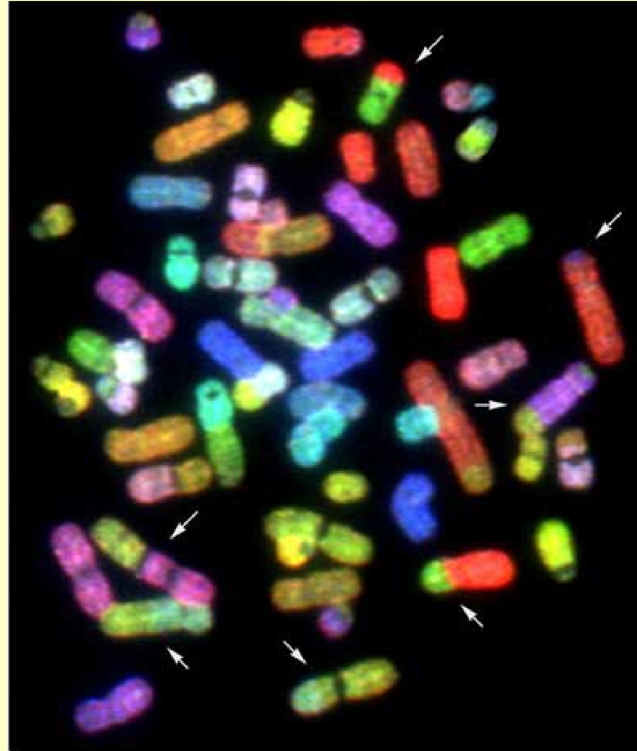
Your Genes and Your Health

http://bio84.stanford.edu/

The Human Genome Project



Doug Brutlag, Professor Emeritus
Biochemistry and Medicine (by courtesy)
brutlag@stanford.edu

© Doug Brutlag 2016

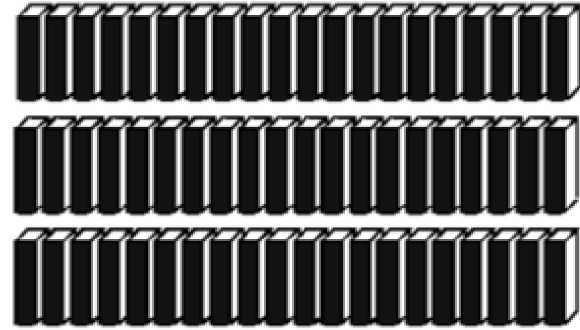
The Human Genome Project: Should we do it?

- Service, R. F. (2001). The human genome: Objection #1: big biology is b
  Science, 291(5507), 1182.
  - Not hypothesis driven.
  - Fishing expedition or stamp collecting.
  - Eliminate funds from investigator initiated science.
- Vogel, G. (2001). The human genome: Objection #2: why sequence the
  Science, 291(5507), 1184.
  - Limit sequencing to 1.5% of genome that codes proteins.
  - Do not sequence intergenic regions “genetic wastelands”.
  - Do not sequence repeated regions (centromeres, telomeres and heterochromatin).
- Service, R. F. (2001). The human genome: Objection #3: impossible to d
  Science, 291(5507), 1186.
  - Technology of the time permitted 500 bp per day per person.
  - Move from radioactively labeled sequencing to fluorescent sequencing permitted complete automation up to 1 gigabyte per year.

© Doug Brutlag 2016

Genome Sizes

Human Genome
Mouse Genome



~3,000,000,000 bp

Fruit Fly Genome

~160,000,000 bp

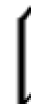
Nematode Genome

~100,000,000 bp

Yeast Genome

~15,000,000 bp

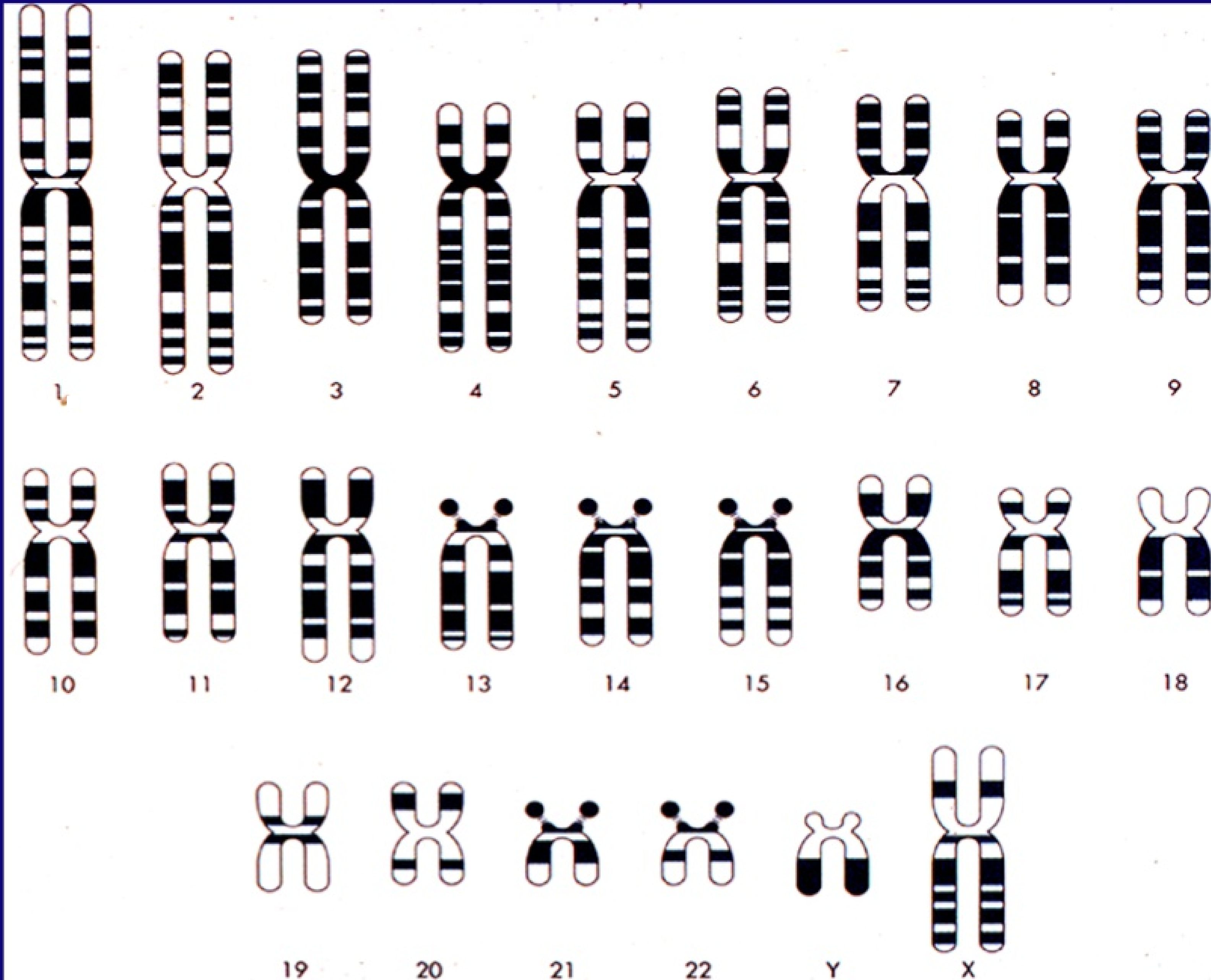
E. coli Genome



~5,000,000 bp

Courtesy
Eric Green

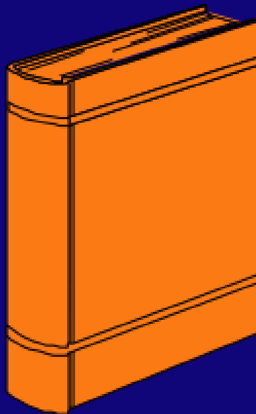
The Human Cytogenetic Map



urtesy
c Green

Human Genome

(~3000 Mb)



Human Chromosome

(~130 Mb)

GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT
TGTGACTAGCCACAGT
TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA
CCGGTTAGACATACAT
GAGGCCCACCGCCGCT
GTGCACGTCCACCACC

YAC

(~0.5-1.0 Mb)

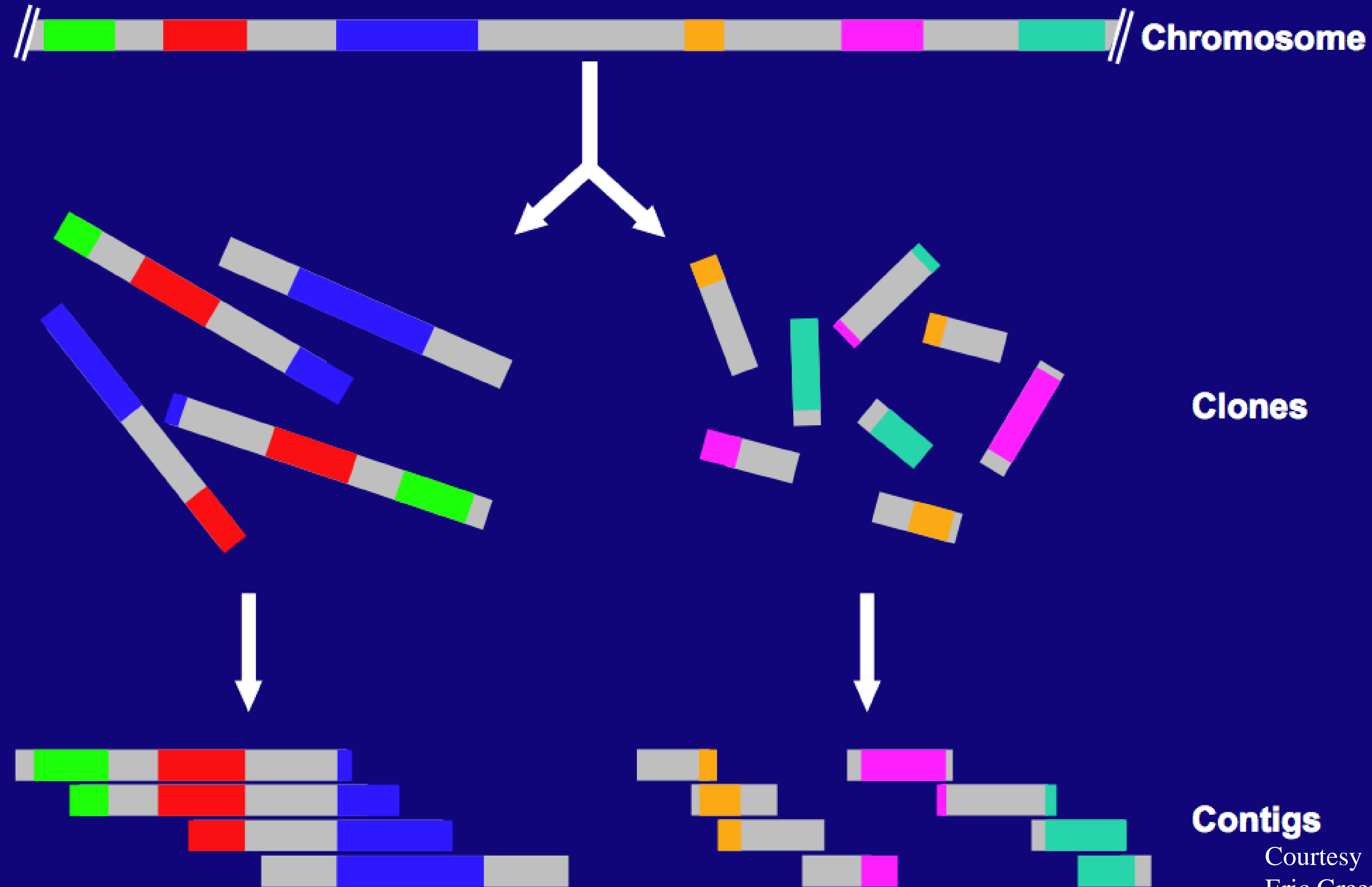
GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT
TAGGTATTGGGGCATT
TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA
CCGGTTAGACATACAT
GAGGCCCACCGCCGCT
GTGCACGTCCACCACC

