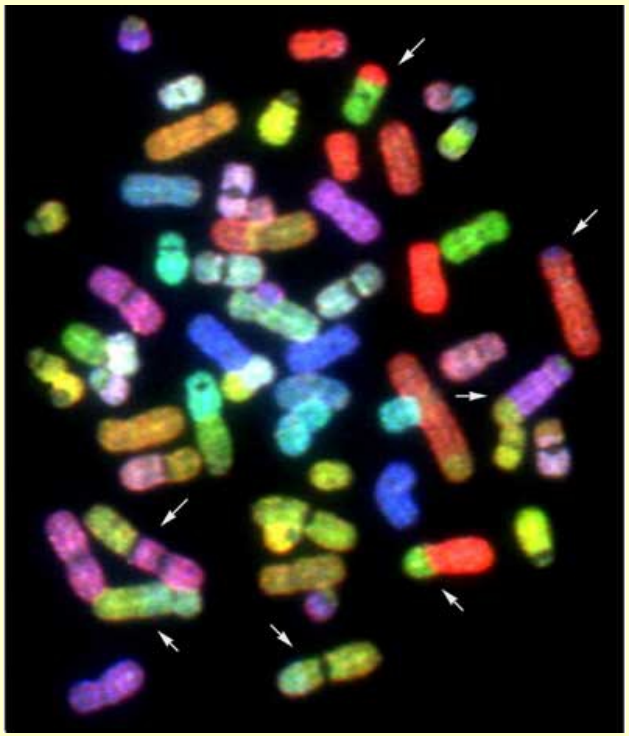
Finishing the Human Genome

http://biochem158.stanford.edu/

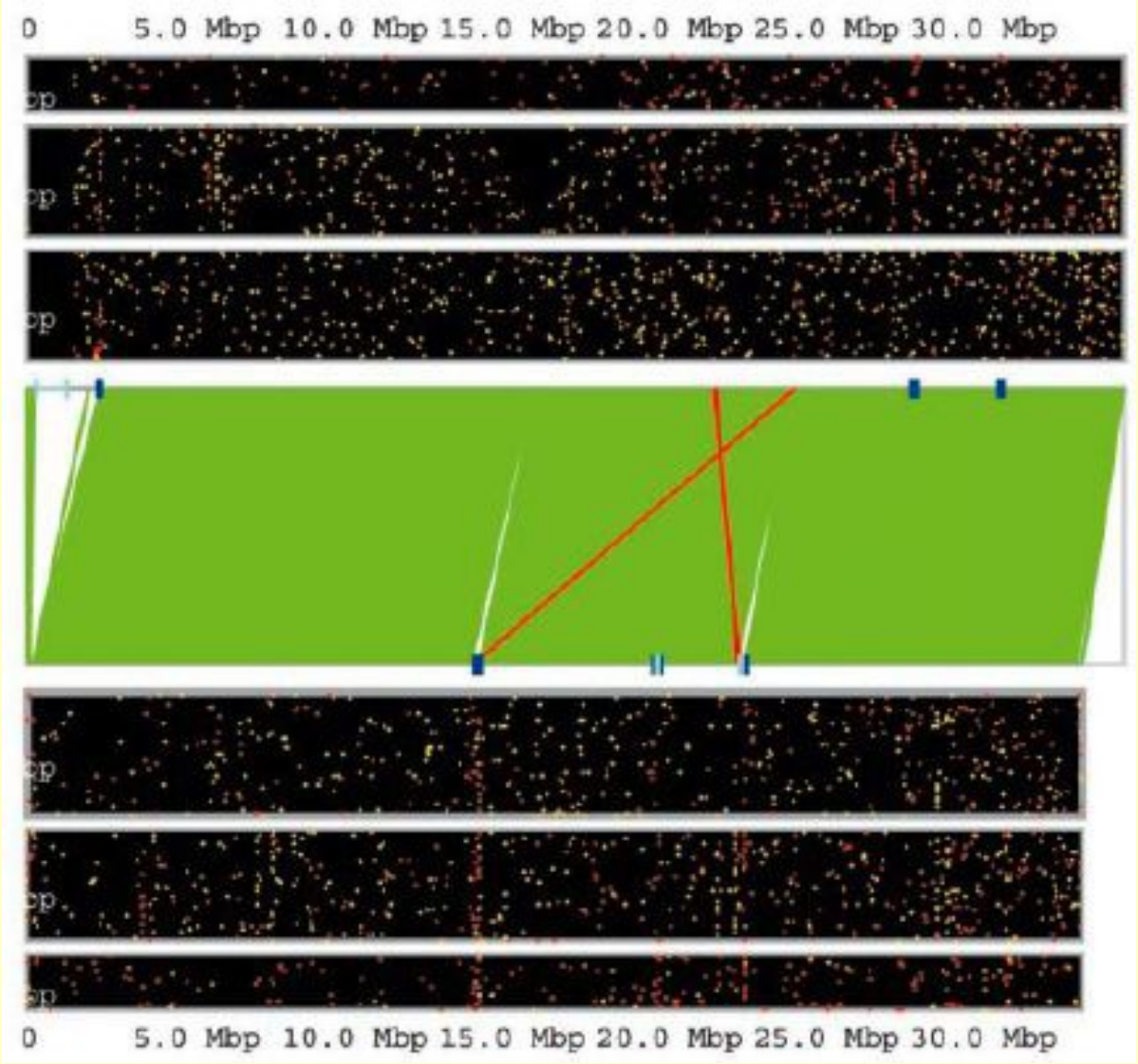
Genomics, Bioinformatics & Medicine



Doug Brutlag
Professor Emeritus of Biochemistry & Medicine
Stanford University School of Medicine

