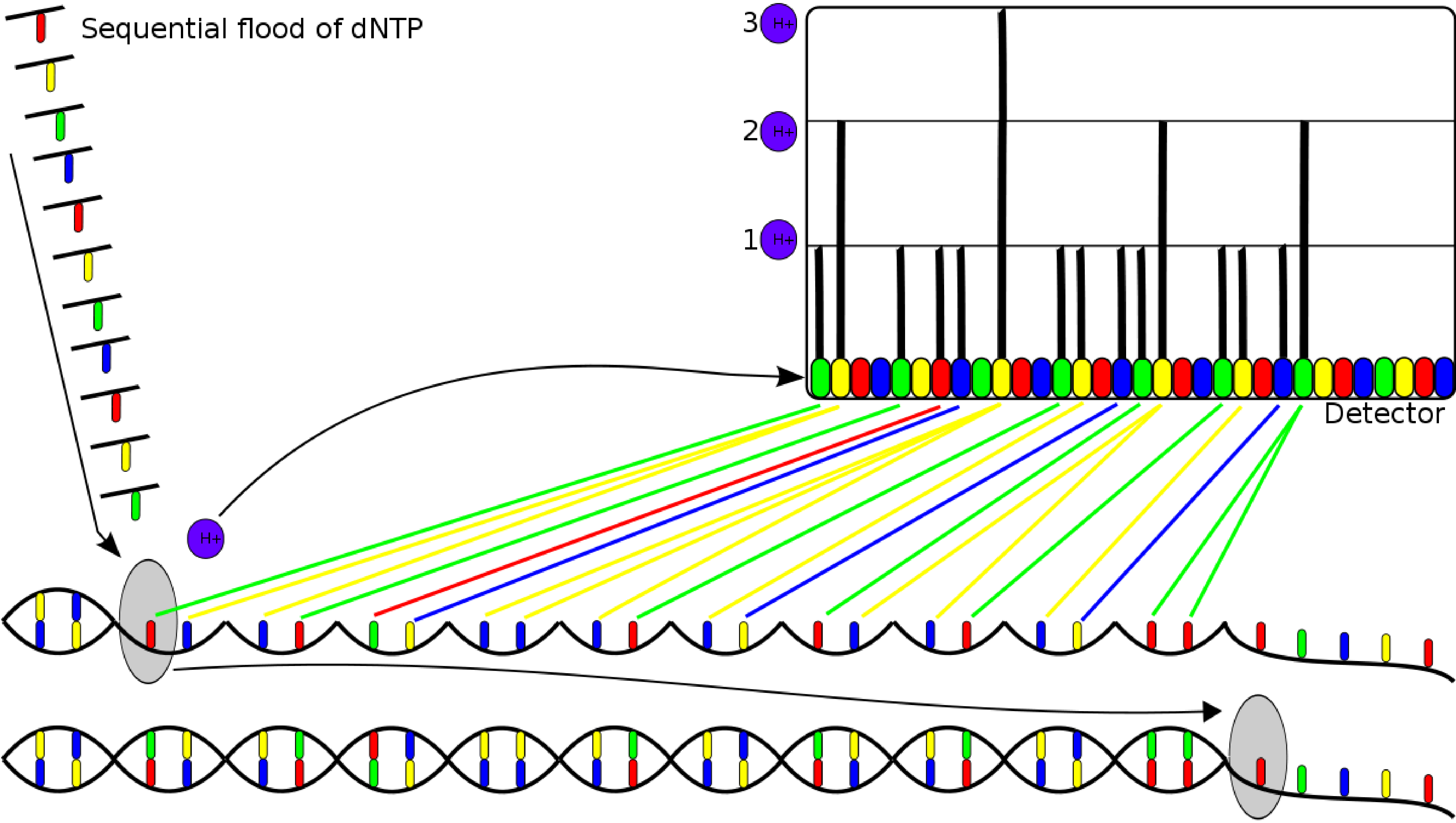
- Nucleotides flow sequentially over Ion semiconductor chip
- One sensor per well per sequencing reaction
- Direct detection of natural DNA extension
- Millions of sequencing reactions per chip
- Fast cycle time, real time detection

No PCR reaction, light emission, CCD camera etc.