BAC

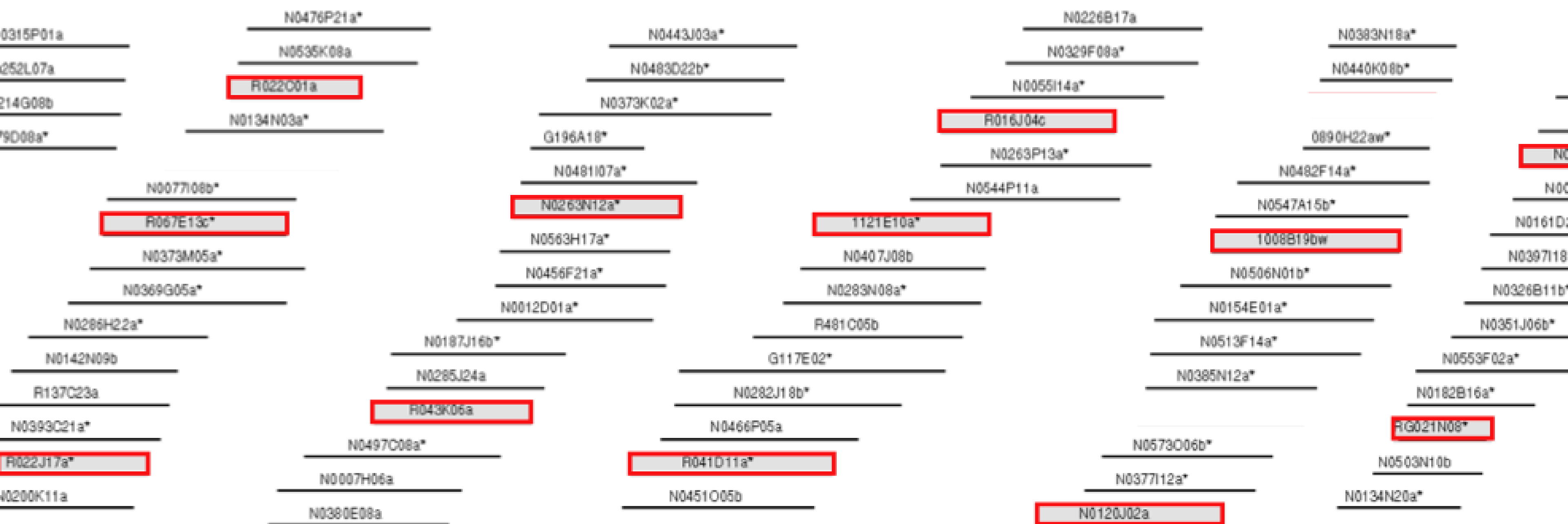
(~0.1-0.2 Mb)

Courtesy
Eric Green

Clone-Based Physical Mapping



Sequence-Ready BAC Contig Map



Courtesy
Eric Green

Subclone Construction

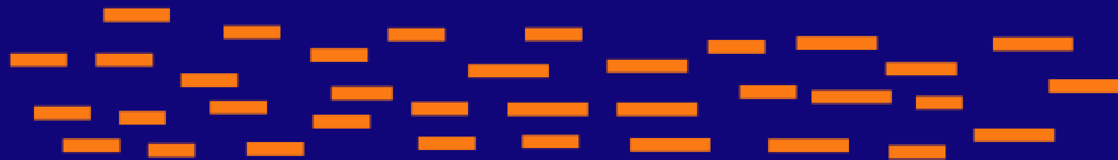
GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT
TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA
GAGGCCCACCGCCGCT
GTGCACGTCCACCACC
GATTATTACCATTTTA
ATCCTTAGGATTGACA

BAC DNA

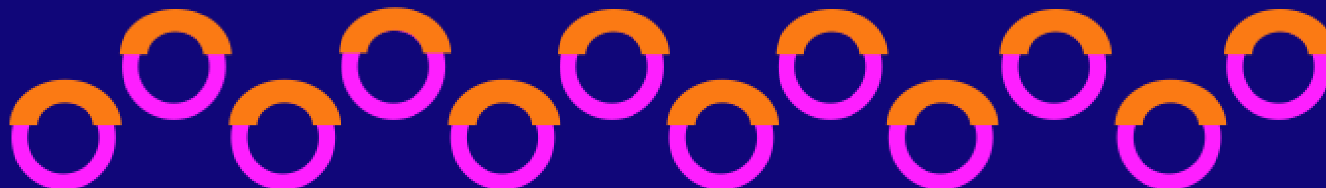
Prepare Multiple Copies

GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT
TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA
GAGGCCCACCGCCGCT
GTGCACGTCCACCACC
GATTATTACCATTTTA
ATCCTTAGGATTGACA