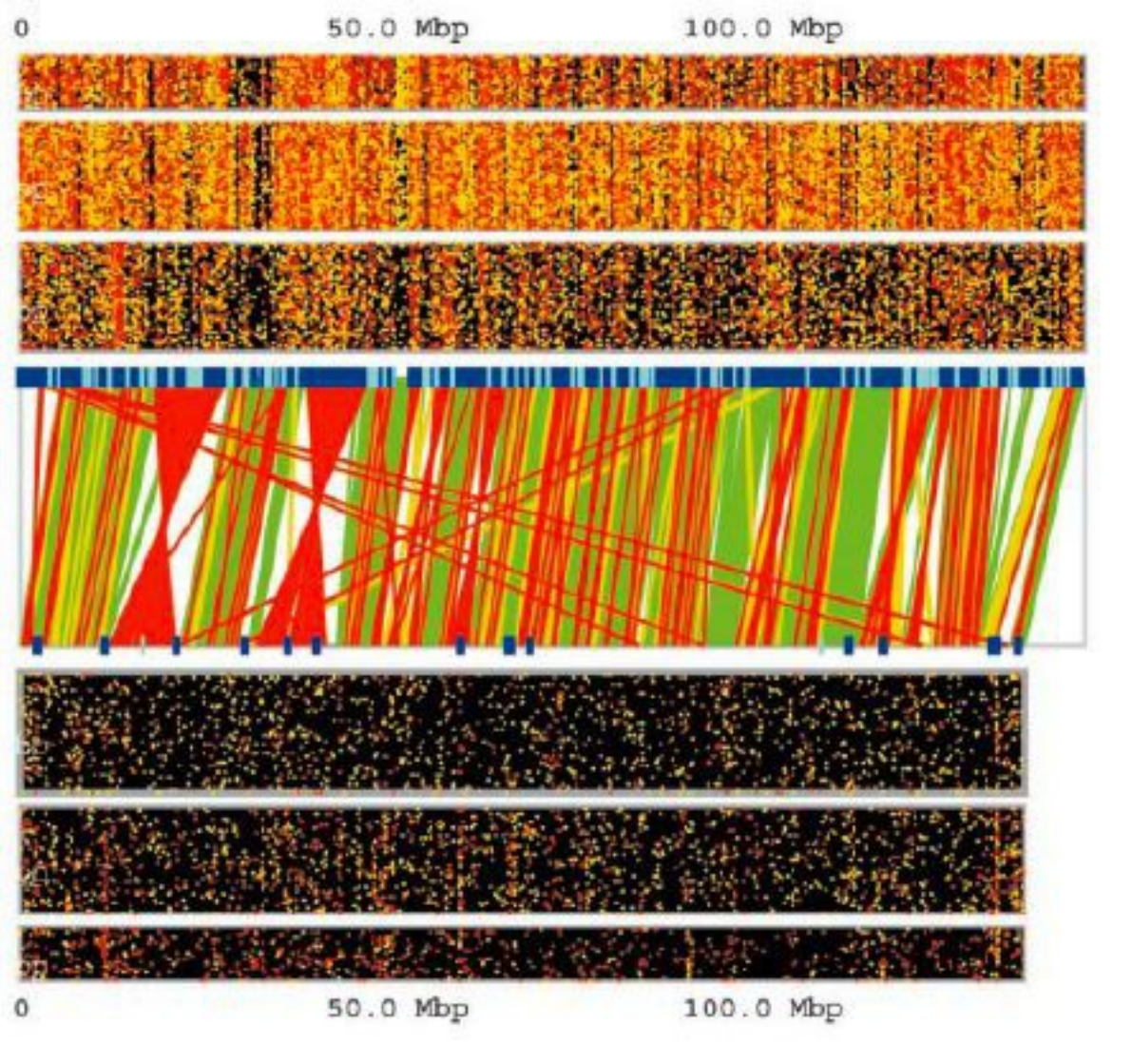
© Doug Brutlag 2015

Chromosome 21:
Public vs Whole Genome Shotgun Assemblies



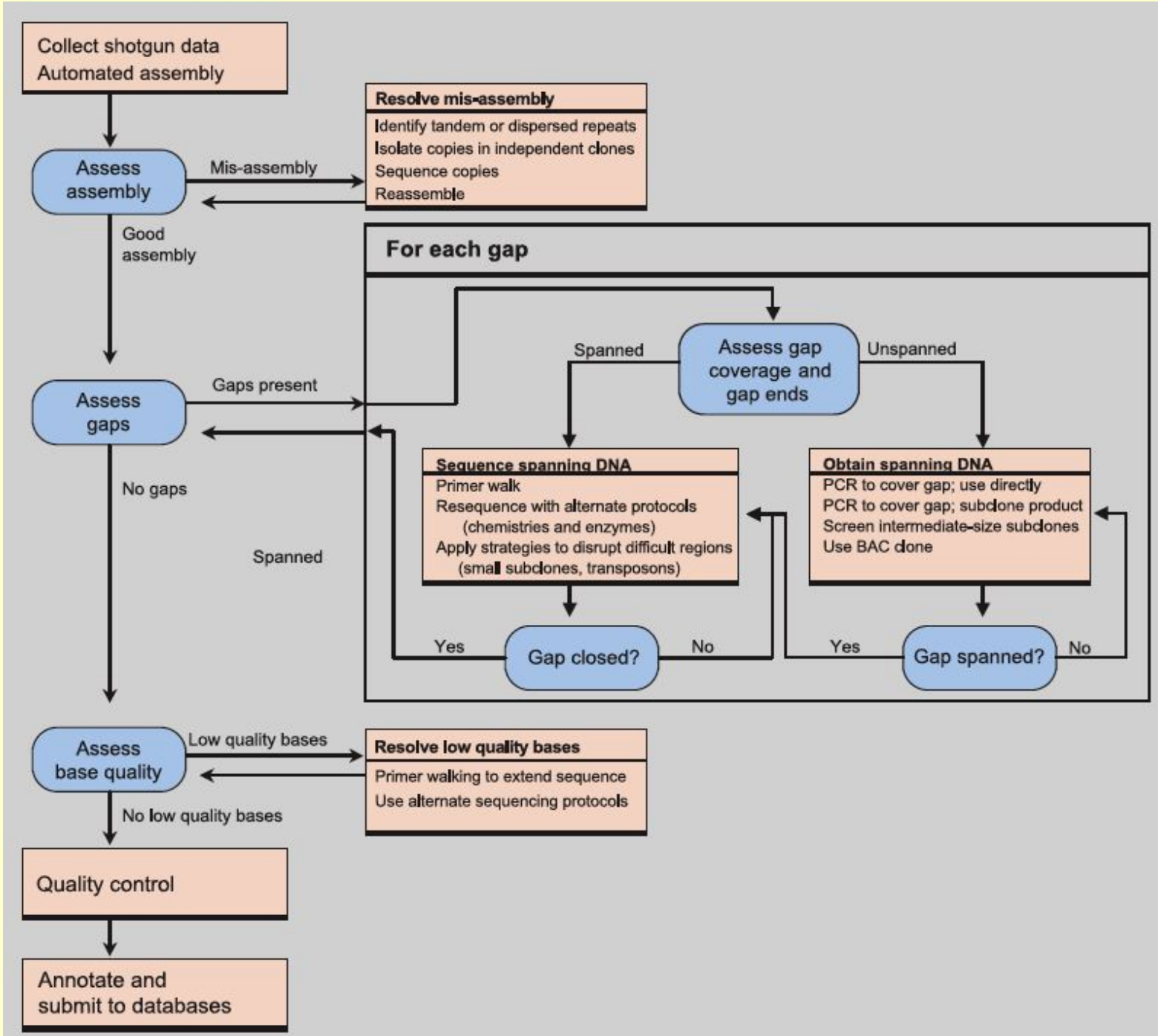
© Doug Brutlag 2015

Chromosome 8:
Public vs Whole Genome Shotgun Assemblies

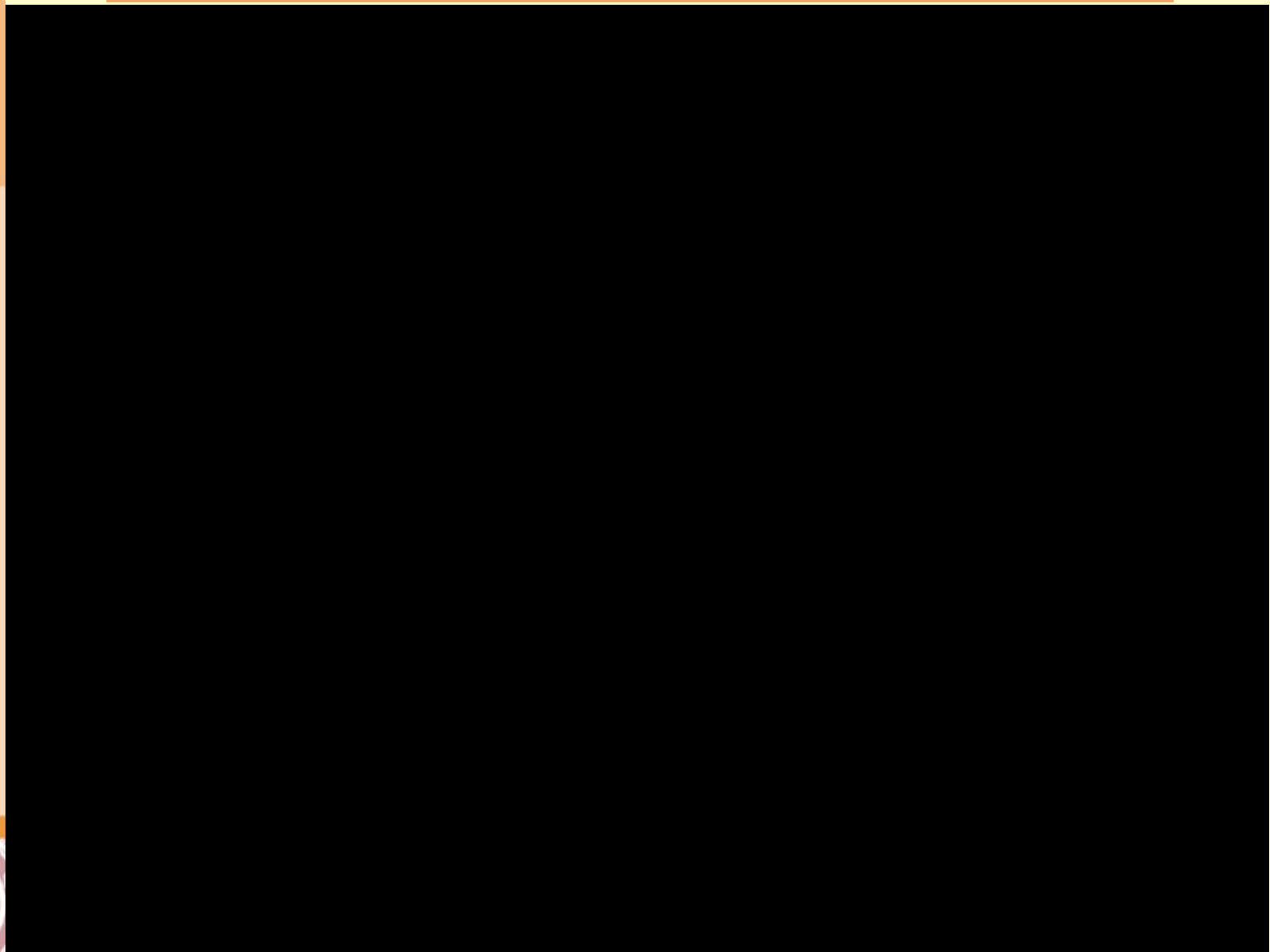


© Doug Brutlag 2015

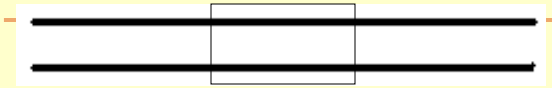
Finishing Strategy for the Public Genome Project



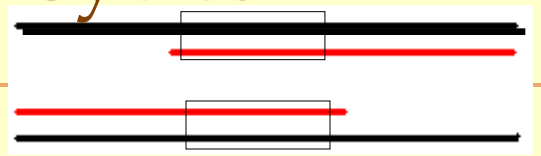
Polymerase Chain Reaction Overview: Exponential Amplification of DNA



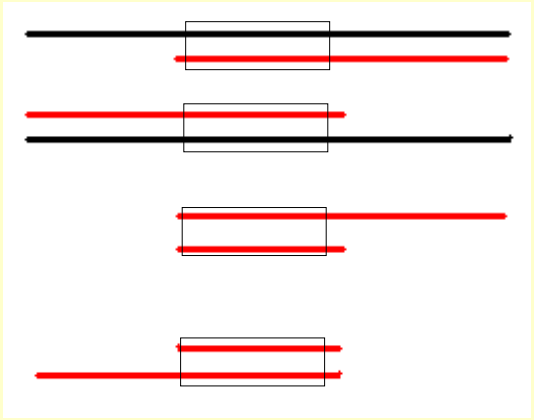
The First Three Cycles



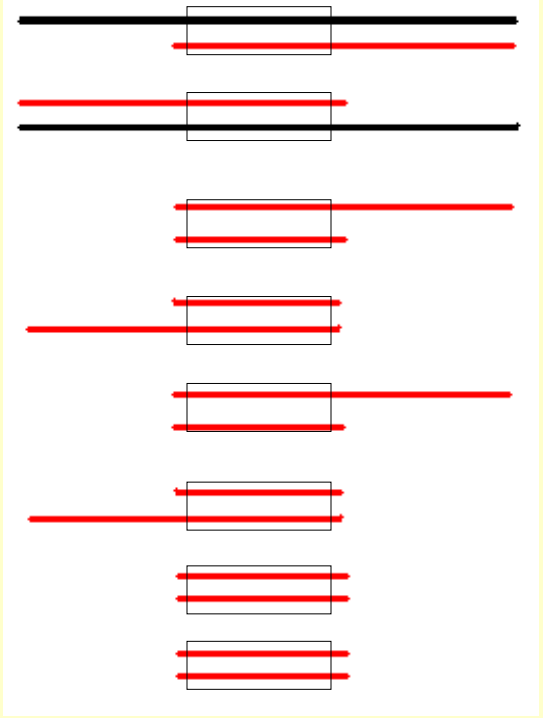
Original DNA



After Cycle 1



After Cycle 2



After Cycle 3

After N cycles, amount of target DNA is 2^N-2N

© Doug Brutlag 2015

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*

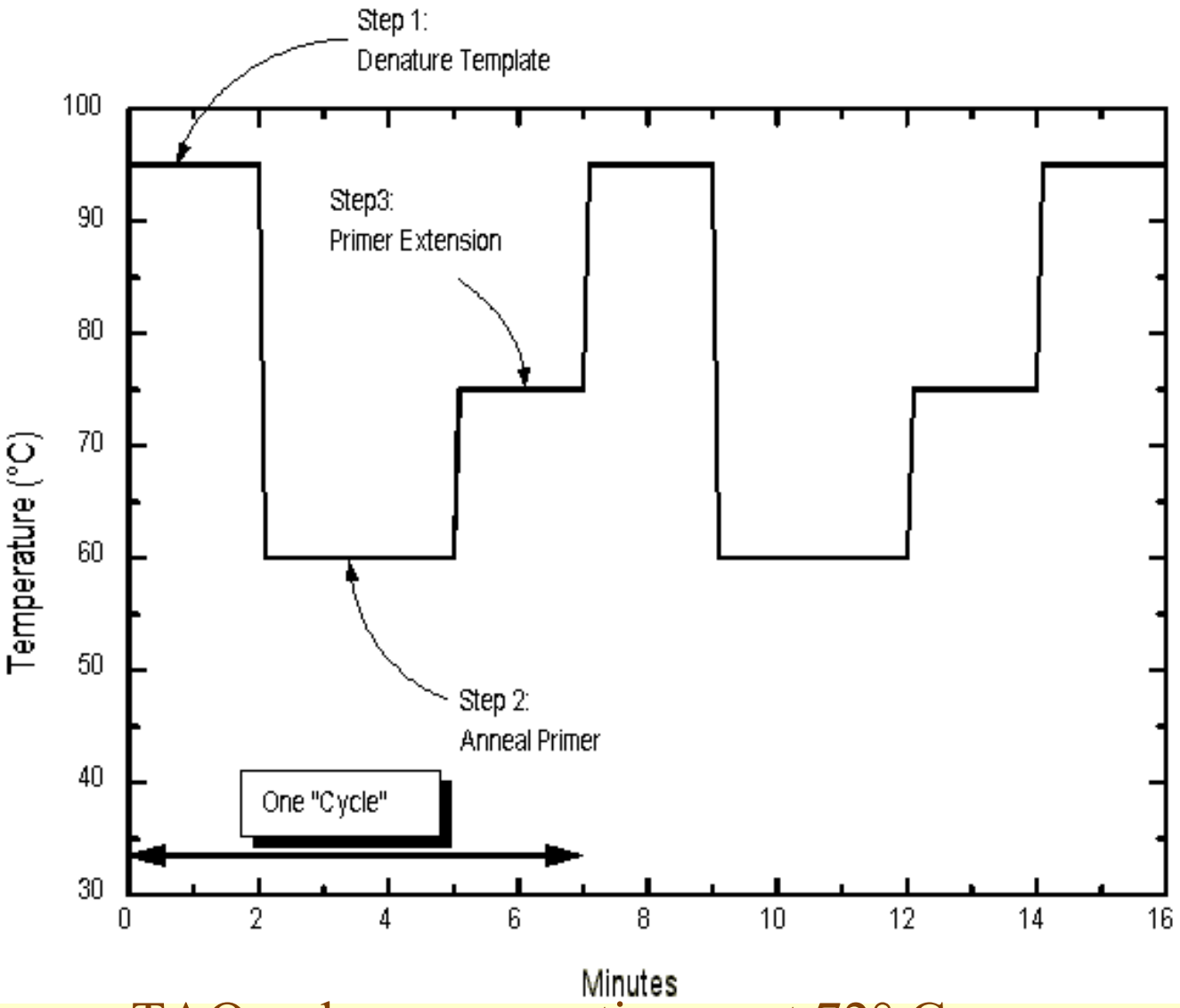
DNA Polymerase Enzyme

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

Thermocycler

© Doug Brutlag 2015

Temperature Cycling



TAQ polymerase optimum at 72° C

© Doug Brutlag 2015

PCR Applications

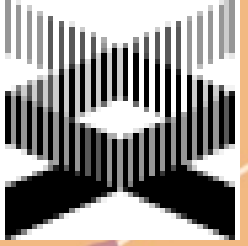
Forensics
- assessment/reassessment of crimes

Archaeology
- determine gene sequences of ancient organisms
- rethinking the past, human origins

Molecular Biology
- PCR ionstead of cloning genes
- Sequencing genes in final DNA synthetic step
- Finishing genome sequences filling in gaps
- Amplification of DNA or cDNA

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

© Doug Brutlag 2015



Finished Sequence in 2004 (Build 35)

Table 2 **Finished sequence and gaps, HGSC Build 35**

Chr	Total finished sequence* (kb)	Euchromatic gaps†		Heterochromatic gaps‡		Estimate of total gap size§ (kb)	Unfinished clones‖	
		Number	Est. size (kb)	Number	Est. size (kb)		Number	Est. size (kb)
1	222,828	32	1,605	2	19,510	21,115	17	850
2	237,503	20	2,512	1	2,900	5,412	0	0
3	194,636	5	1,935	1	1,500	3,435	0	0
4	187,161	14	1,250	1	3,000	4,250	0	0
5	177,703	5	92	1	340	432	0	0
6	167,318	10	658	1	2,300	2,958	0	0
7	154,759	11	869	1	4,630	5,499	0	0
8	142,613	9	662	1	2,190	2,852	0	0
9	117,781	40	1,955	2	18,000	19,955	12	600
10	131,614	12	1,020	1	2,515	3,535	8	400
11	131,131	7	322	1	4,760	5,082	0	0
12	130,259	8	795	1	4,300	5,095	0	0
13	95,560	6	715	2	17,200	17,915	0	0
14	88,291	1	8	2	17,220	17,228	0	0
15	81,342	10	737	2	18,260	18,997	0	0
16	78,885	4	143	2	10,000	10,143	0	0
17	77,800	9	875	1	7,500	8,375	0	0
18	74,656	3	97	1	1,368	1,465	0	0
19	55,786	5	5,015	1	340	5,355	0	0
20	59,505	4	1,157	1	1,766	2,923	0	0
21	34,170	3	53	2	11,620	11,673	0	0
22	34,765	11	460	2	14,330	14,790	0	0
X	150,394	12	750	1	3,000	3,750	14	700
Y	24,872	9	1,480	2	31,618	33,098	7	350
Total	2,851,331	250	25,165	33	200,167	225,332	58	2,900

*The total length of tiling paths including only finished bases of clones in Build 35. Roughly 2.19 Mb of sequence on chromosome Y was derived directly from the equivalent pseudoautosomal region on chromosome X.