Instead pH measures in microfluids



Ion semiconductor DNA sequencing



Ion semiconductor DNA sequencing: Ion Torrent

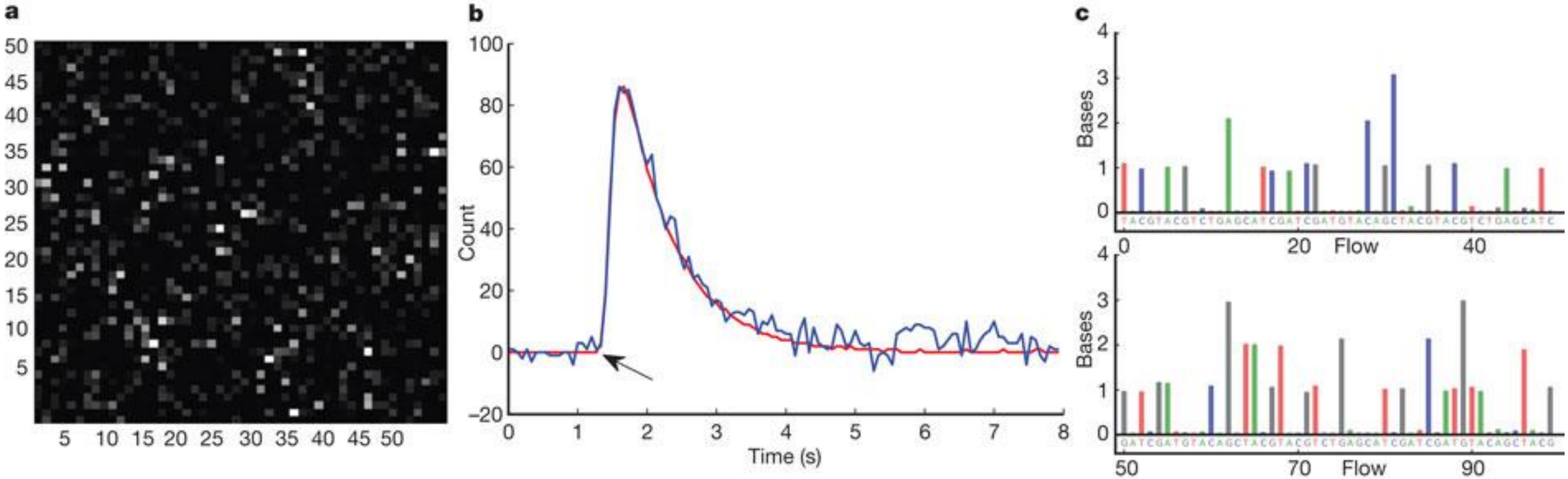


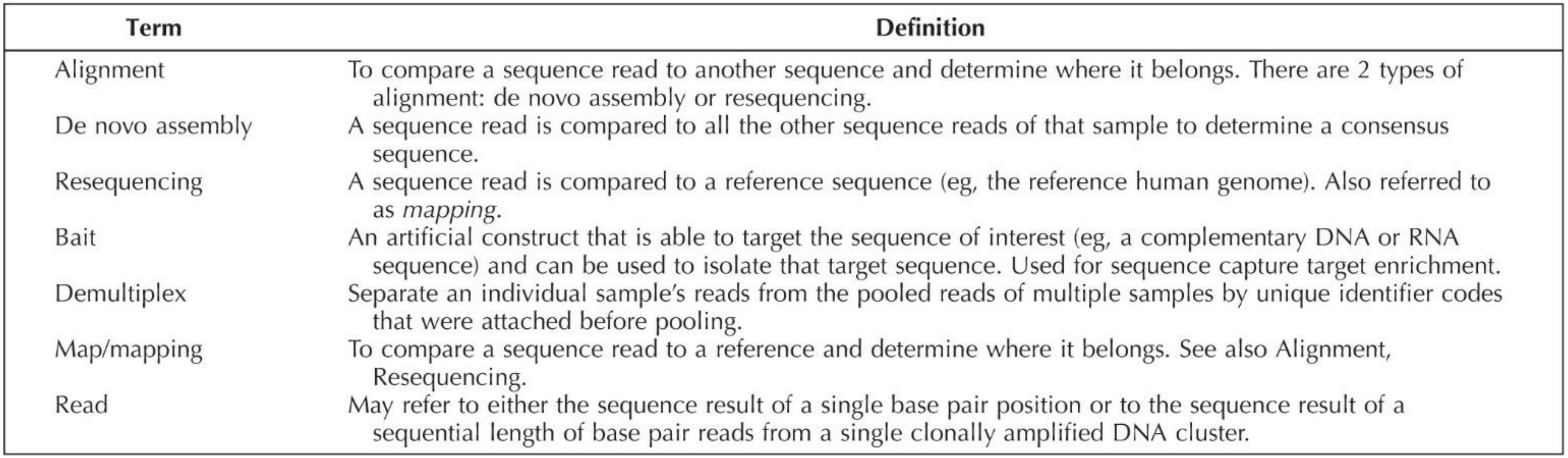
Table 1 | *Vibrio fisheri*, *E. coli*, *Rhodopseuomanas palustris* and *Homo sapiens*

	V. fisheri	*R. palustris*	*E. coli*	*E. coli*	*E. coli*	*H. sapiens*
GC content	38%	65%	51%	51%	51%	41%
Genome size	4.2 Mb	5.5 Mb	4.7 Mb	4.7 Mb	4.7 Mb	2.9 Gb
Number of runs x ion chip size	1 × 1.2 M	1 × 1.2 M	1 × 1.2 M	1 × 6.1 M	1 × 11 M	1,601 × 1.2 M 267 × 6.1 M 28 × 11.1 M