Randomly Fragment



Subclone Fragments

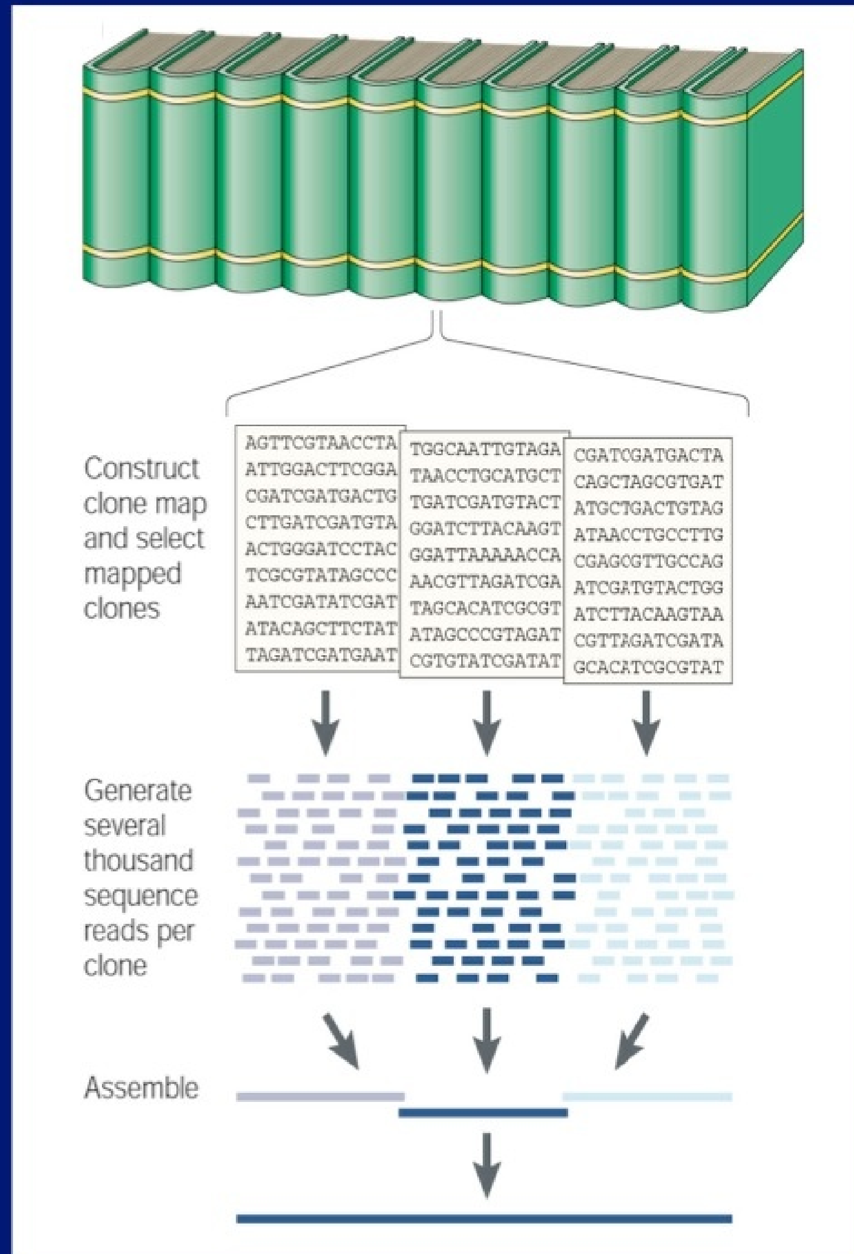


Courtesy
Eric Green

Shotgun Sequencing Strategy

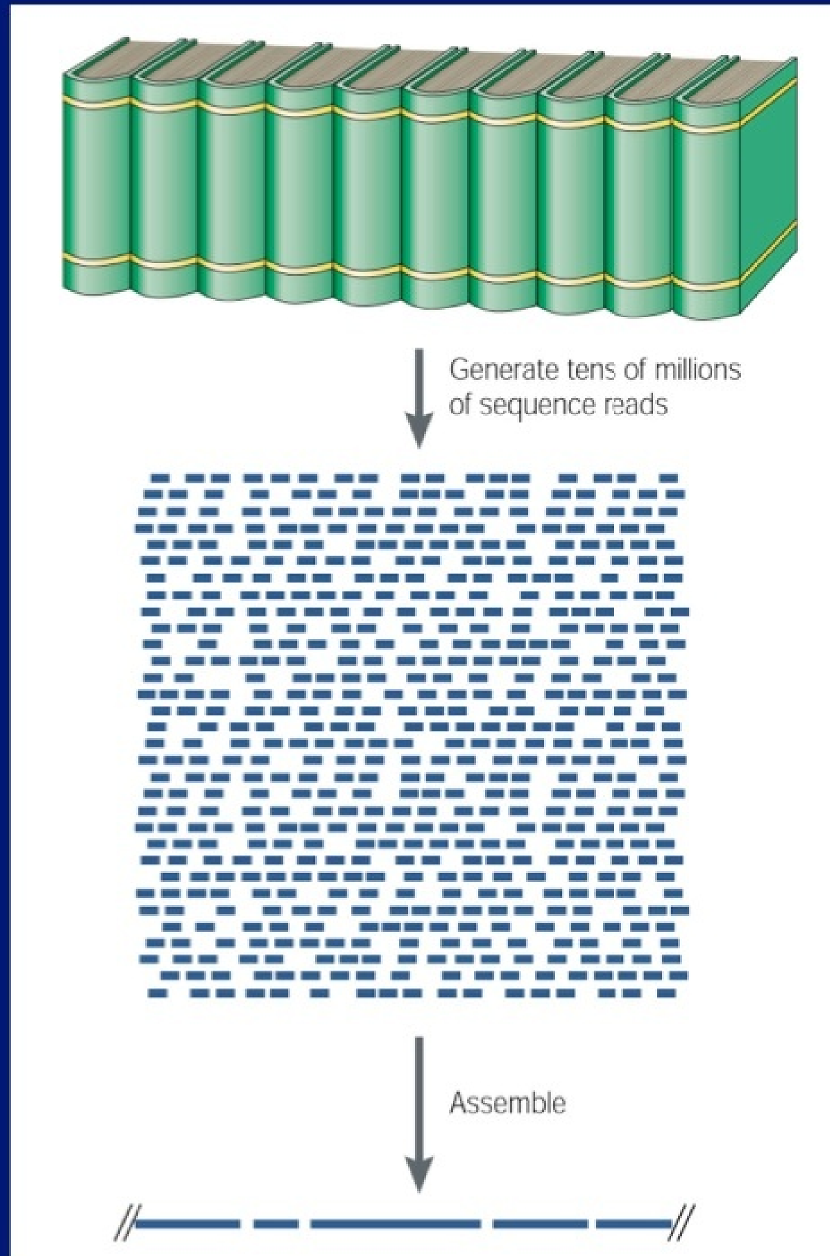
Courtesy
Eric Green

Clone-Based Shotgun Sequencing



Courtesy
Eric Green

Whole-Genome Shotgun Sequencing

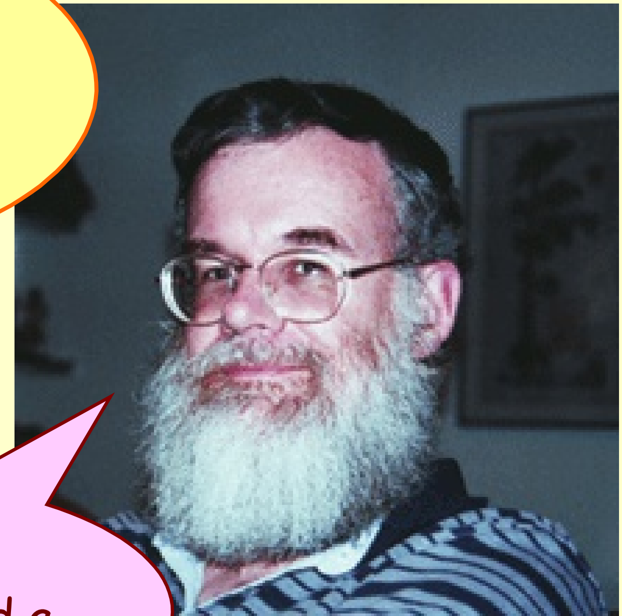


Courtesy
Eric Green

Whole Genome Shotgun versus Clone Sequencing

1997

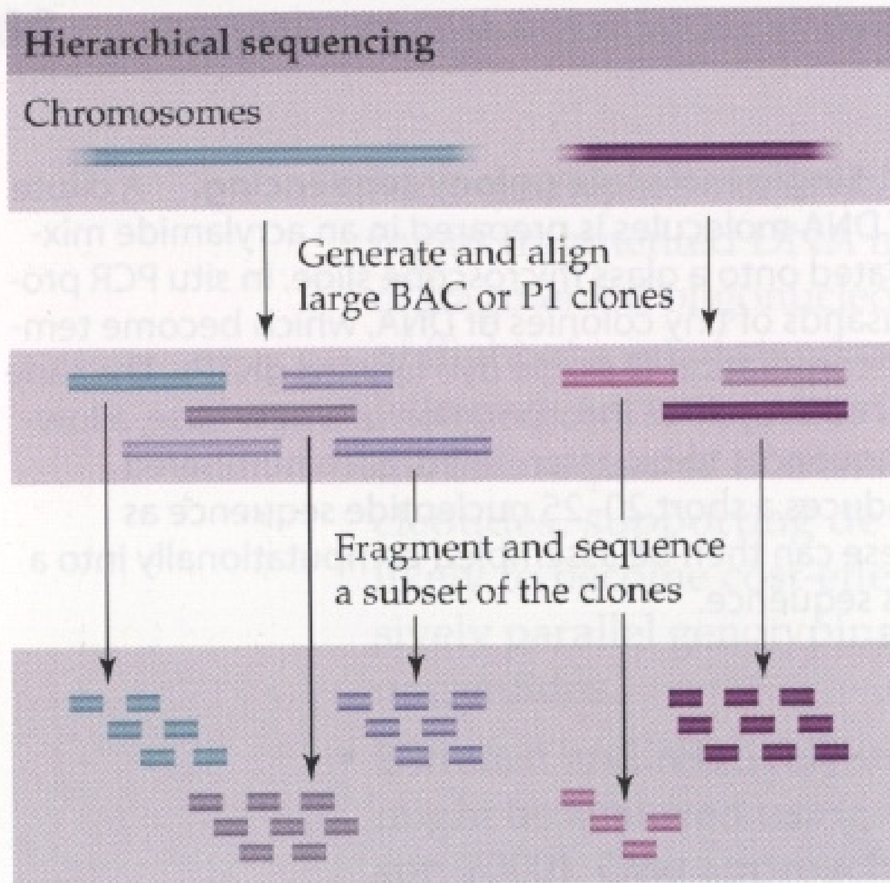
Gene Myers

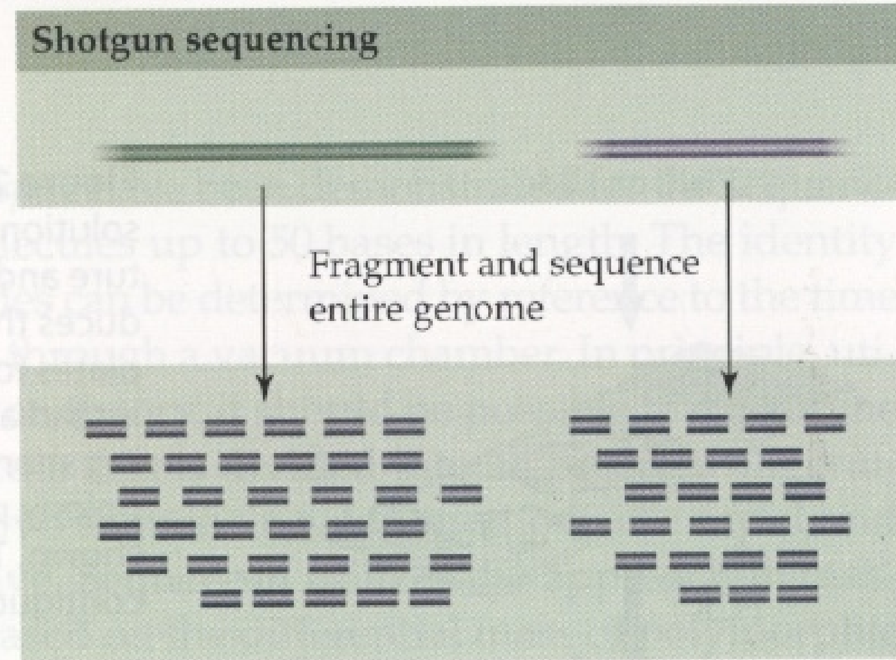


Phil Green

Thanks to Seraf im Batzoglou

Hierarchical Sequencing Vs. Whole Genome Shotgun Sequencing





(from Gibson & Muse, A Primer of Genome Science)