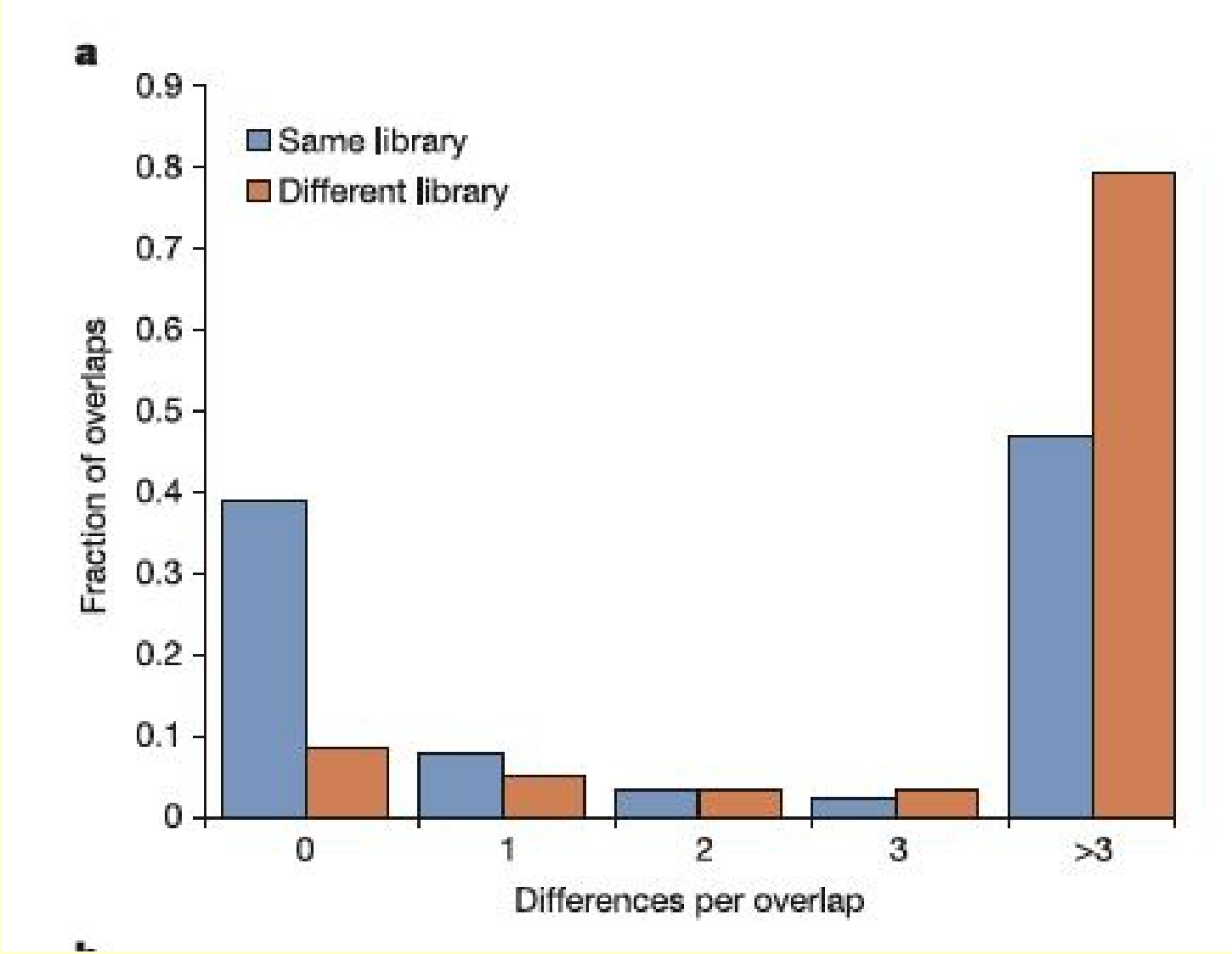
†Defined as gaps in euchromatic regions, including junctions with heterochromatic/centromeric sequences, for which no clone was available (see text).

‡Defined here as gaps in heterochromatic regions (see text and Supplementary Note 2 on heterochromatic sequence). Separate gaps were counted for centromeres and pericentric heterochromatin, even when the two were contiguous. Centromere sizes were taken from ref. 62 or in some cases provided directly by the sequencing centres (see Supplementary Note 2). Acrocentric sizes are based on centromere ratios from ref. 63. The sizes of large heterochromatic gaps are typically difficult to estimate accurately owing to their repeat structure and polymorphic nature[62,64]. Other regions might arguably be called heterochromatin (for example, the pericentric regions of chromosomes 19 and 3 and a ~400-kb gap on the Y chromosome[23]), but are classified as euchromatin here.

§The sum of lengths for finished sequence, estimated heterochromatic gaps, euchromatic gaps and unfinished clone gaps. The total length is only approximate because of uncertainty in gap sizes, particularly for heterochromatic gaps and centromeres.

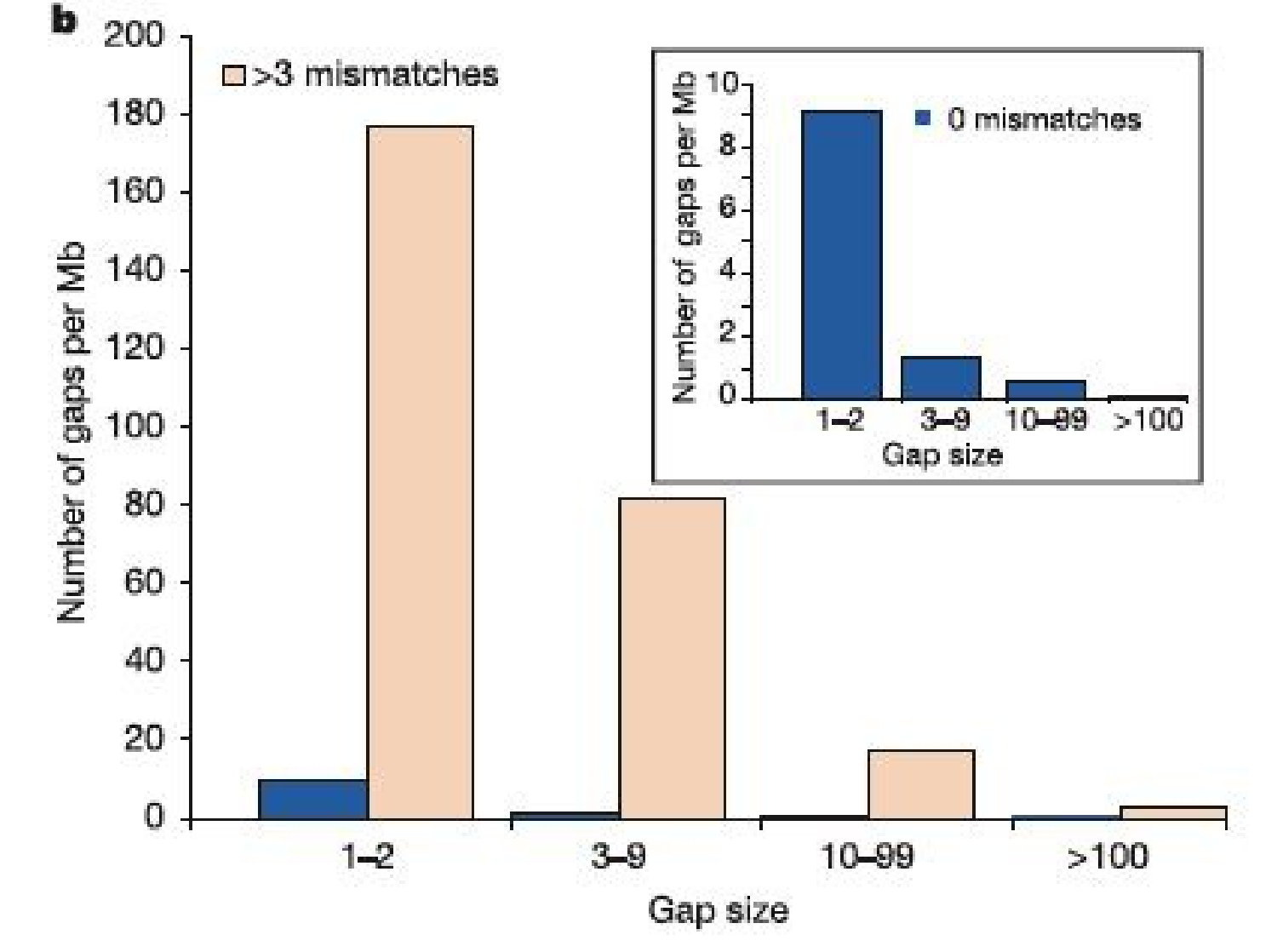
‖Those in the tiling path but for which it has not been possible to obtain finished sequence. Unfinished sequence from these clones is deposited in public databases. These gaps are all listed at 50 kb, reflecting the approximate average size of the gap.

Substitutions in BAC Overlaps with BACs



© Doug Brutlag 2015

Gaps in BAC Overlaps with BACs from Same or Different Libraries



© Doug Brutlag 2015

Duplications and Deletions in the Human Genome

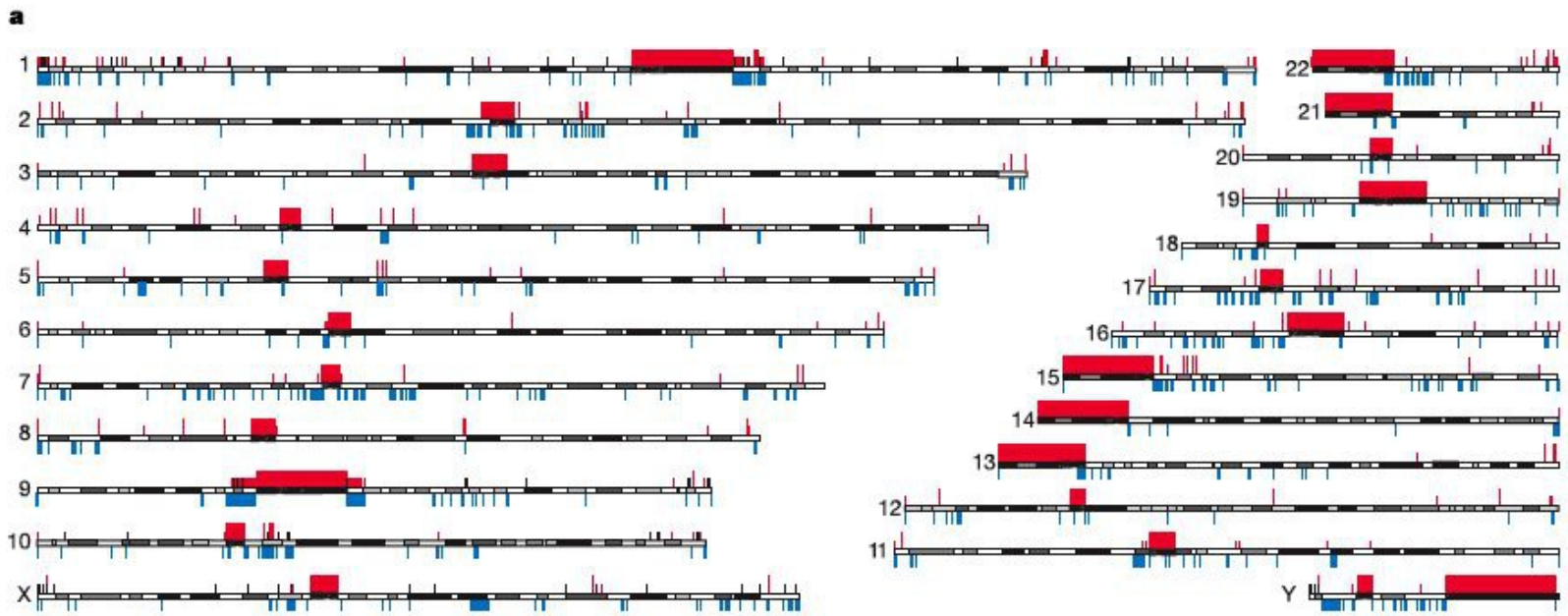
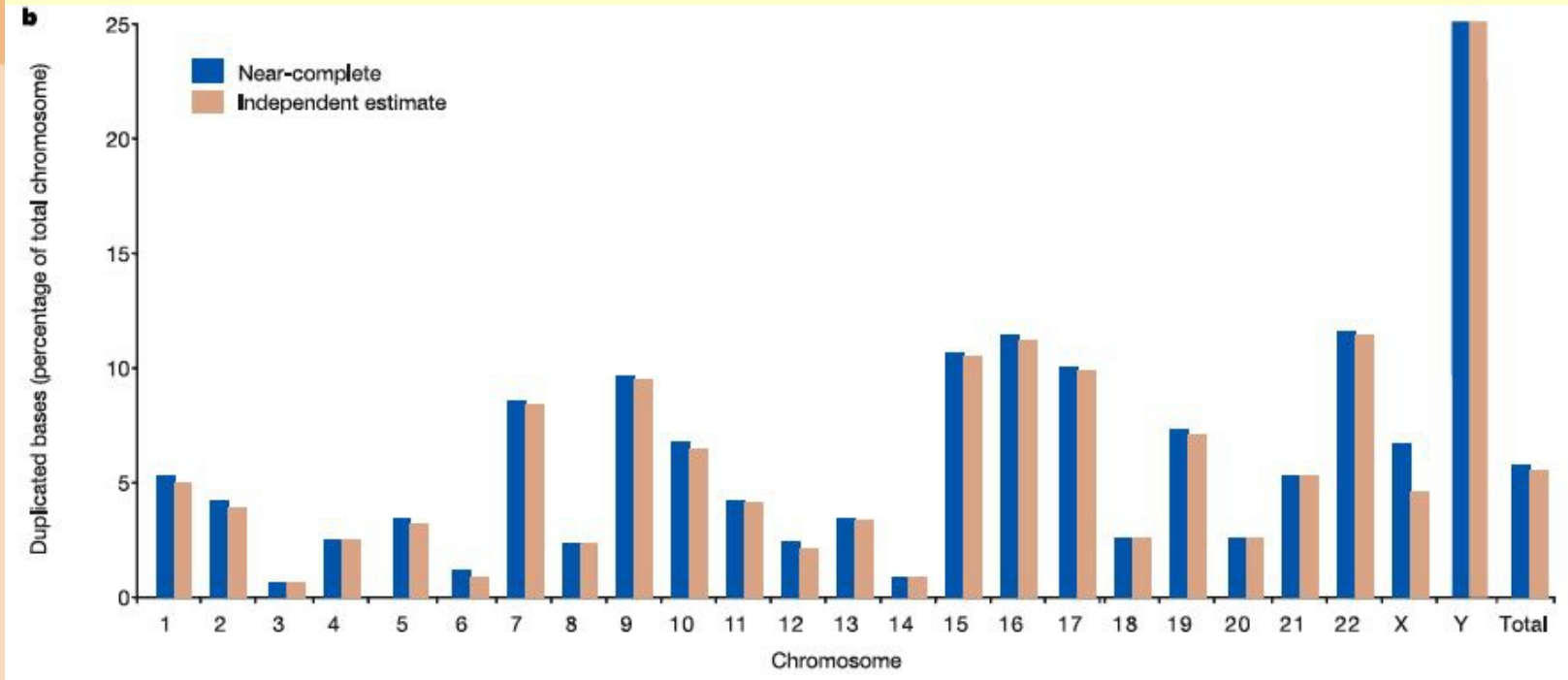