Fold coverage	6.2-fold	6.9-fold	11.3-fold	36.2-fold	58.4-fold	10.6-fold
Coverage	96.80%	99.64%	99.99%	100.00%	100.00%	99.21%
Reads ≥21 bases	261,313	444,750	507,198	1,852,931	2,594,031	366,623,578
Reads ≥50 bases	233,049	399,360	487,420	1,698,852	2,343,880	306,042,650
Reads ≥100 bases	156,391	160,726	400,743	1,012,918	1,779,237	139,624,090
Mapped bases	26.0 Mb	37.8 Mb	47.6 Mb	169.6 Mb	273.9 Mb	30.2 Gb

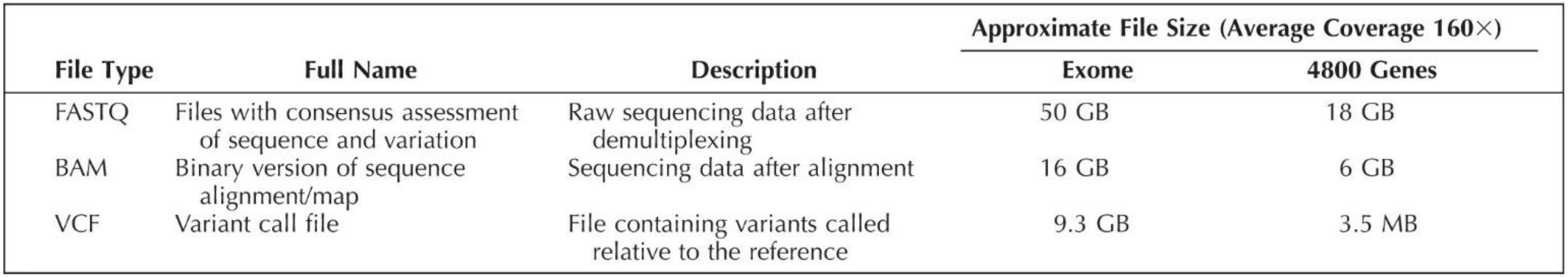
Coverage shows percentage of genome covered based on one or more reads mapping to each base of the reference genome. Reads align with 98% or greater accuracy.