© Gibson & Muse A Primer of Genome Science

M13mp18 Sequencing Vector

http://www.mikeblaber.org/oldwine/bch5425/lect33/lect33.htm

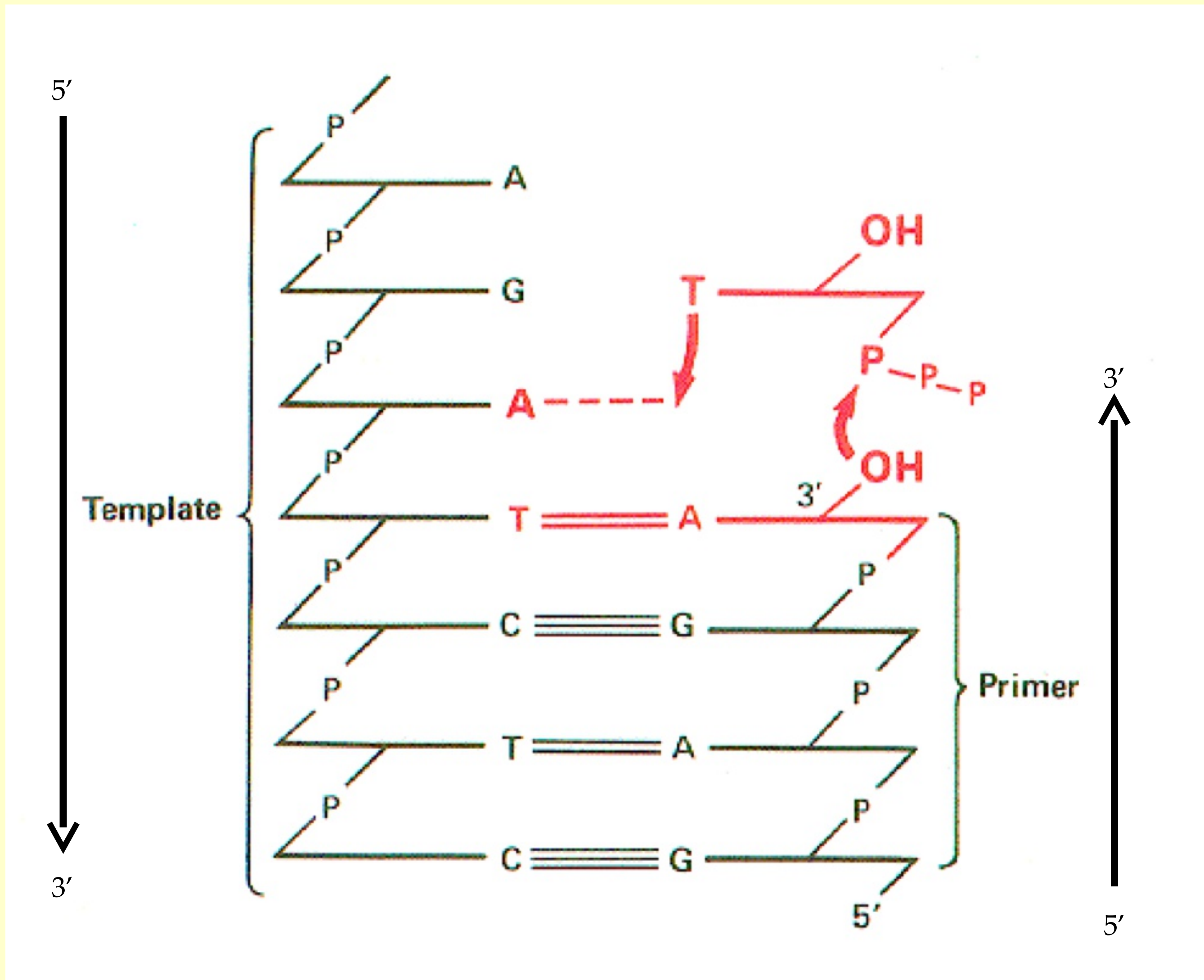
Insert 1 kb Human DNA Segment

LacZ Gene

M13 Circular DNA With LacZ gene

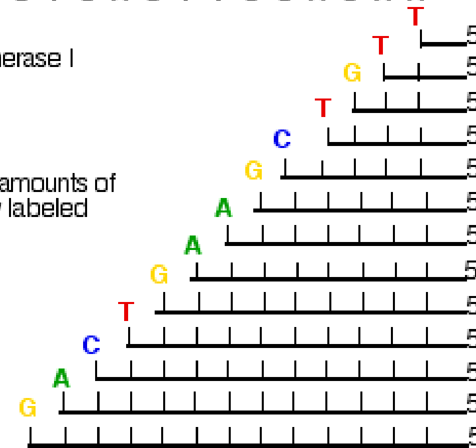
© Doug Brutlag 2016

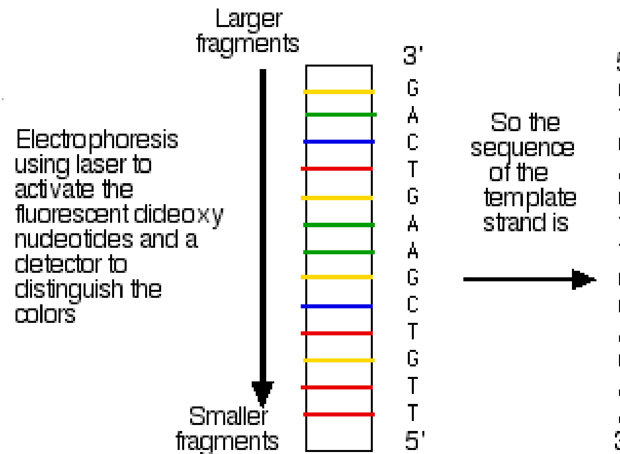
DNA Synthesis by DNA polymerases



© Doug Brutlag 2016

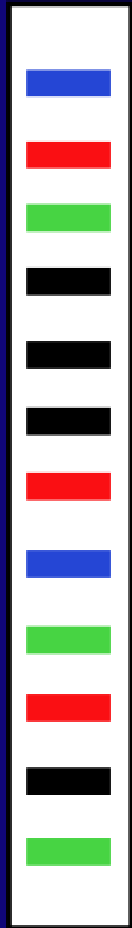
DNA Sequencing by Chain Termination





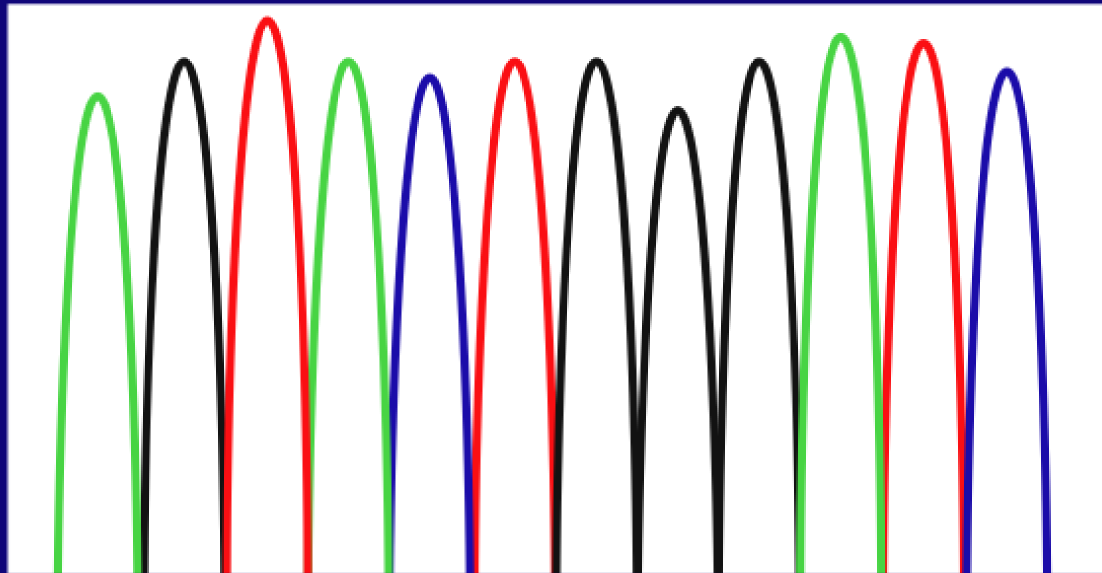
© Doug Brutlag 2016

Analyzing Fluorescent DNA Sequencing Data



Computer Analysis

A G T A C T G G G A T C



Courtesy
Eric Green

Fluorescent DNA Sequencing Results

Sequence Assembly

(a) Sequence reads

Read 1 CACATACACATGG
Read 2 TCAATGGGGCTAA
Read 3 AGCACGGACTTGTCACATACACATG
Read 4 ACACATGGAAATA
Read 5 GGGCTAATGATTGTCAC
Read 6 TGATTGTCACATA
Read 7 ATTCATGAAGCACGGA
Read 8 GTCACATACACATGATCAATGGGG

Use computer to assemble sequence reads

(b)

7 ATTCATGAAGCACGGA
3 AGCACGGACTTGTCACATACACATG
8 GTCACATACACATGATCAATGGGG
2 TCAATGGGGCTAA
5 GGGCTAATGATTGTCAC
6 TGATTGTCACATA
1 CACATACACATGG
4 ACACATGGAAATA

Assembled sequence

Contig

(c) ATTCATGAAGCACGGACTTGTCACATACACATGATCAATGGGGCTAATGATTGTCACATACACATGGAAATA

Courtesy Eric Green

February, 2001 Draft Sequence

15 February 2001

nature

www.nature.com

the human genome

Nuclear fission
Five-dimensional energy landscapes

Seafloor spreading
The view from under the Arctic icepack

Career prospects
Sequence creates new opportunities

naturejobs
genomics special

16 February 2001

Science

Vol. 291 No. 5507
Pages 1145–1434 $9

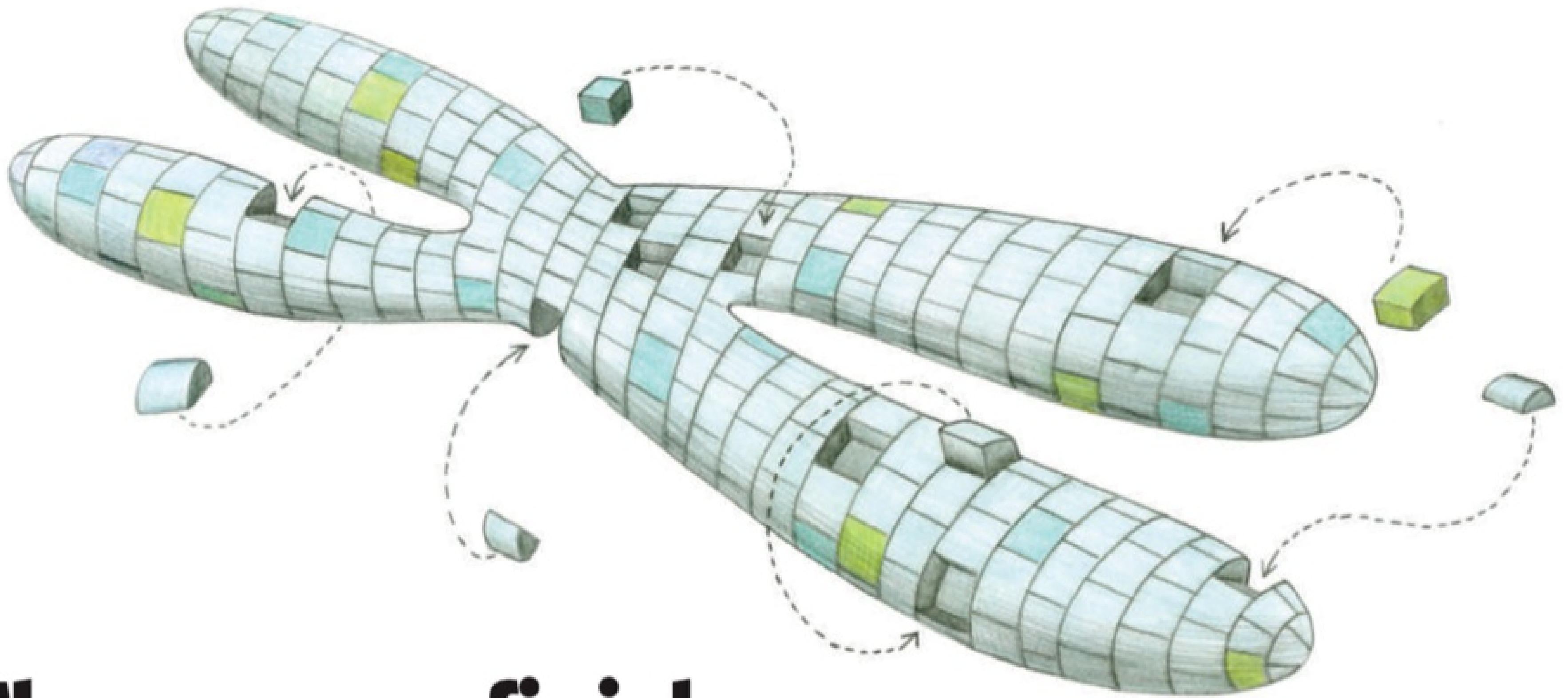
THE HUMAN GENOME

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

International Human Genome Sequencing Consortium (2001)

Venter et al. (2001)

Courtesy
Eric Green

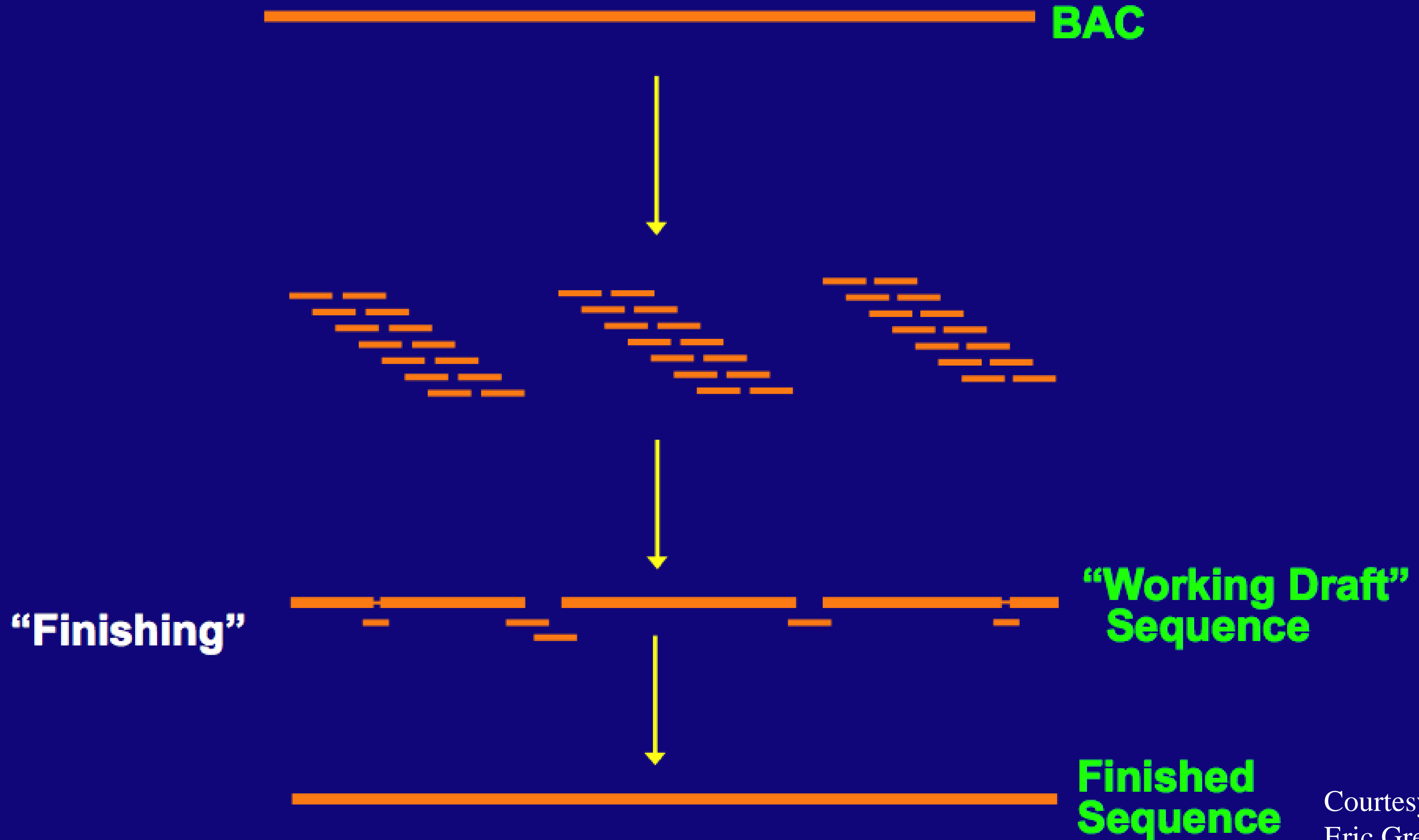


The genome finishers

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.