Figure 4 Segmental duplications across the genome. **a**, Segmental duplications and sequence gaps across the genome. Segmental duplications are indicated below the chromosomes in blue (length ≥10 kb and sequence identity ≥95%). Large duplications are shown to approximate scale; smaller ones are indicated 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

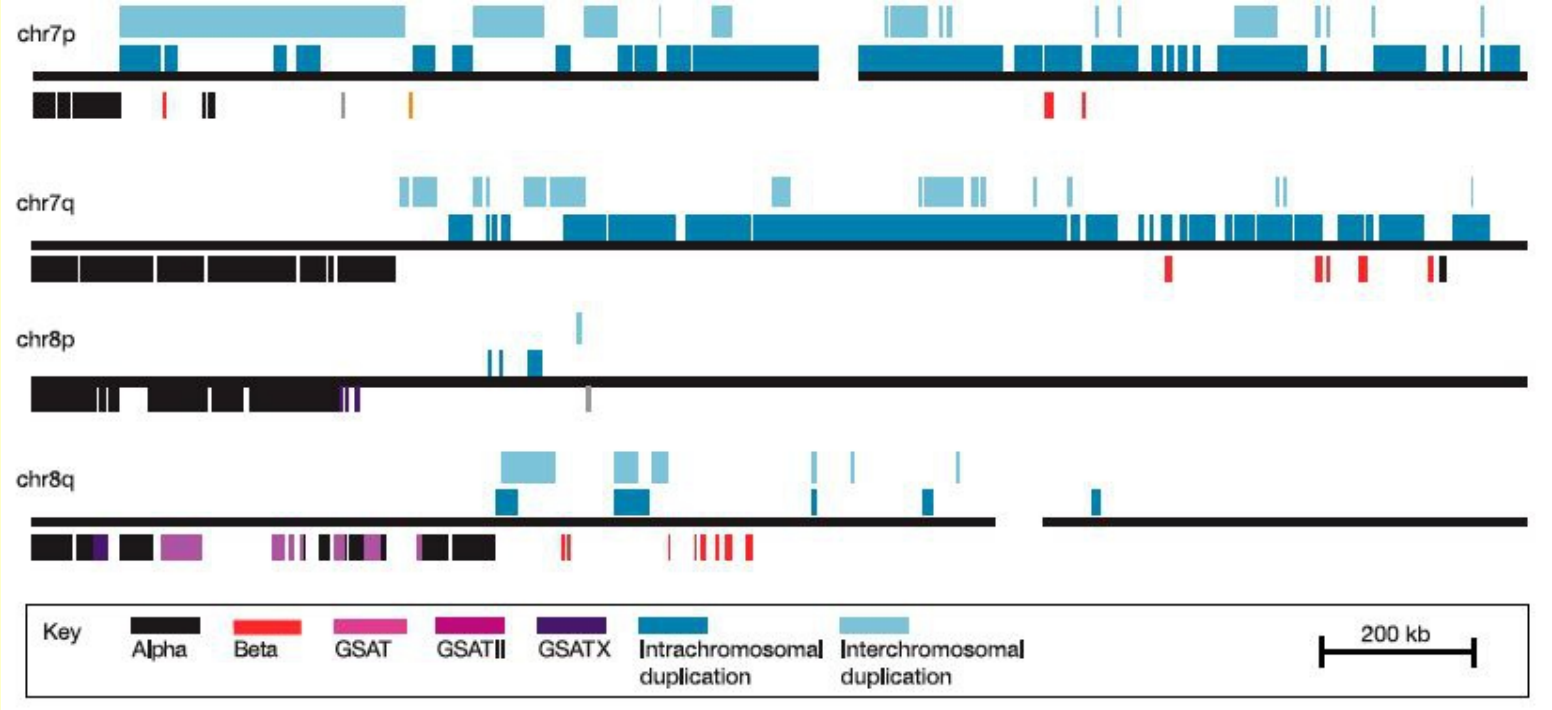
© Doug Brutlag 2015

Percentage of Chromosomes Duplicated



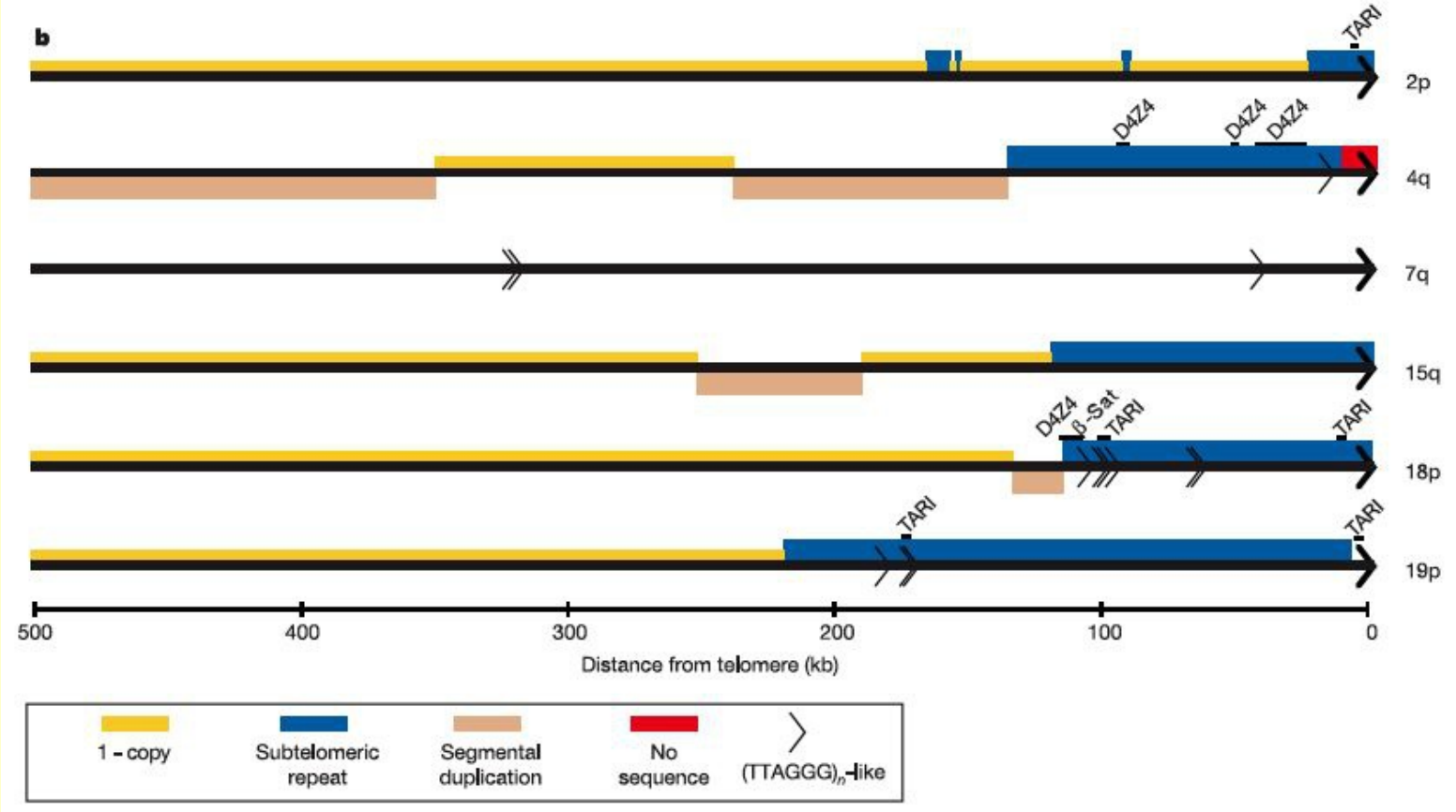
© Doug Brutlag 2015

Duplications near Centromeres



© Doug Brutlag 2015

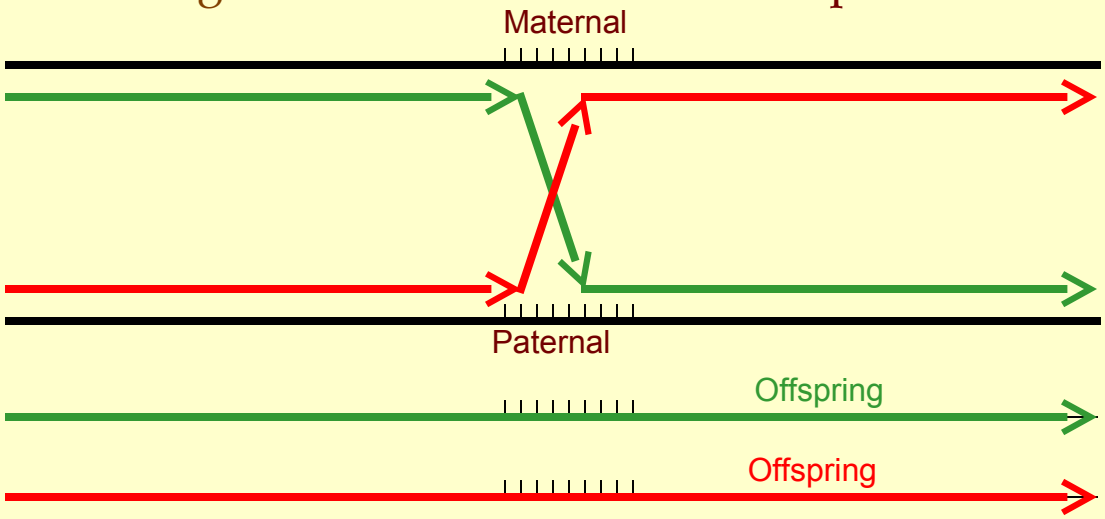
Duplications near Telomeres



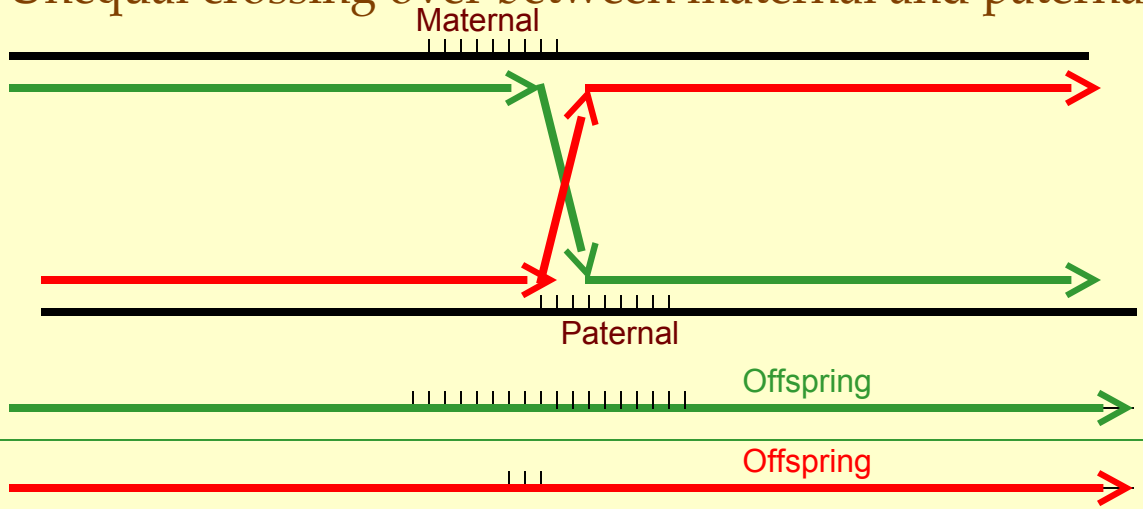
© Doug Brutlag 2015

Deletions and Duplications can Arise from Unequal Crossing Over in Repeated Regions

- Crossing over between maternal and paternal chromosomes



- Unequal crossing over between maternal and paternal chromosomes



© Doug Brutlag 2015

The Diploid Sequence of an Individual Human (HuRef)