JM Rothberg *et al. Nature* **475**, 348-352 (2011) doi:10.1038/nature10242

Terms and definitions used in genome sequencing

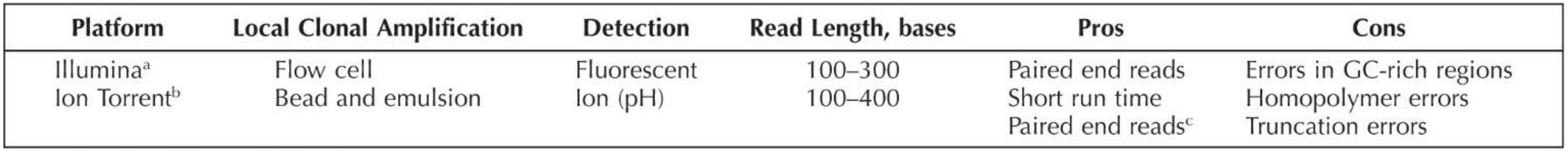


Term	Definition
Alignment	To compare a sequence read to another sequence and determine where it belongs. There are 2 types of alignment: de novo assembly or resequencing.
De novo assembly	A sequence read is compared to all the other sequence reads of that sample to determine a consensus sequence.
Resequencing	A sequence read is compared to a reference sequence (eg, the reference human genome). Also referred to as *mapping*.
Bait	An artificial construct that is able to target the sequence of interest (eg, a complementary DNA or RNA sequence) and can be used to isolate that target sequence. Used for sequence capture target enrichment.
Demultiplex	Separate an individual sample's reads from the pooled reads of multiple samples by unique identifier codes that were attached before pooling.
Map/mapping	To compare a sequence read to a reference and determine where it belongs. See also Alignment, Resequencing.
Read	May refer to either the sequence result of a single base pair position or to the sequence result of a sequential length of base pair reads from a single clonally amplified DNA cluster.



			Approximate File Size (Average Coverage 160×)	
File Type	Full Name	Description	Exome	4800 Genes
FASTQ	Files with consensus assessment of sequence and variation	Raw sequencing data after demultiplexing	50 GB	18 GB
BAM	Binary version of sequence alignment/map	Sequencing data after alignment	16 GB	6 GB
VCF	Variant call file	File containing variants called relative to the reference	9.3 GB	3.5 MB

Abbreviations: GB, gigabytes; MB, megabytes.