Courtesy
Eric Green

Shotgun Sequencing Strategy



Polymerase Chain Reaction Overview: Exponential Amplification of DNA

© Doug Brutlag 2016

PCR Requirements

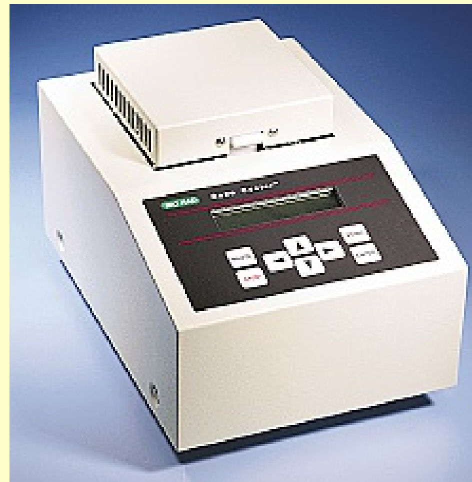
DNA

- Need to know at least the beginning and end of DNA sequence
- These flanking regions have to be unique to strand interested in amplifying
- Region of interest can be present in as little as one copy
- *Enough DNA in 0.1 microliter of human saliva to use PCR*

DNA Polymerase Enzyme

- DNA polymerase from *Thermus aquaticus*--Yellowstone
- Alternatives: *Thermococcus litoralis*, *Pyrococcus furiosus*

Thermocycler



© Doug Brutlag 2016

PCR Applications

Forensics

- assessment/reassessment of crimes
- 13 FBI CODIS markers

Archaeology

- determine gene sequences of ancient organisms
  - Neandertals
  - Denisovans
  - Otzi
- rethinking the past, human origins

Molecular Biology

- Cloning genes
- Sequencing genes
- Finishing genome sequences
- Amplification of DNA or RNA

- Medicine
  - Diagnostics for inherited disease
  - Diagnostics for gene expression
  - Diagnostics for epigenetics

© Doug Brutlag 2016

October, 2004 Publication

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

*A list of authors and their affiliations appears in the Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human enome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

The Human Genome Project (HGP) was launched in 1990 with the goal of obtaining a highly accurate sequence of the vast majority of the euchromatic portion of the human genome. The initial work followed a two-pronged approach: (1) the mapping of the human and mouse genomes to allow the study of inherited disease and provide a crucial scaffold for genome assembly; and (2) the sequencing of organisms with smaller, simpler genomes to serve as a testbed for method development and assist in interpreting the human genome. With success along both paths, the sequencing of the human genome itself eventually became feasible. The International Human Genome Sequencing Consortium (IHGSC), an open collaboration involving twenty centres in six countries, was formed to carry out this component of the HGP.

In February 2001, the IHGSC and Celera Genomics each reported draft sequences providing a first overall view of the human genome. These sequences allowed systematic study of the human genome itself, including identification of genes, combinatorial architecture of proteins, regional differences in genome composition, distribution and history of transposable elements, distribution of polymorphism and relationship between genetic recombination and physical distance. Moreover, systematic knowledge of the human genome has enabled new tools and approaches that have markedly accelerated biomedical research.

Both draft sequences, however, had important shortcomings. The IHGSC sequence, for example, omitted ~10% of the euchromatic genome; it was interrupted by ~150,000 gaps; and the order and orientation of many segments within local regions had not been established. The IHGSC thus turned to the challenge of completing the sequence of the euchromatic genome. Operationally, a finished sequence was defined as having an error rate of, at most, one event per 10^4 bases, and the goal for completion was coverage in finished sequence of at least 95% of the euchromatic genome, with the only gaps being those refractory to all available techniques (see http://www.genome.gov/10000923). The goal was challenging because the human genome is replete with such features as dispersed repeats and large segmental duplications, which greatly complicate the determination of genome structure and sequence. In fact, near-complete sequences have been obtained so far only for three multicellular organisms: the nematode, mustard weed and the fruitfly. These genomes are all roughly 30-fold smaller than the human genome and have much simpler structure.

We describe here the results of a multiyear effort by the IHGSC towards the goal of a complete human sequence. The number of gaps has been reduced 400-fold to only 341, most of which are associated with segmental duplications and will require new methods for resolution. The assembled near-complete genome sequence has an error rate of only ~1 event per 100,000 bases; it contains 2.85 billion nucleotides and covers ~99% of the euchromatic genome. This paper describes the current genome sequence and the process used to produce it; examines the accuracy and completeness of the sequence; and illustrates biological analyses made possible by the sequence. We do not attempt here a comprehensive analysis of the contents of the human genome. An initial analysis was previously reported and a series of papers is being written describing the individual chromosomes, including annotation of genes and other features.

Current genome sequence

Finishing process

The process of converting the initial draft sequence into a near-complete sequence is referred to as 'finishing'. It is a complex iterative process that proceeds simultaneously at multiple scales, ranging from single nucleotides to the integrity of whole chromosomes. The fundamental challenge is that genomic regions that are not well represented or readily resolved through random shotgun sequencing tend to be highly enriched in problematic sequences. Resolving such regions required the development of special approaches, which evolved substantially over time and varied among centres.

Broadly, the finishing process involved two distinct components: (1) producing finished maps, consisting of continuous and accurate paths of overlapping large-insert clones spanning the euchromatic region of each chromosome arm; and (2) producing finished clones, consisting of continuous and accurate nucleotide sequence across each large-insert clone. In practice, these two components were tightly intertwined in that progress in each often depended on results from the other. The components are described in Boxes 1 and 2. Further information about the finishing process and finishing standards can be found in the Supplementary Information (Note 1) and at http://www.genome.gov/10000923.

In total, we generated a shotgun sequence from 59,208 large-insert clones (total length ~5.84 gigabases (Gb)) and finished the sequence from 45,742 of these clones (total length ~3.67 Gb). The clones consisted primarily of bacterial artificial chromosomes

©2004 Nature Publishing Group

International Human Genome Sequencing Consortium (2004)

Courtesy Eric Green

Duplications and Deletions in the Human Genome

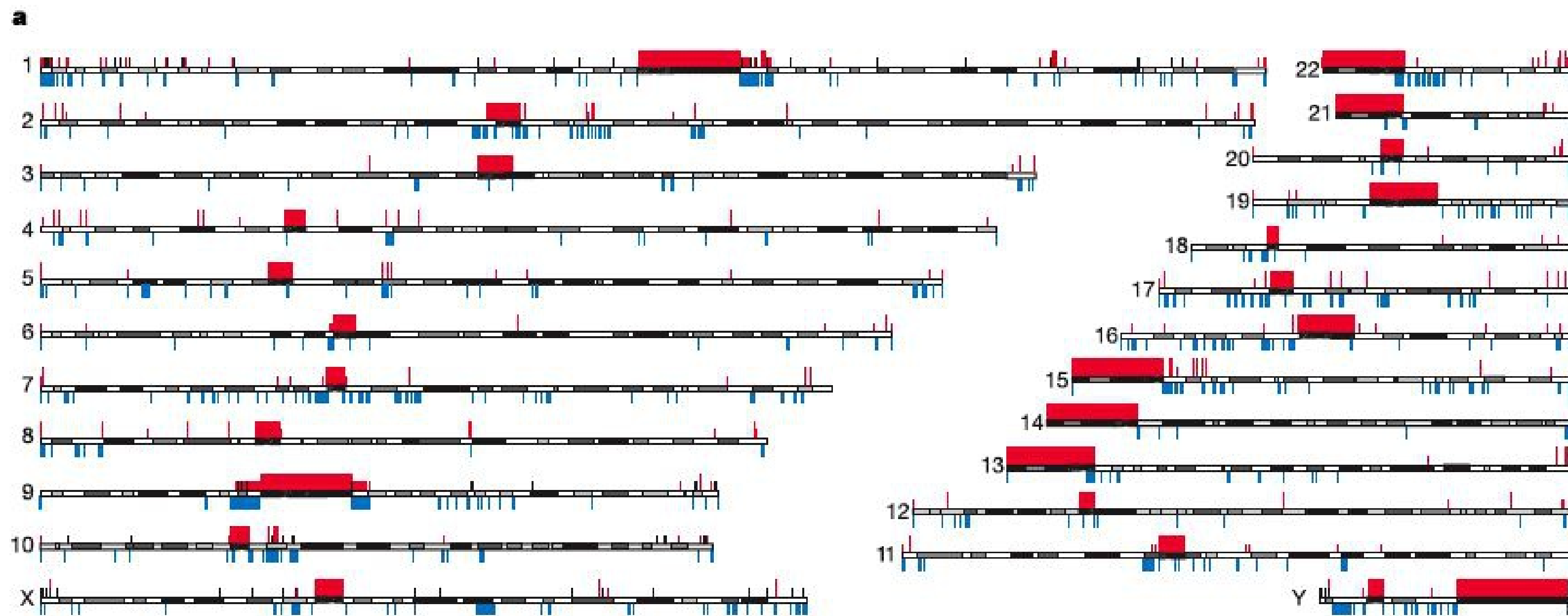
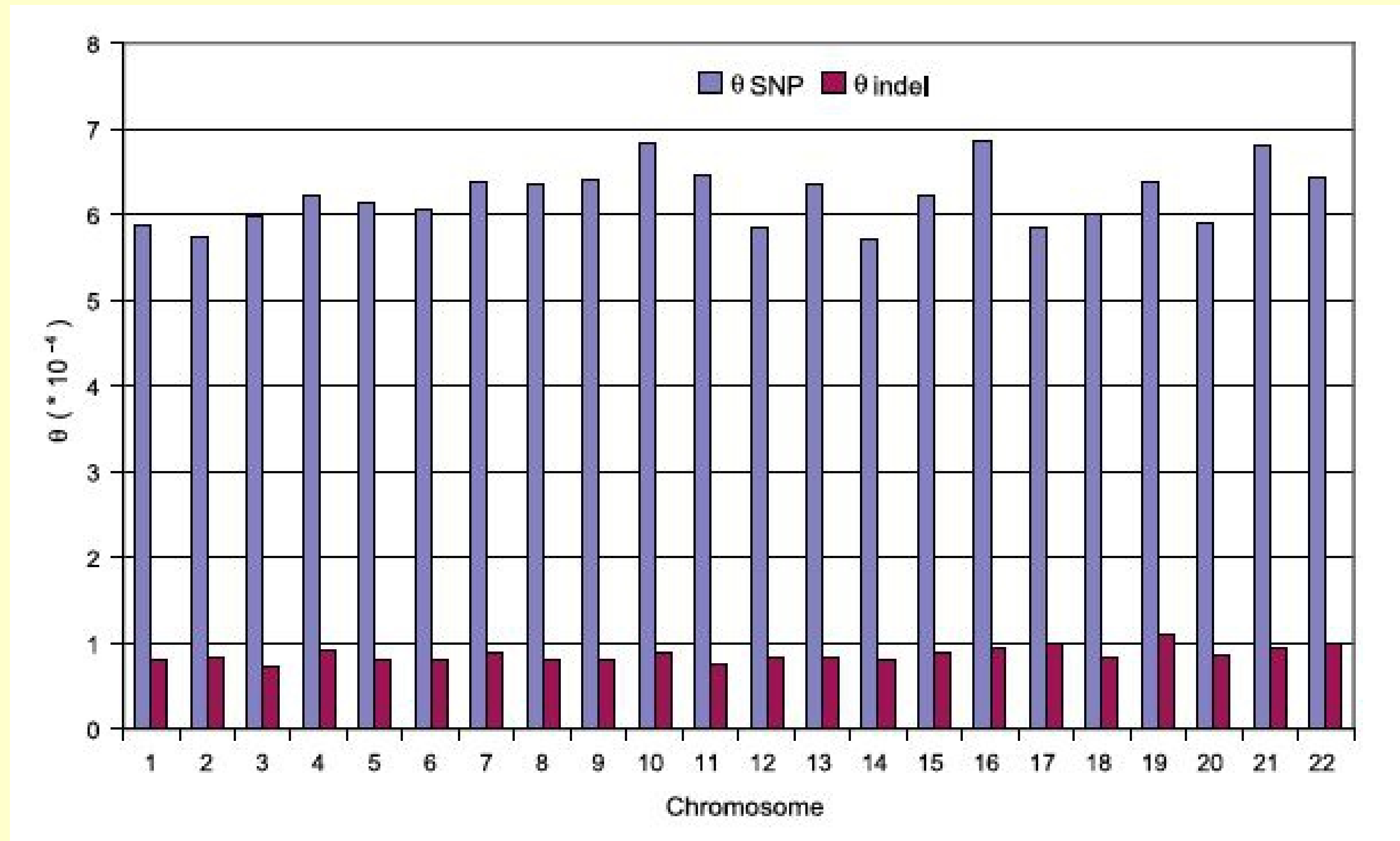