The Diploid Genome Sequence of an Individual Human

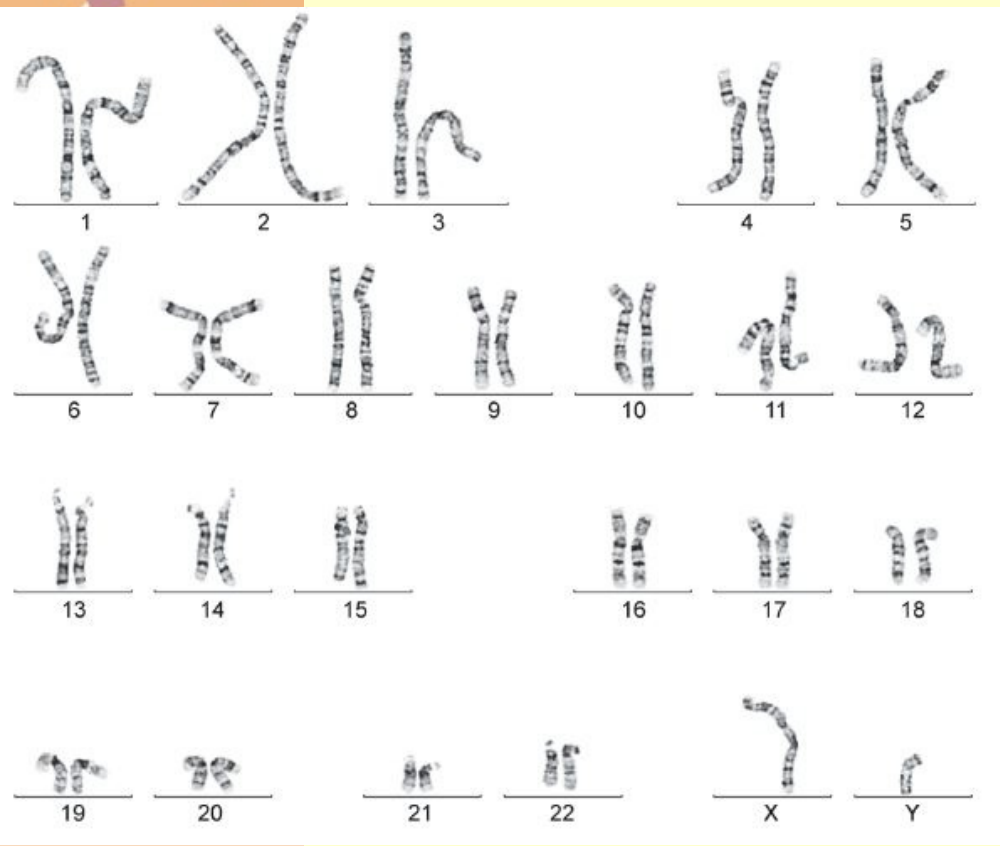
Samuel Levy[1*], Granger Sutton[1], Pauline C. Ng[1], Lars Feuk[2], Aaron L. Halpern[1], Brian P. Walenz[1], Nelson Axelrod[1], Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1], Yuan Lin[1], Jeffrey R. MacDonald[2], Andy Wing Chun Pang[2], Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamouri[1], Vineet Bafna[3], Vikas Bansal[3], Saul A. Kravitz[1], Dana A. Busam[1], Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remington[1], Josep F. Abril[4], John Gill[1], Jon Borman[1], Yu-Hui Rogers[1], Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. Strausberg[1], J. Craig Venter[1]

1 J. Craig Venter Institute, Rockville, Maryland, United States of America, 2 Program in Genetics and Genomic Biology, The Hospital for Sick Children, and Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada, 3 Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, United States of America, 4 Genetics Department, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain

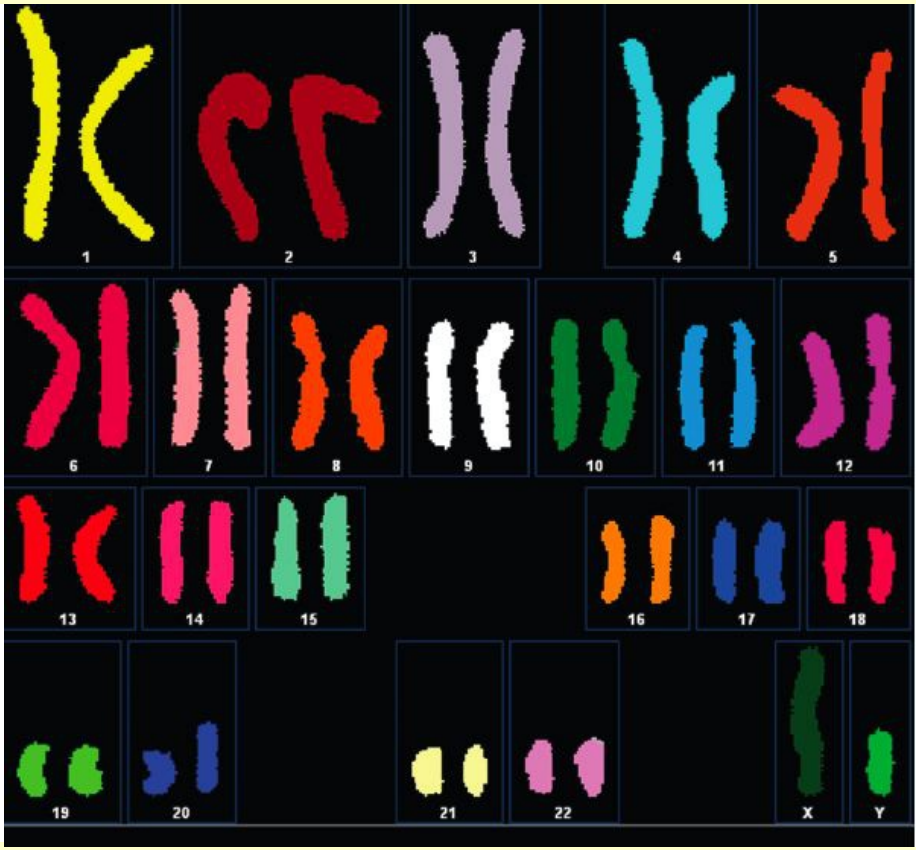
Presented here is a genome sequence of an individual human. It was produced from ~32 million random DNA fragments, sequenced by Sanger dideoxy technology and assembled into 4,528 scaffolds, comprising 2,810 million bases (Mb) of contiguous sequence with approximately 7.5-fold coverage for any given region. We developed a modified version of the Celera assembler to facilitate the identification and comparison of alternate alleles within this individual diploid genome. Comparison of this genome and the National Center for Biotechnology Information human reference assembly revealed more than 4.1 million DNA variants, encompassing 12.3 Mb. These variants (of which 1,288,319 were novel) included 3,213,401 single nucleotide polymorphisms (SNPs), 53,823 block substitutions (2–206 bp), 292,102 heterozygous insertion/deletion events (indels)(1–571 bp), 559,473 homozygous indels (1–82,711 bp), 90 inversions, as well as numerous segmental duplications and copy number variation regions. Non-SNP DNA variation accounts for 22% of all events identified in the donor, however they involve 74% of all variant bases. This suggests an important role for non-SNP genetic alterations in defining the diploid genome structure. Moreover, 44% of genes were heterozygous for one or more variants. Using a novel haplotype assembly strategy, we were able to span 1.5 Gb of genome sequence in segments >200 kb, providing further precision to the diploid nature of the genome. These data depict a definitive molecular portrait of a diploid human genome that provides a starting point for future genome comparisons and enables an era of individualized genomic information.

© Doug Brutlag 2015

Karyotype of J.Craig Venter



Giemsa Stain



FISH Stain

© Doug Brutlag 2015

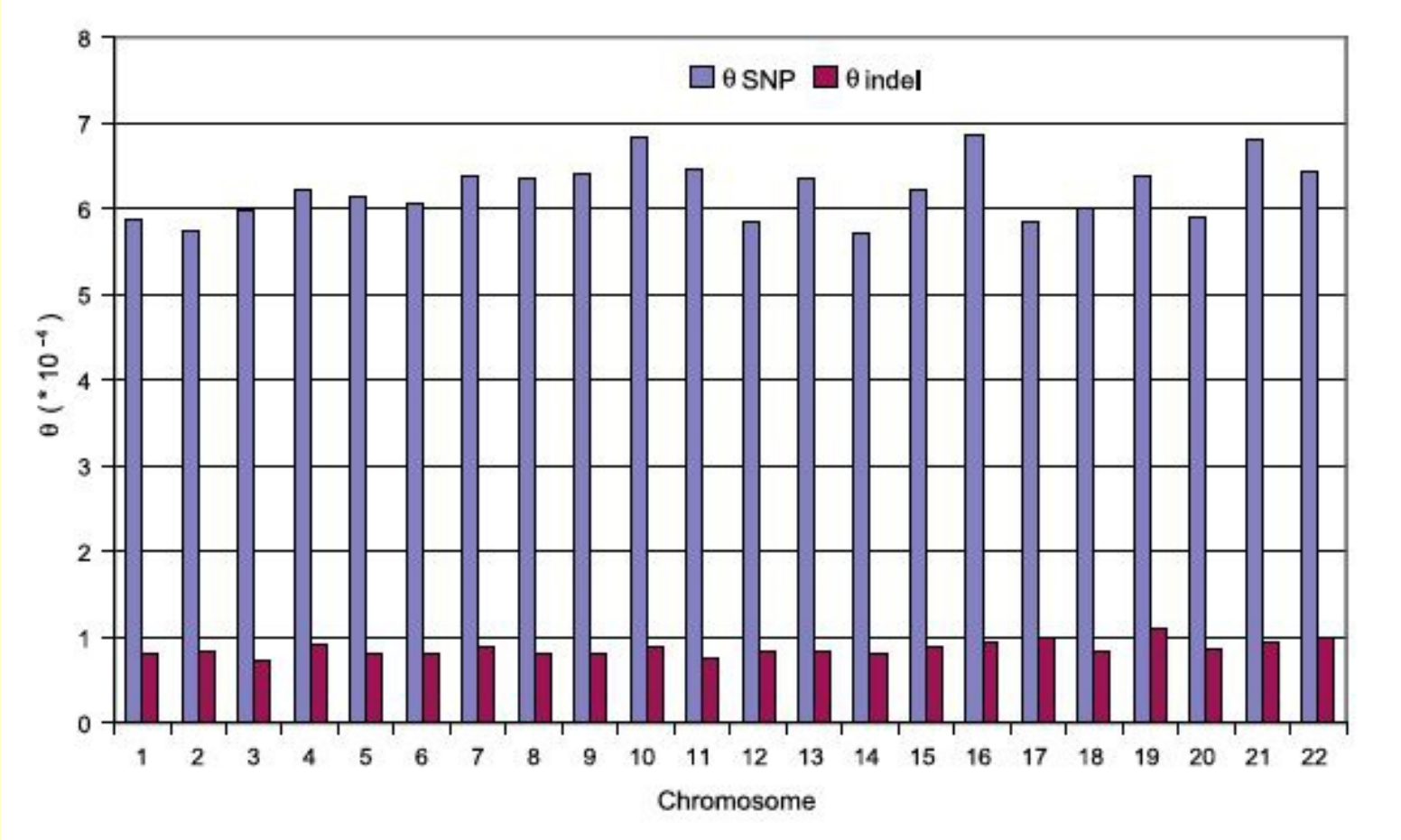
Comparing Reference Genome Assembly to HuR

Table 2. Summary of HuRef Assembly Statistics and Comparison to the Human NCBI Genome

Assembly	Assembly Subset	Number of Scaffolds	Number of Contigs	Gaps within Scaffolds	ACGT Bases	Span
NCBI Chromosomes	N/A	279	N/A	N/A	2,858,012,806	3,080,419,480
NCBI All	N/A	367	N/A	N/A	2,870,607,502	3,093,104,542
WGSA Chromosomes	N/A	4,940	211,493	206,553	2,659,468,408	2,993,154,503
HuRef Assembly	Chromosomes	1,408	66,762	66,354	2,782,357,138	2,809,547,336
	Scaffolds $\geq$ 100 kb	553	65,932	65,379	2,779,929,229	2,806,091,853
	Scaffolds $\geq$ 3 kb	4,528	71,343	66,815	2,809,774,459	2,844,046,670
	All scaffolds	188,394	255,300	66,906	3,002,932,476	3,037,726,076