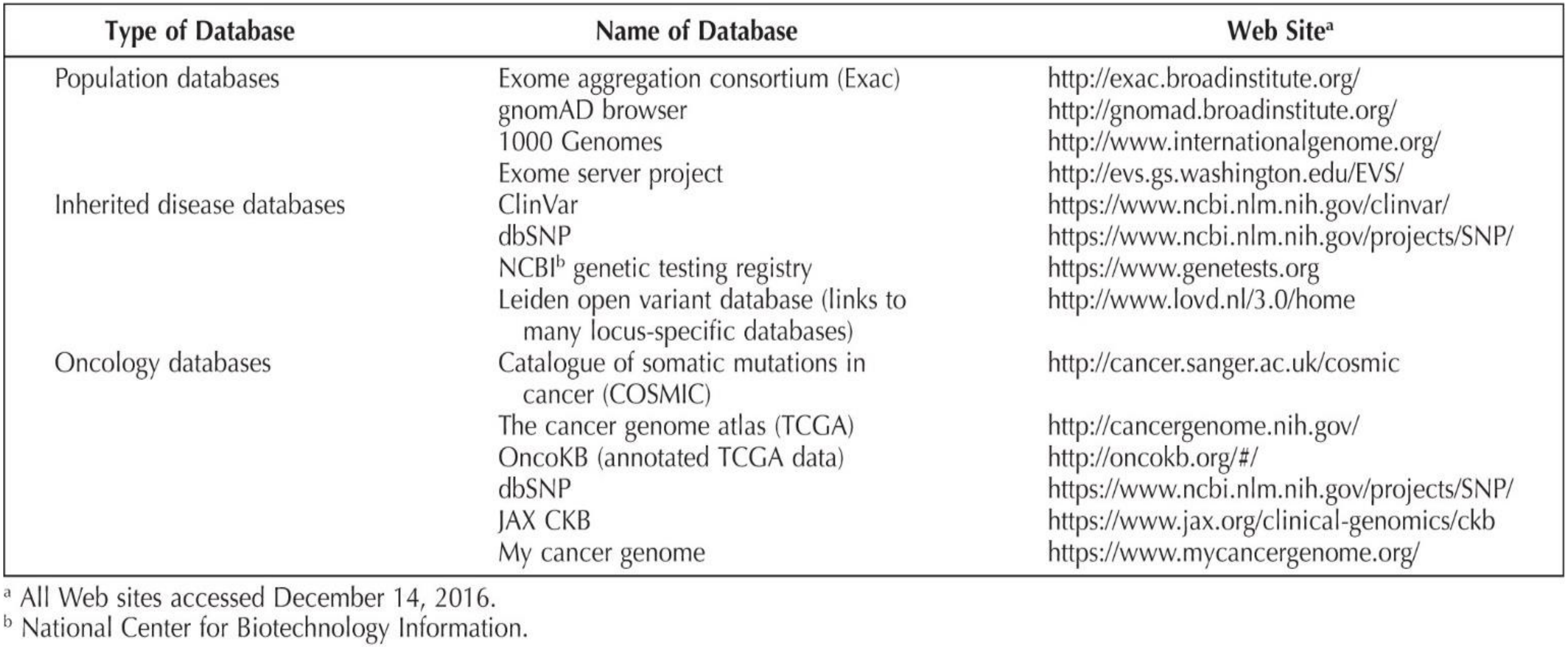
Platform	Local Clonal Amplification	Detection	Read Length, bases	Pros	Cons
Illumina[a]	Flow cell	Fluorescent	100–300	Paired end reads	Errors in GC-rich regions
Ion Torrent[b]	Bead and emulsion	Ion (pH)	100–400	Short run time Paired end reads[c]	Homopolymer errors Truncation errors

[a] Illumina, San Diego, California.

[b] ThermoFisher, Waltham, Massachusetts.

[c] Available on newer instruments only.