Figure 4 Segmental duplications across the genome. **a**, Segmental duplications and sequence gaps across the genome. Segmental duplications are indicated below the chromosomes in blue (length ≥10 kb and sequence identity ≥95%). Large duplications are shown to approximate scale; smaller ones are indicated as ticks. Sequence gaps are indicated above the chromosomes in red. Large gaps (>300 kb) are shown to approximate scale; smaller gaps are indicated as ticks with those that are 50 kb or smaller shown as shorter ticks. Unfinished clones are indicated as black ticks. **b**, Percentage of

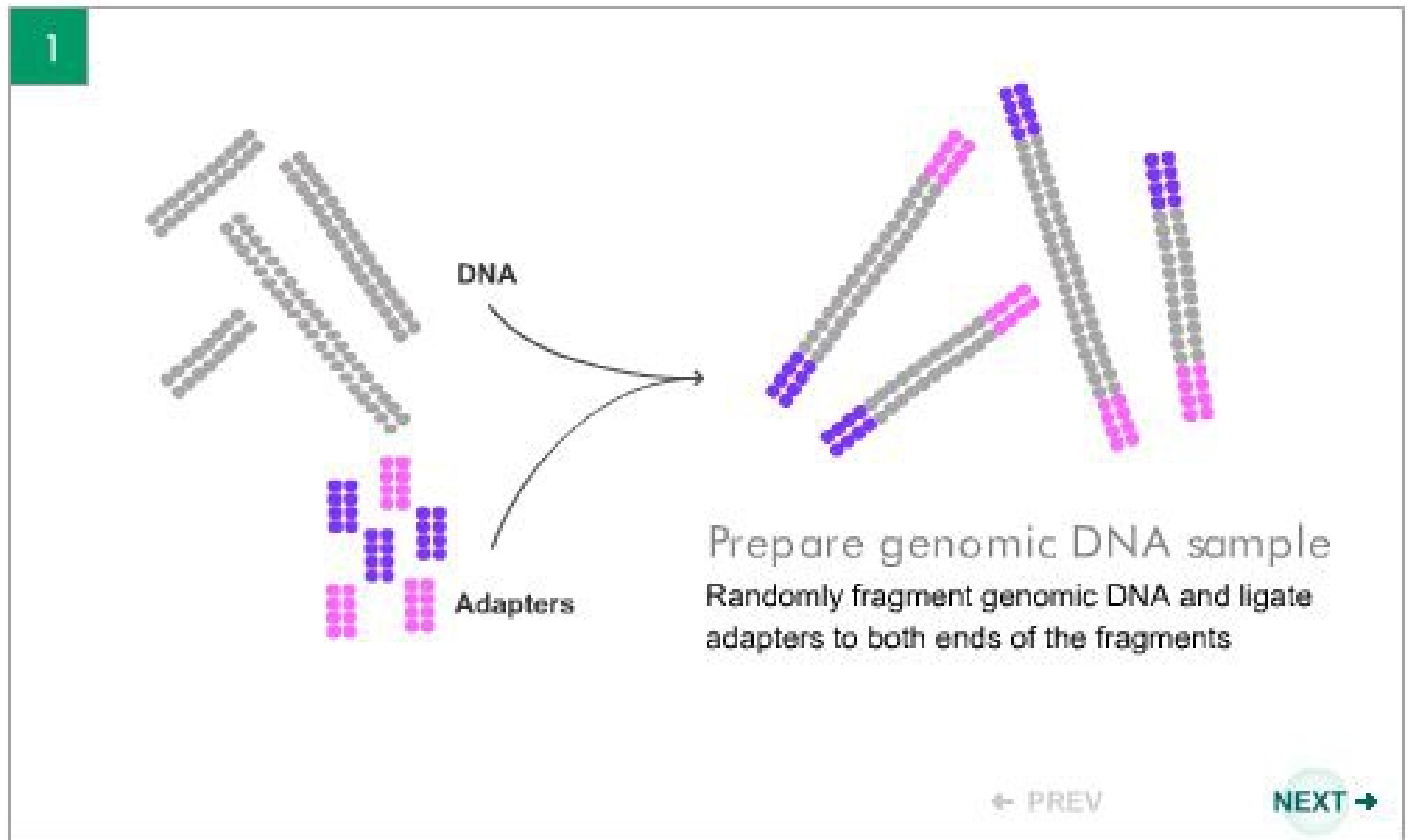
© Doug Brutlag 2016

Single Nucleotide Polymorphisms (SNPs) and Short Insertion/Deletions in The Human Genome

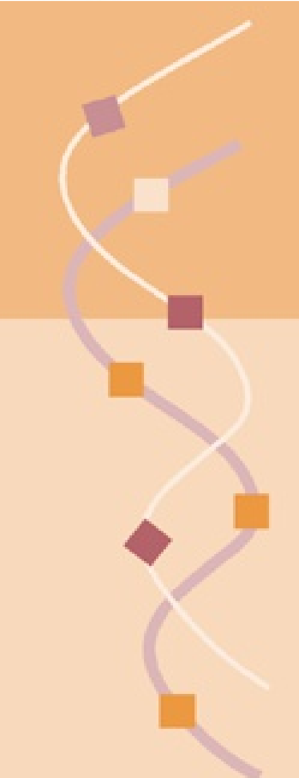


© Doug Brutlag 2016

Illumina Solexa Sequencing Technology



© Doug Brutlag 2016



Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

2

Adapter

DNA fragment

Dense lawn of primers

Adapter

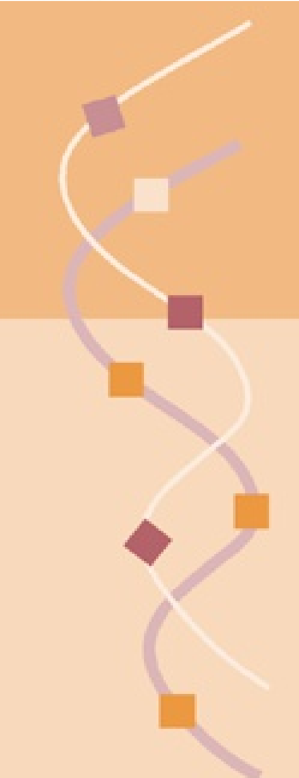
Attach DNA to surface

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

← PREV

NEXT →

© Doug Brutlag 2016



Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

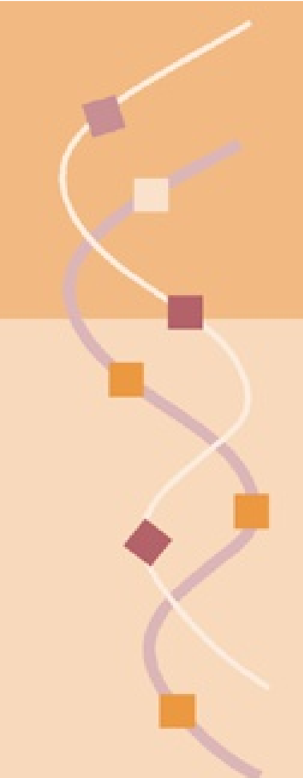
3

Bridge amplification

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

← PREV NEXT →

© Doug Brutlag 2016



Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

4

Attached terminus

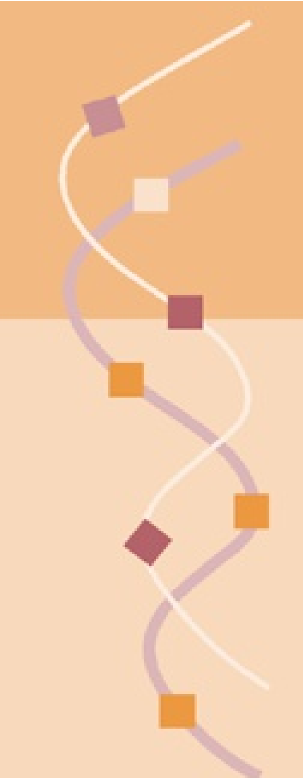
Free terminus

Fragments become double stranded

← PREV

NEXT →

© Doug Brutlag 2016



Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

5

Attached

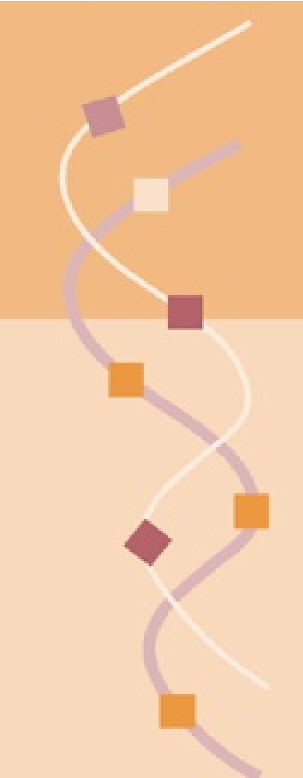
Denature the double stranded molecules

← PREV

NEXT →

© Doug Brutlag 2016

illumina



Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

6

Clusters

Completion of amplification

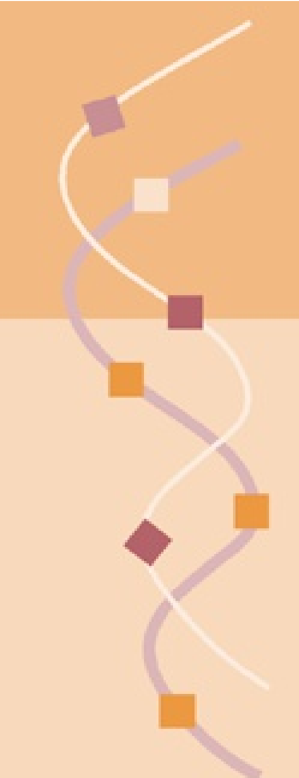
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.