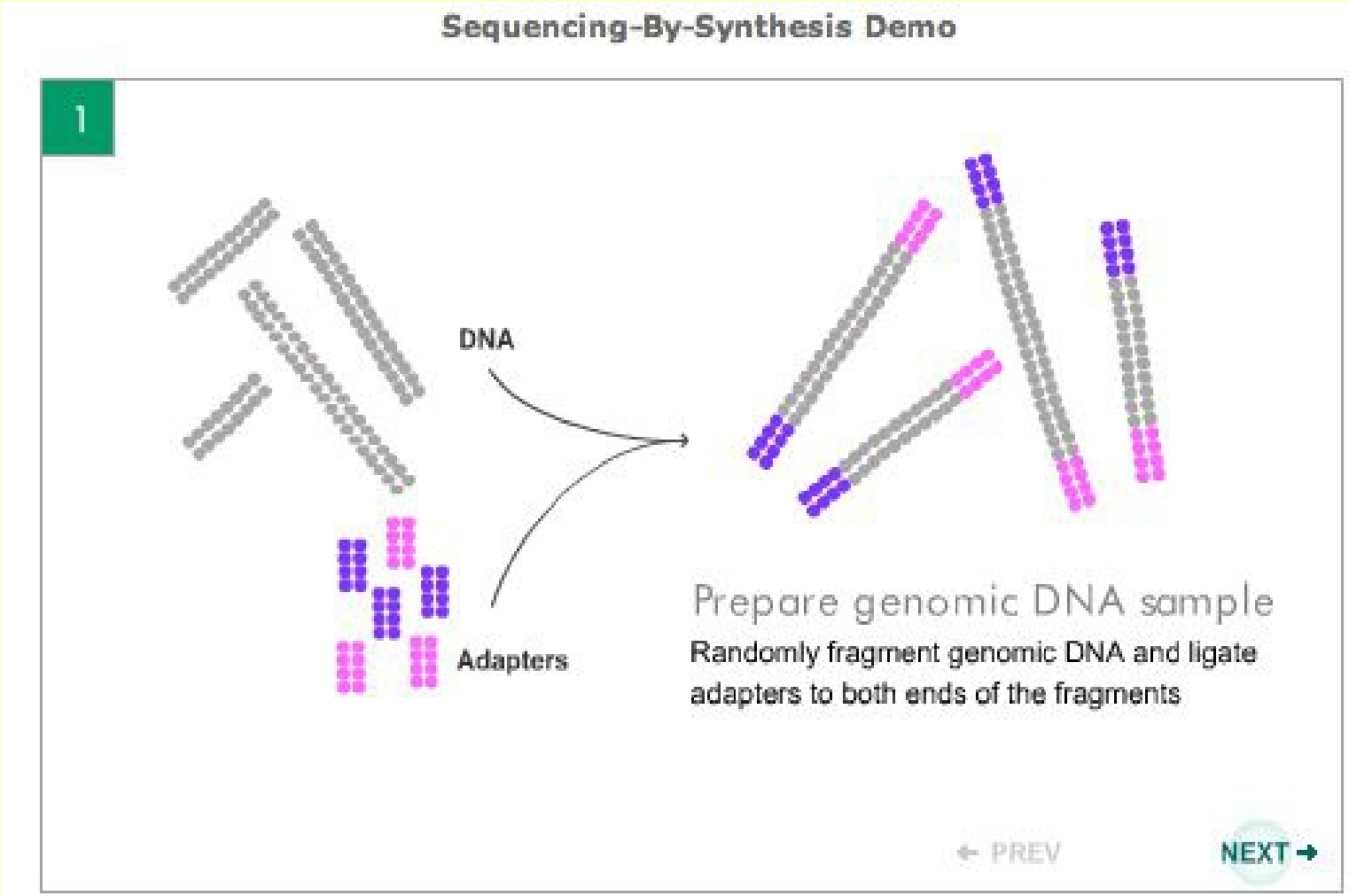
© Doug Brutlag 2015

SNPs & InDels in HuRef Autosomes



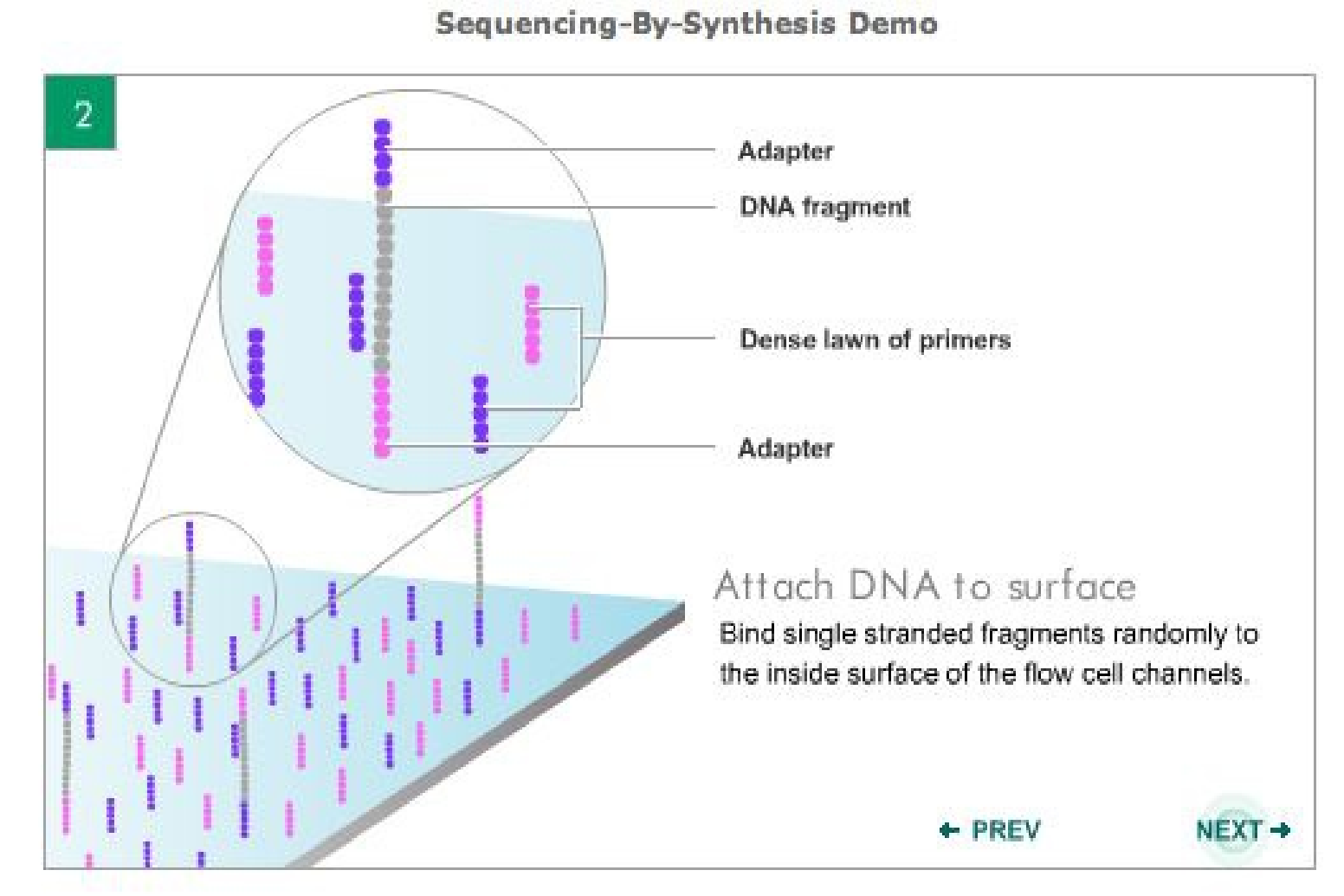
© Doug Brutlag 2015

Illumina Solexa Sequencing Technology



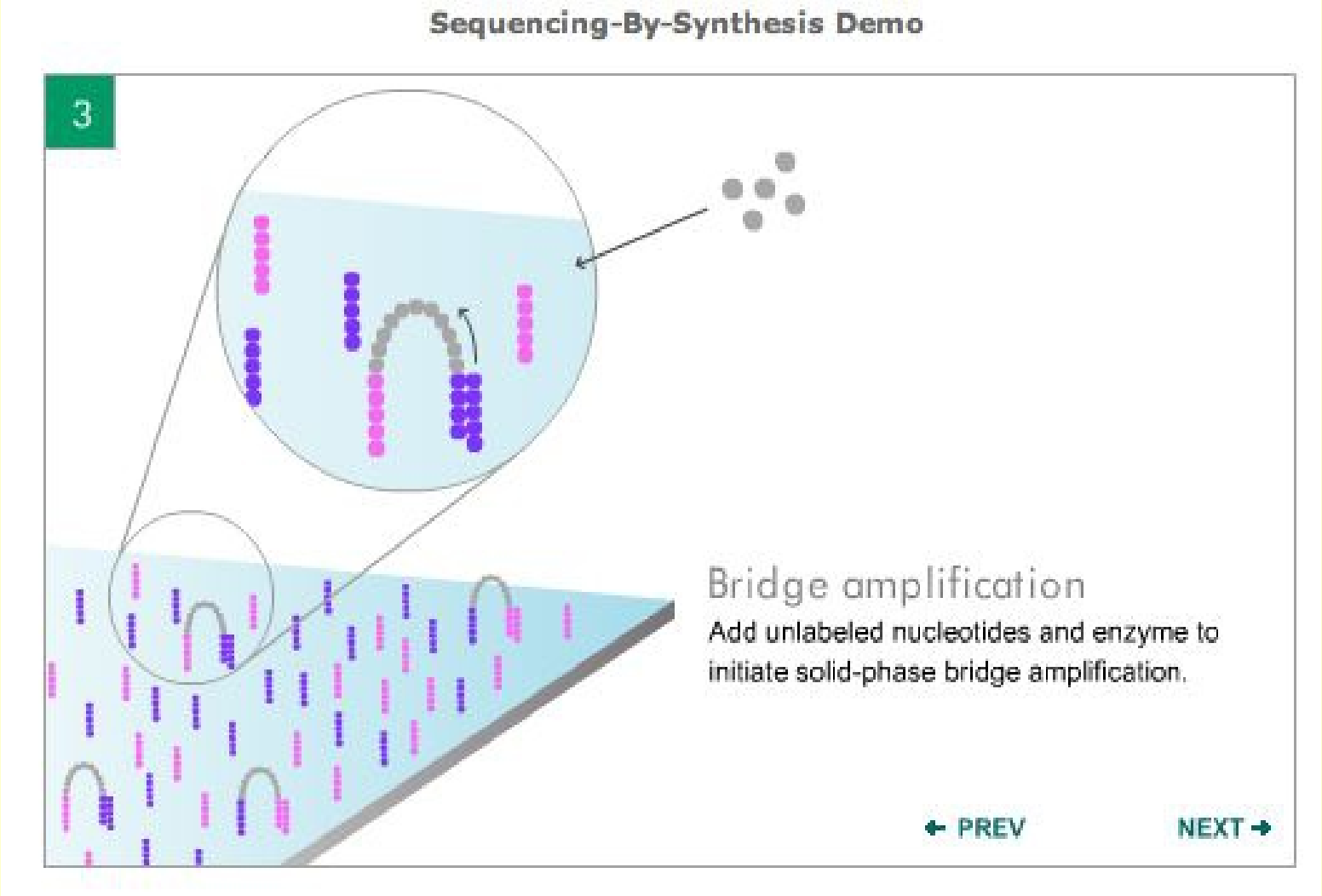
© Doug Brutlag 2015

Illumina Solexa Sequencing Technology



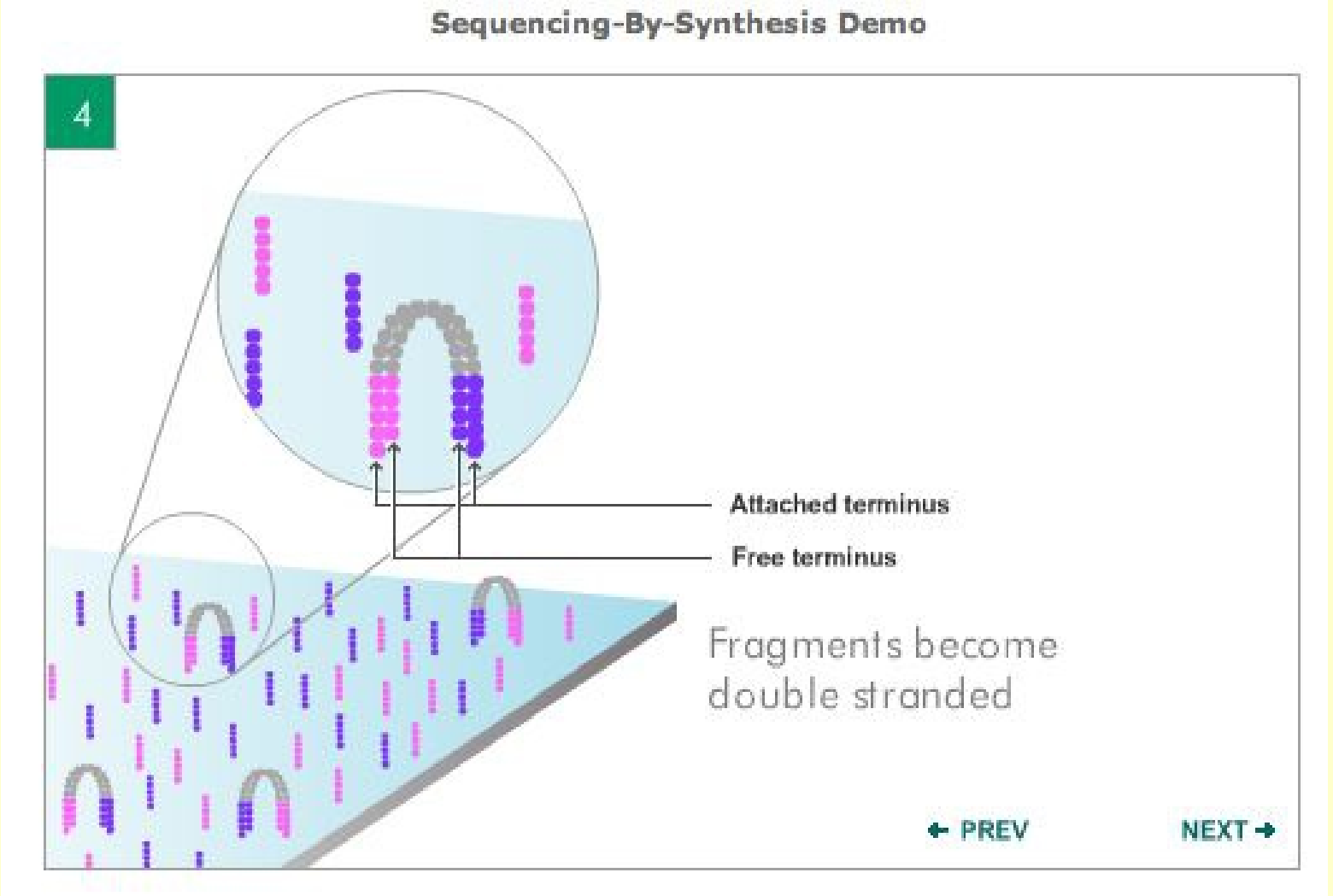
© Doug Brutlag 2015