Recent genomic databases

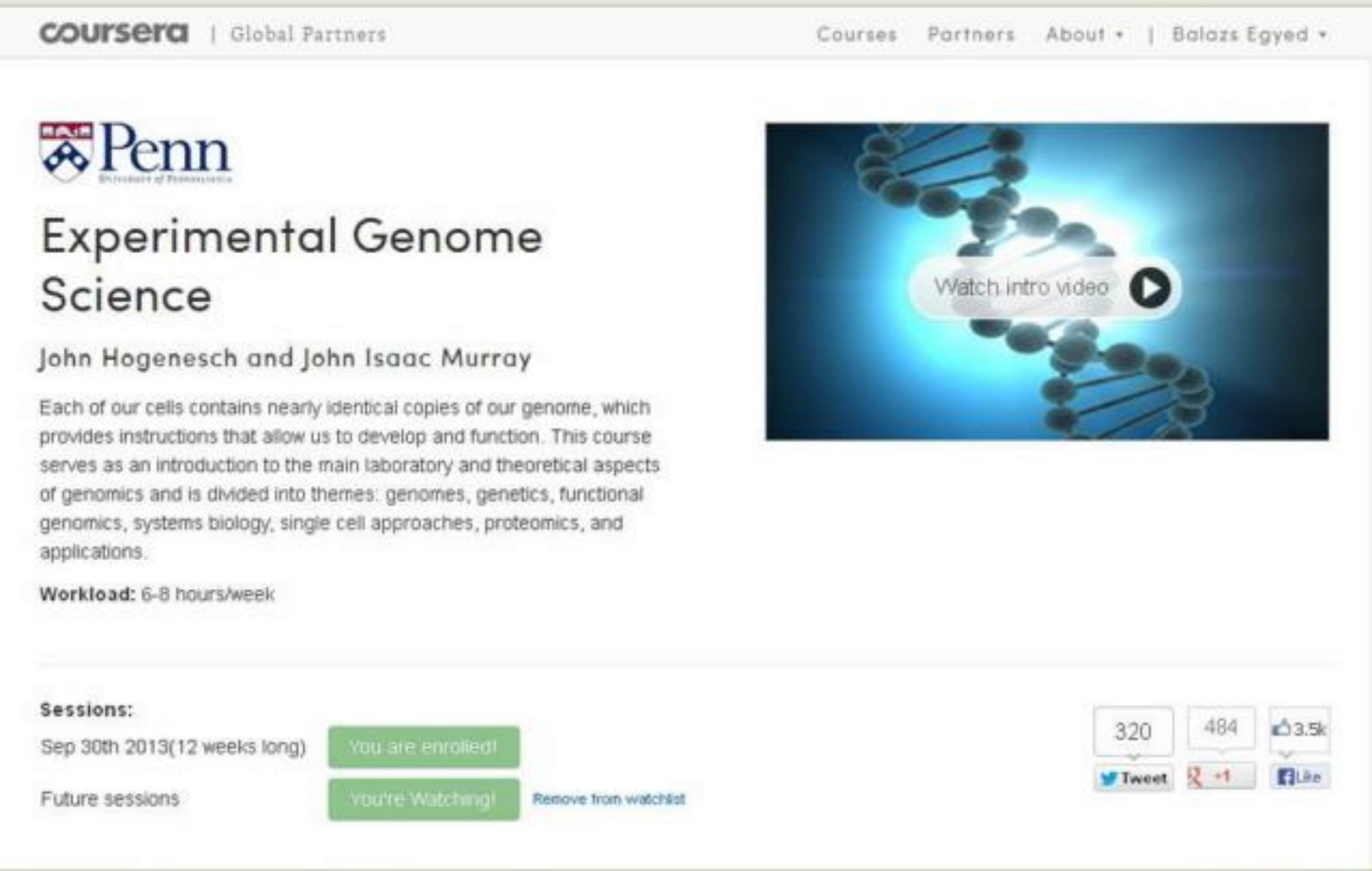


Type of Database	Name of Database	Web Site[a]
Population databases	Exome aggregation consortium (Exac)	http://exac.broadinstitute.org/
	gnomAD browser	http://gnomad.broadinstitute.org/
	1000 Genomes	http://www.internationalgenome.org/
	Exome server project	http://evs.gs.washington.edu/EVS/
Inherited disease databases	ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/
	dbSNP	https://www.ncbi.nlm.nih.gov/projects/SNP/
	NCBI[b] genetic testing registry	https://www.genetests.org
	Leiden open variant database (links to many locus-specific databases)	http://www.lovd.nl/3.0/home
Oncology databases	Catalogue of somatic mutations in cancer (COSMIC)	http://cancer.sanger.ac.uk/cosmic
	The cancer genome atlas (TCGA)	http://cancergenome.nih.gov/
	OncoKB (annotated TCGA data)	http://oncokb.org/#/
	dbSNP	https://www.ncbi.nlm.nih.gov/projects/SNP/
	JAX CKB	https://www.jax.org/clinical-genomics/ckb
	My cancer genome	https://www.mycancergenome.org/

[a] All Web sites accessed December 14, 2016.

[b] National Center for Biotechnology Information.

https://www.coursera.org/course/genomescience



Experimental Genome Science

John Hogenesch and John Isaac Murray

Each of our cells contains nearly identical copies of our genome, which provides instructions that allow us to develop and function. This course serves as an introduction to the main laboratory and theoretical aspects of genomics and is divided into themes: genomes, genetics, functional genomics, systems biology, single cell approaches, proteomics, and applications.

Workload: 6-8 hours/week

Sessions:

Sep 30th 2013(12 weeks long) — You are enrolled!

Future sessions — You're Watching! — Remove from watchlist