← PREV NEXT →

© Doug Brutlag 2016

illumina

Illumina Solexa Sequencing Technology



Sequencing-By-Synthesis Demo

7

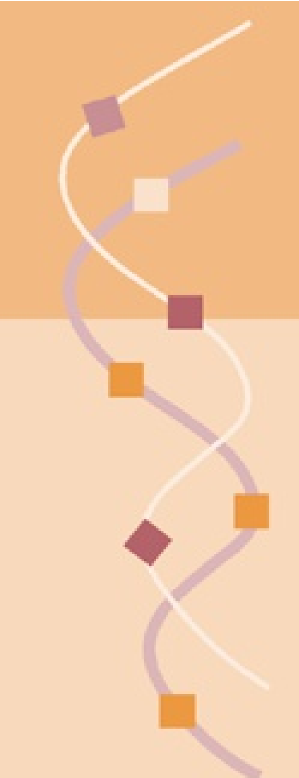
First chemistry cycle: determine first base

To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

Laser

← PREV NEXT →

© Doug Brutlag 2016



Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

8

T
G
C
A

Image of first chemistry cycle

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle

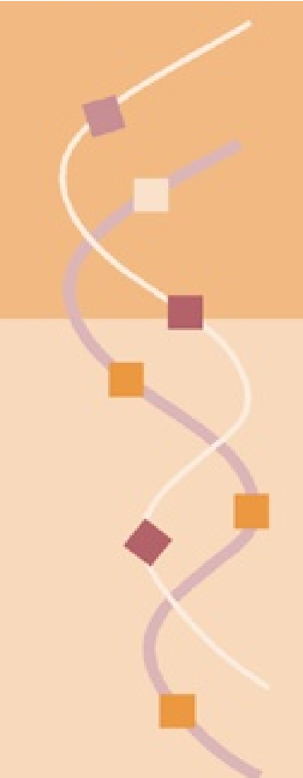
The blocked 3' terminus and the fluorophore from each incorporated base are removed.

← PREV NEXT →

© Doug Brutlag 2015

illumina

Illumina Solexa Sequencing Technology



Sequencing-By-Synthesis Demo

9

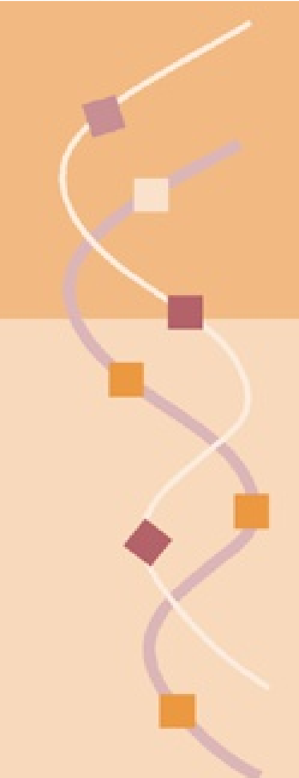
Second Chemistry Cycle:
determine second base

To initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

Laser

← PREV   NEXT →

© Doug Brutlag 2016



Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

10

A
C
T
G

Image of second chemistry cycle is captured by the instrument

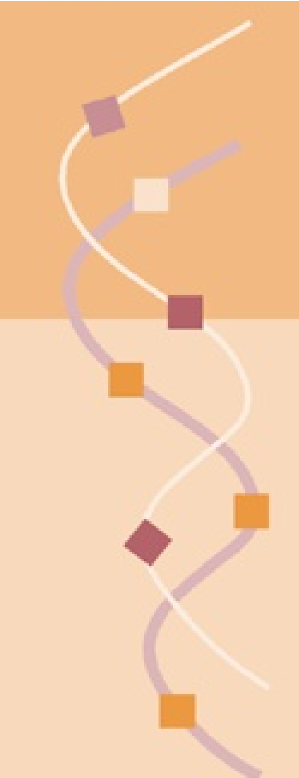
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

← PREV NEXT →

© Doug Brutlag 2016

illumina

Illumina Solexa Sequencing Technology



Sequencing-By-Synthesis Demo

11

Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5
T, G, C, A	A, C, T, G	A, U, T, C	G, G, A, C	T, A, G, G

GCTGA....

Sequence read over multiple chemistry cycles

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

← PREV NEXT →

© Doug Brutlag 2016

Pacific Biosciences SMRT Sequencing

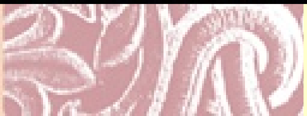
PACIFIC BIOSCIENCES™

© Doug Brutlag 2016

PACIFIC BIOSCIENCES™

Pacific Biosciences SMRT Sequencing

SMRT® Cell



© Doug Brutlag 2016

PACIFIC BIOSCIENCES™

Pacific Biosciences Sequencing

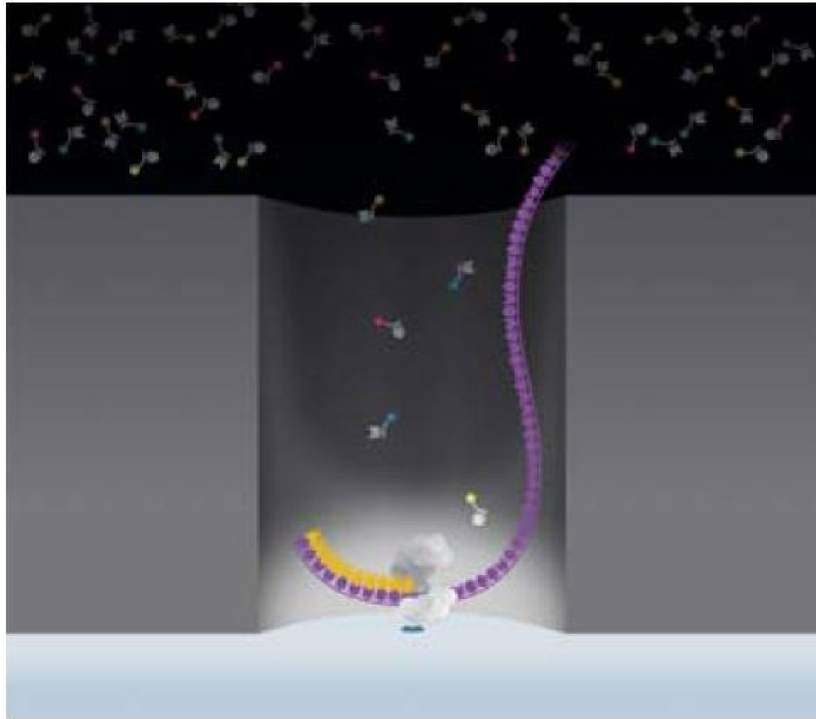


Figure 6. ZMW with DNA polymerase and phospholinked nucleotides

Phospholinked nucleotides are added into the ZMW at the high concentrations required for proper enzyme functioning.

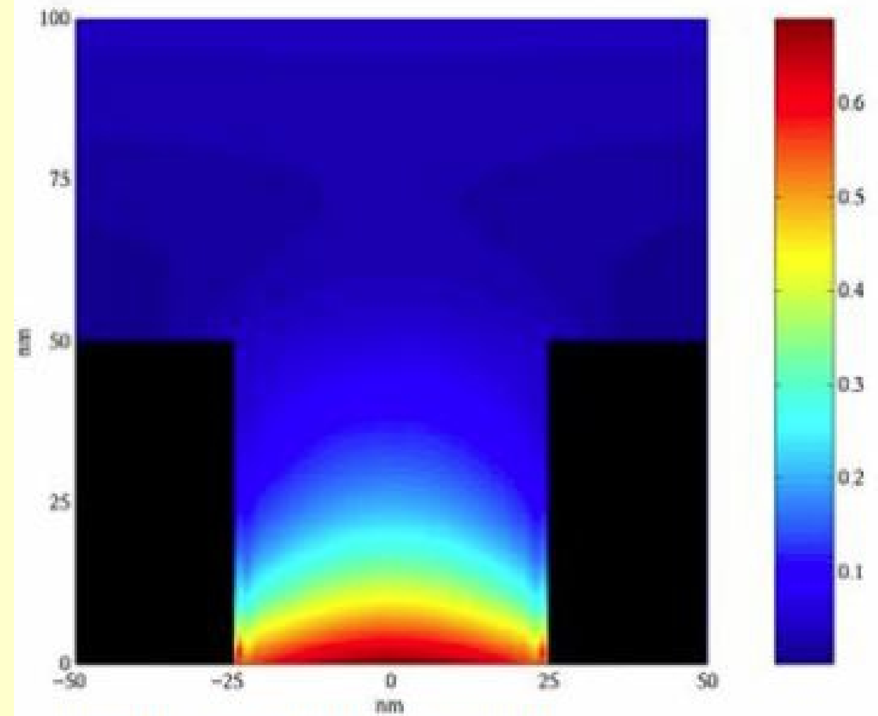
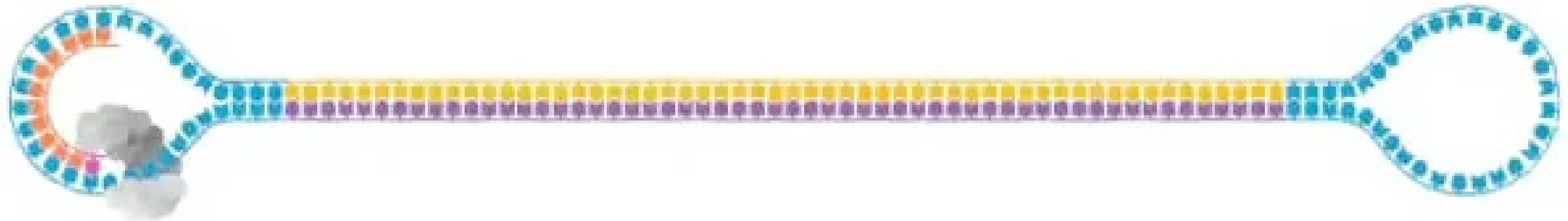


Figure 4. Detection volume

Attenuated light from the excitation beam penetrates only the lower 20-30 nm of each waveguide, creating a detection volume of 20 zeptoliters (10^{-21} liters).

PACIFIC BIOSCIENCES

Circular Templates Gives Redundant Sequencing and Accuracy



© Doug Brutlag 2016

ion torrent

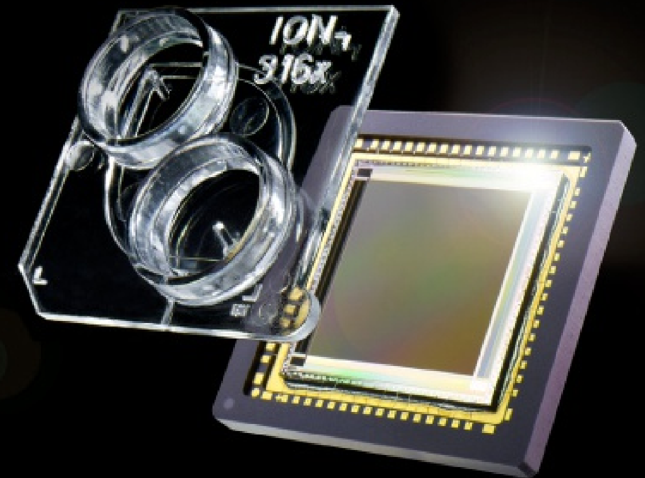
Ion Torrent Sequencing

Semiconductor Sequencing for Life™

Search

HOME | TECHNOLOGY | PRODUCTS | APPLICATIONS | COMMUNITY | SUPPORT | STORE

fastest next-gen workflow
10X more throughput
fastest-selling sequencer
all in six months



Ion Torrent
Ion 316™ -- Everything moves faster
when The Chip is the Machine™

Watch the video >
Read more >

Publications

nature
An integrated semiconductor device enabling non-optical genome sequencing

Application Note

The Ion PGM™ sequencer exhibits superior long-read accuracy
Learn more >

Life Grand Challenges

life
grand challenges

News

August 2011
Combating Superbug...more >

Ion Torrent long read accuracy, download app note & data...more >

View more Ion News...go >

© Doug Brutlag 2016

Ion Torrent Sequencing

ion torrent

ion torrent

Semiconductor Sequencing for Life™

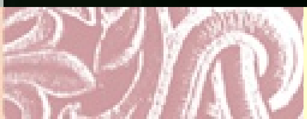
www.iontorrent.com

1 87 SEQUENCE = 1 877-378-3623

© Doug Brutlag 2016

Ion Torrent Sequencing

ion torrent



© Doug Brutlag 2016

Ion Torrent Promises $1,000 Genome

http://uk.reuters.com/article/2012/01/10/us-dna-reader-idUKTRE8090B820120110

Insight: New DNA reader to bring promise

Related Topics

Health »

Special Reports »

By Sharon Begley

NEW YORK | Tue Jan 10, 2012 7:06pm GMT

(Reuters) - After years of predictions that the "$1,000 genome" - a read-out of a person's complete genetic information for about the cost of a dental crown - was just around the corner, a U.S. company is announcing Tuesday that it has achieved that milestone and taken the technology several steps ahead.