Illumina Solexa Sequencing Technology



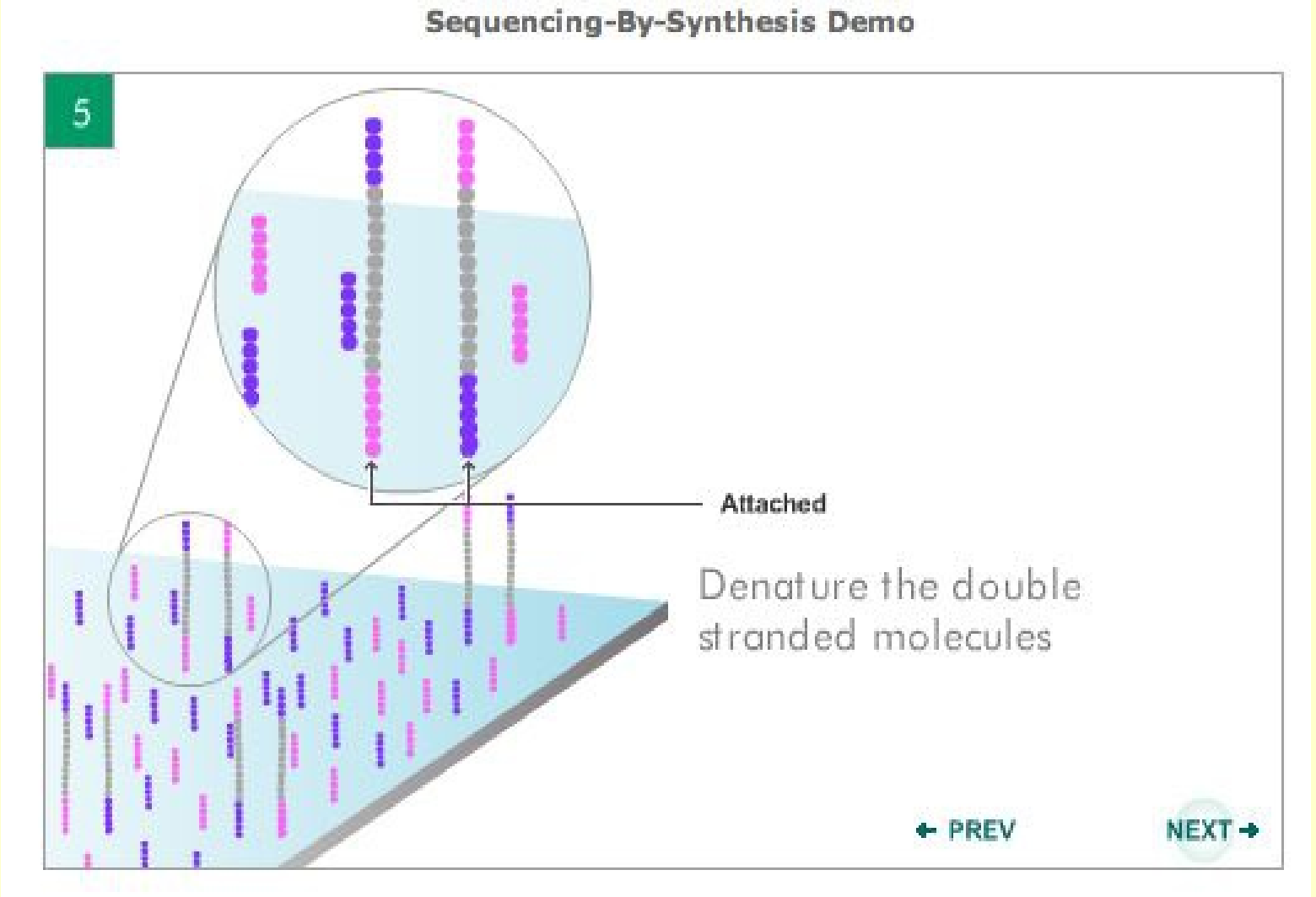
© Doug Brutlag 2015

Illumina Solexa Sequencing Technology



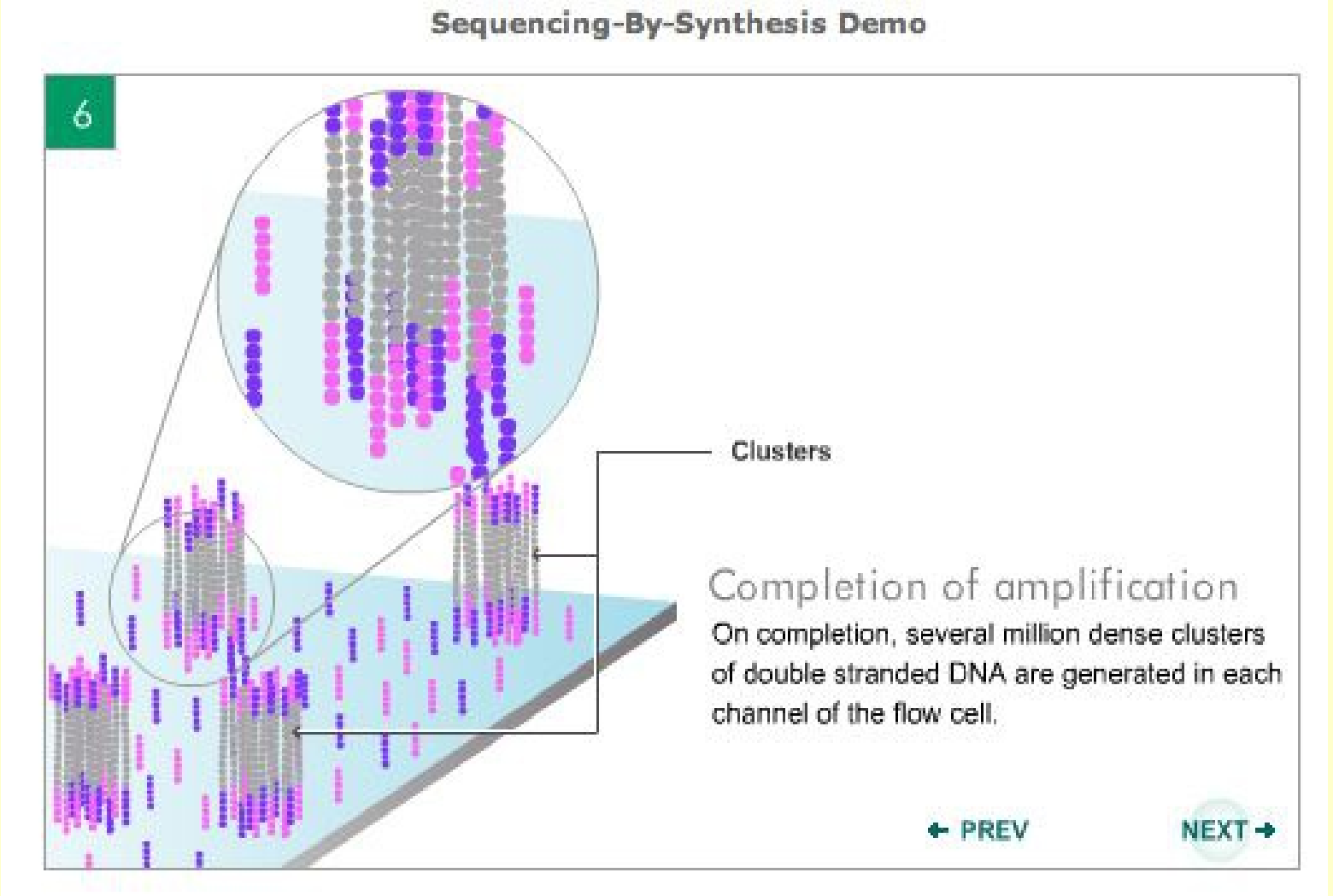
© Doug Brutlag 2015

Illumina Solexa Sequencing Technology



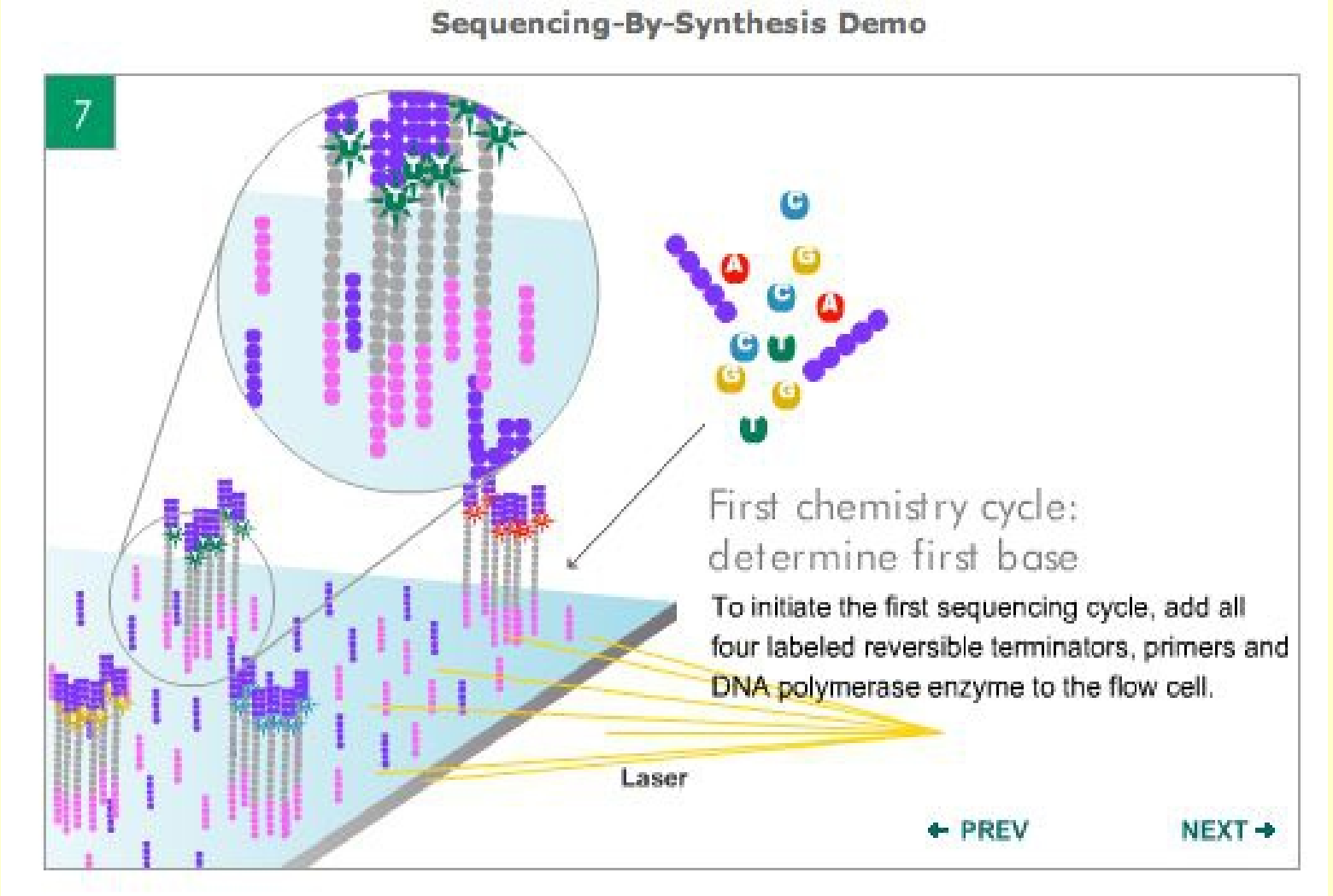
© Doug Brutlag 2015

Illumina Solexa Sequencing Technology



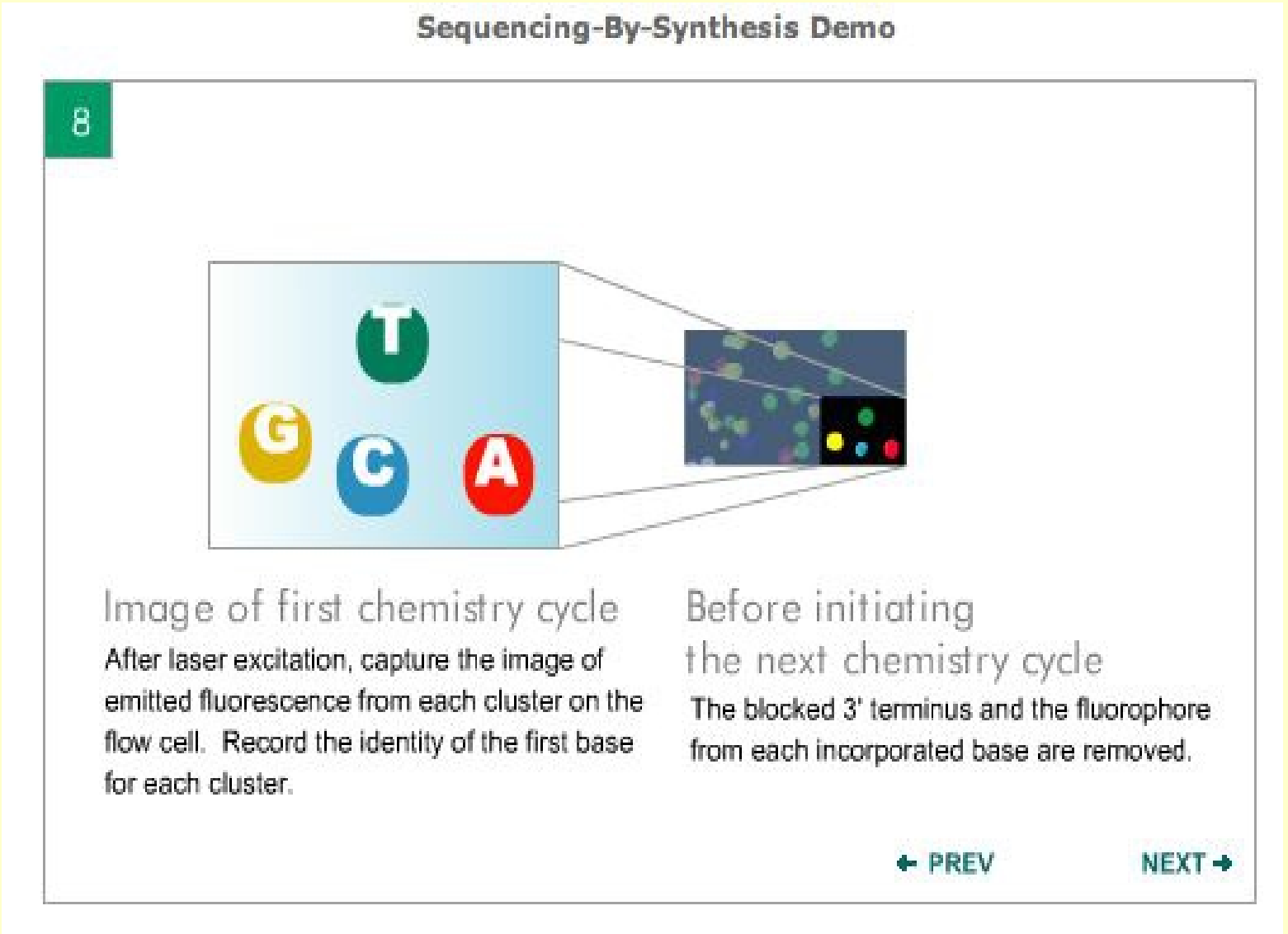