Quotes

Amgen Inc
AMGN.O
$69.14
▼ -0.08 ▼ -0.12%
21:00:09 BST

Life Technologies Corp
LIFE.O
$46.90
▼ -0.15 ▼ -0.32%
21:00:03 BST

Pfizer Inc
PFE.N
$21.84
▼ -0.13 ▼ -0.59%
21:01:26 BST

Illumina Announces $1,000 Genome
J.P. Morgan Tech Show 1-16-2014

illumina

Contact Us MyIllumina Tools

APPLICATIONS SYSTEMS CLINICAL SERVICES SCIENCE SUPPORT COMPANY

Search

Systems / **HiSeq X Ten**

Subscribe | Follow us:
| Select Language

Overview System GET A QUOTE

Population power. Extreme throughput. $1,000 human genome.

The HiSeq X Ten is a set of ten ultra-high-throughput sequencers, purpose-built for large-scale human whole-genome sequencing.

HAVE QUESTIONS ABOUT THE HISEQ X TEN? CONTACT A SALES REPRESENTATIVE »

The First $1000 Genome

Discover how HiSeq X Ten breaks the $1000 genome barrier for human whole-genome sequencing.
Learn more »

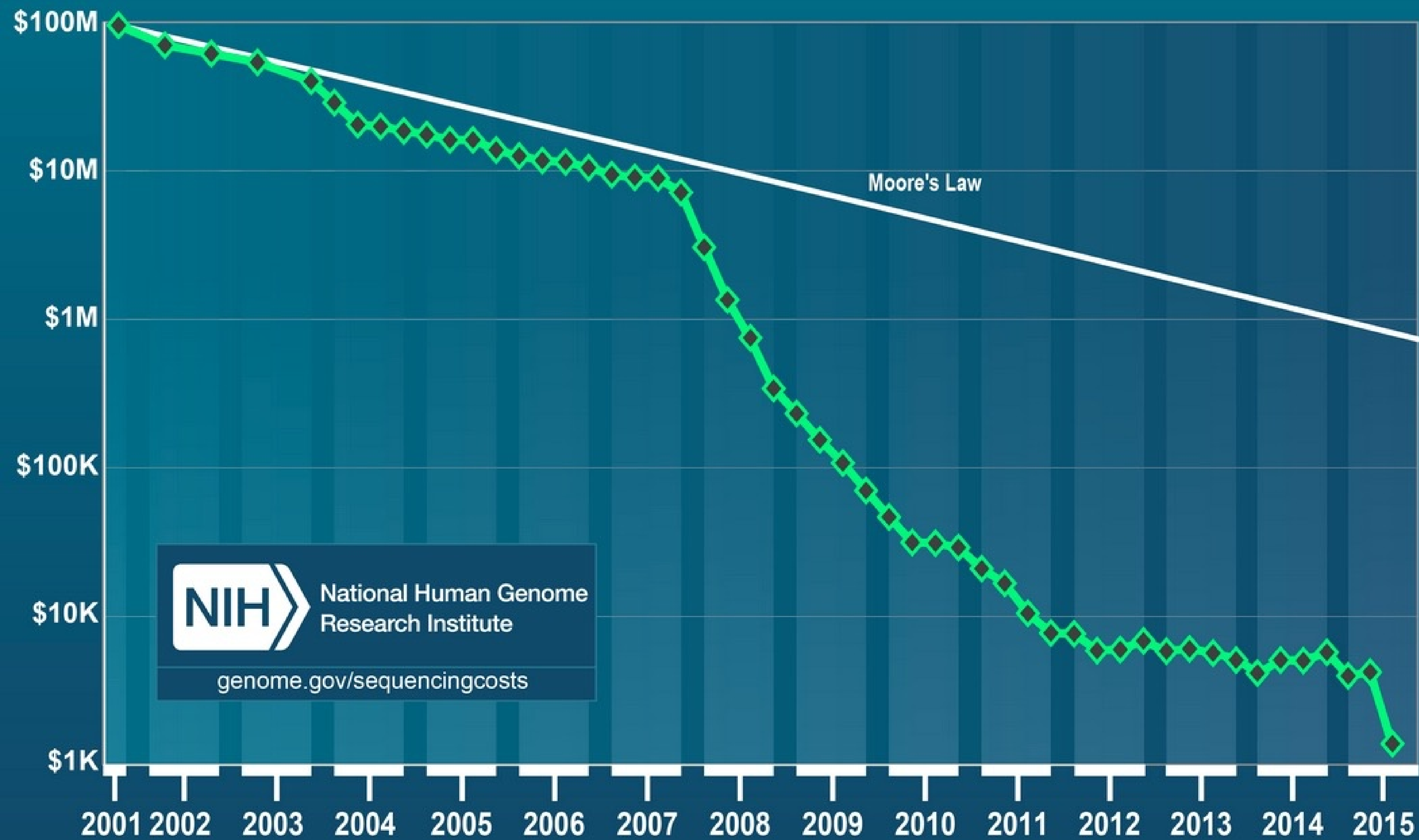
Population Scale Studies

Learn how the HiSeq X Ten can benefit communities by enabling them to sequence their entire population.
Read blog post »

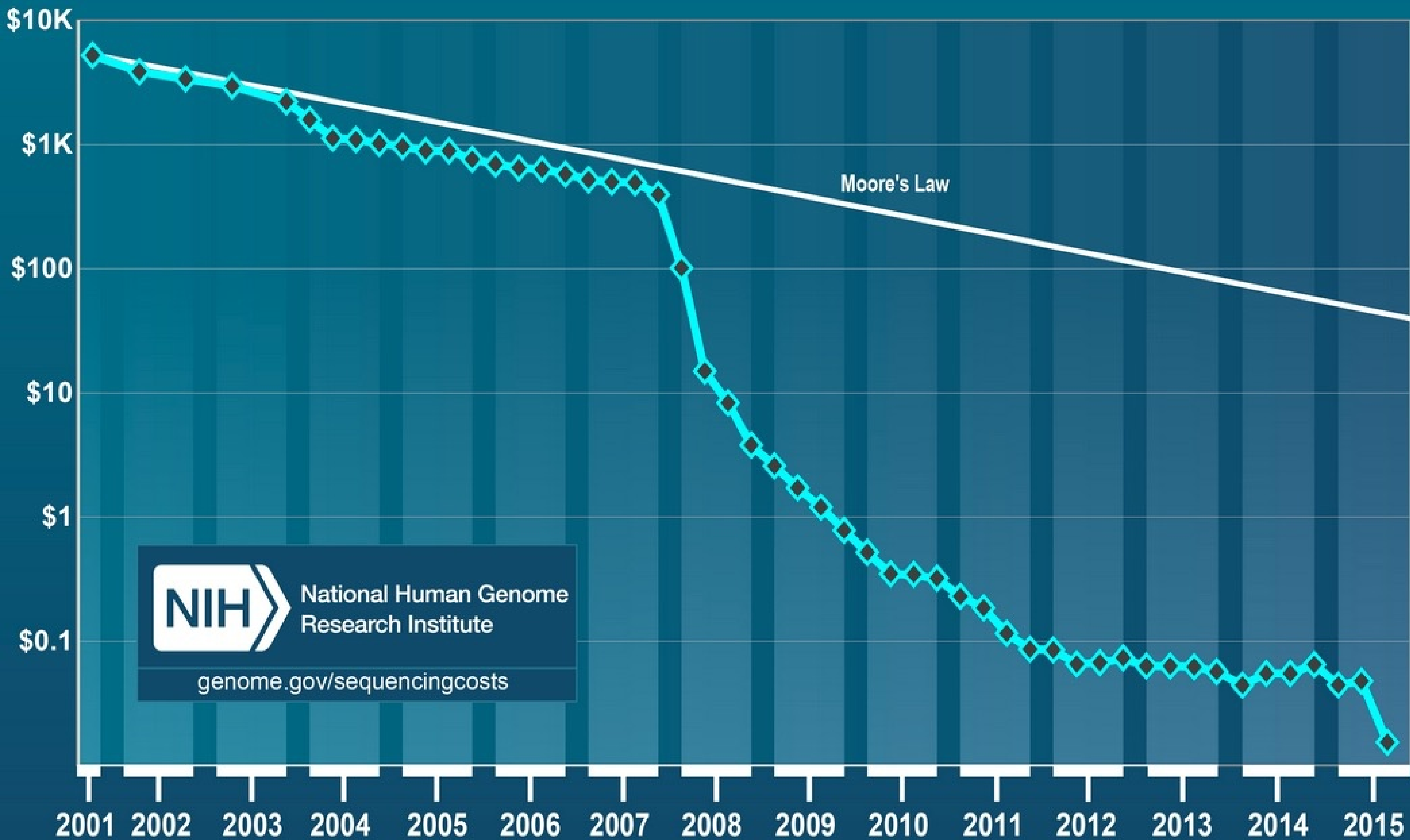
COMPARE SEQUENCING PLATFORMS

© Doug Brutlag 2016

Cost per Genome

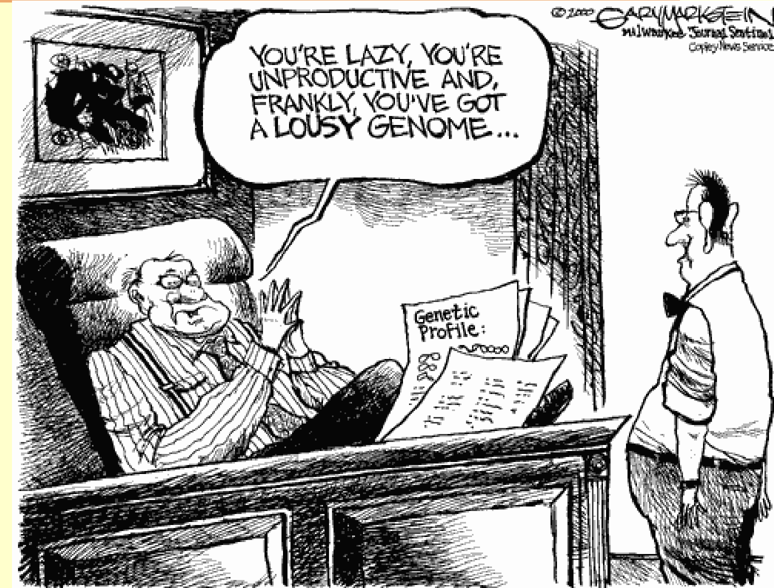


Cost per Raw Megabase of DNA Sequence



The Human Genome
How fast is the cost going down?

- 2006: $ 50 million
- 2008: $500,000
- 2009: $50,000
- 2010: $20,000
- 2011: $5,000
- 2012: $4,000
- 2013: $3,000
- 2014 $1,400
- 2015 $1,000



"The locket contains a strand of my DNA."

Thanks to Seraf im Batzoglou