© Doug Brutlag 2015

Illumina Solexa Sequencing Technology



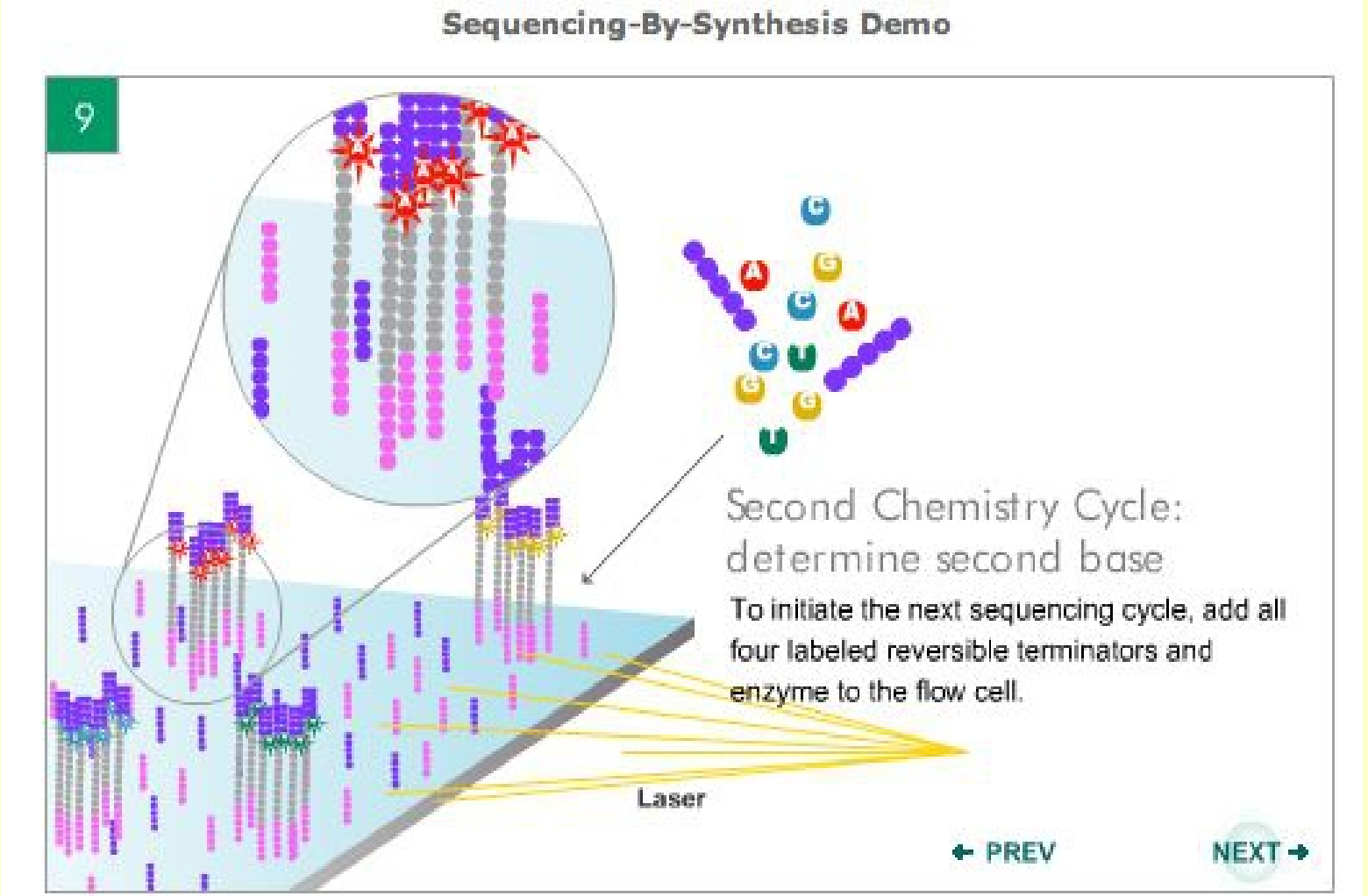
© Doug Brutlag 2015

Illumina Solexa Sequencing Technology



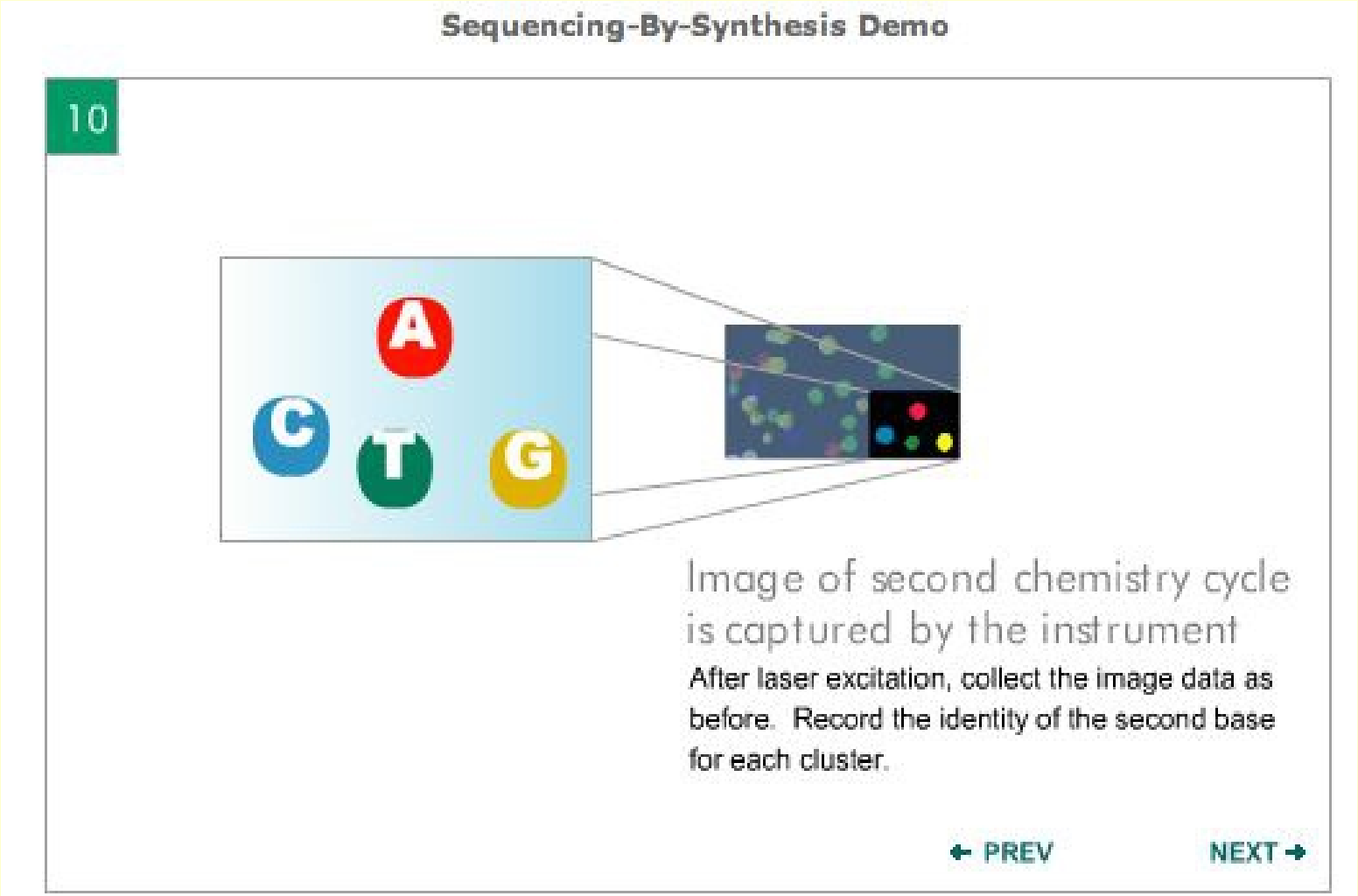
© Doug Bradag 2015

Illumina Solexa Sequencing Technology



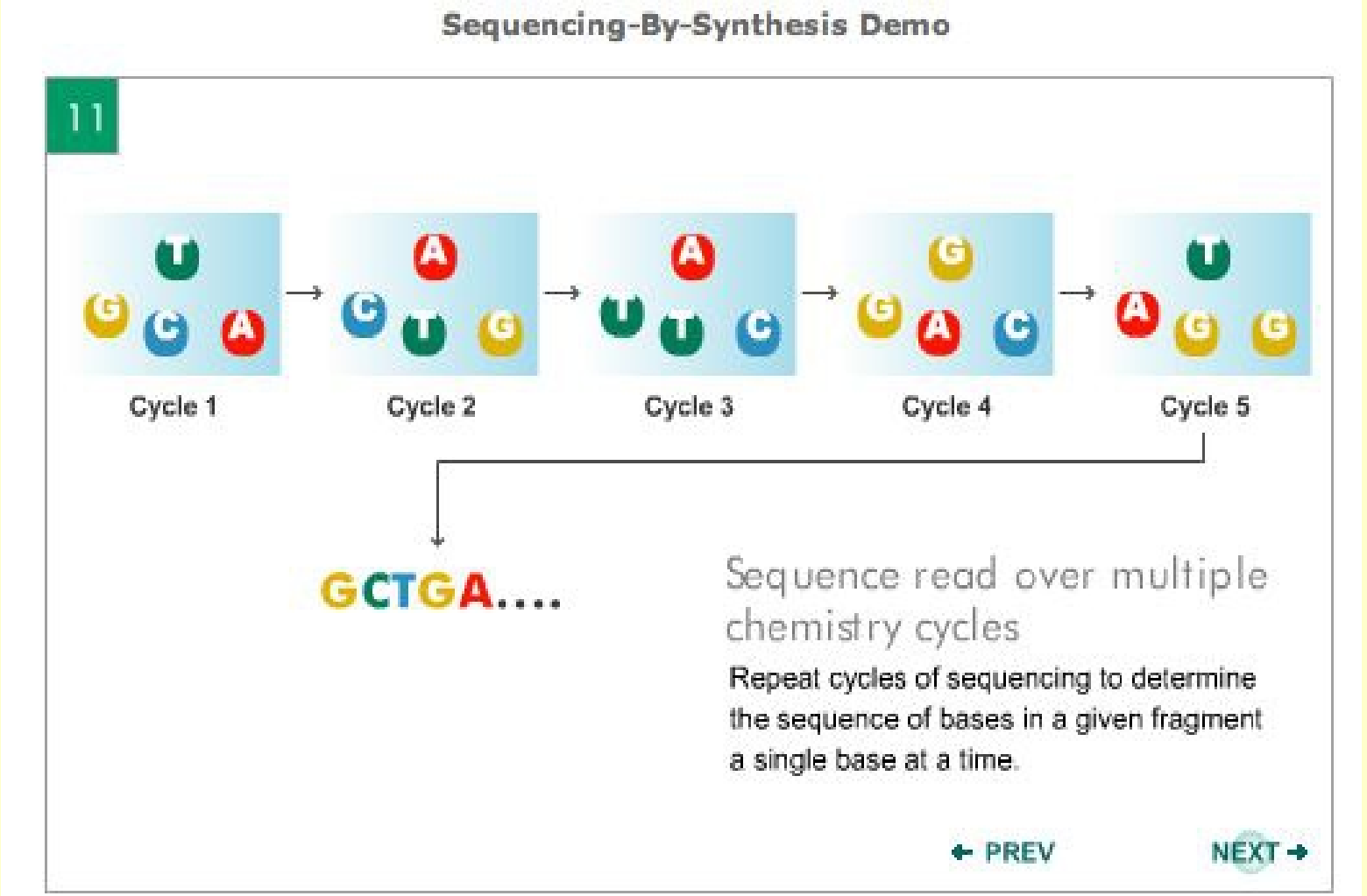
© Doug Brutlag 2015

Illumina Solexa Sequencing Technology



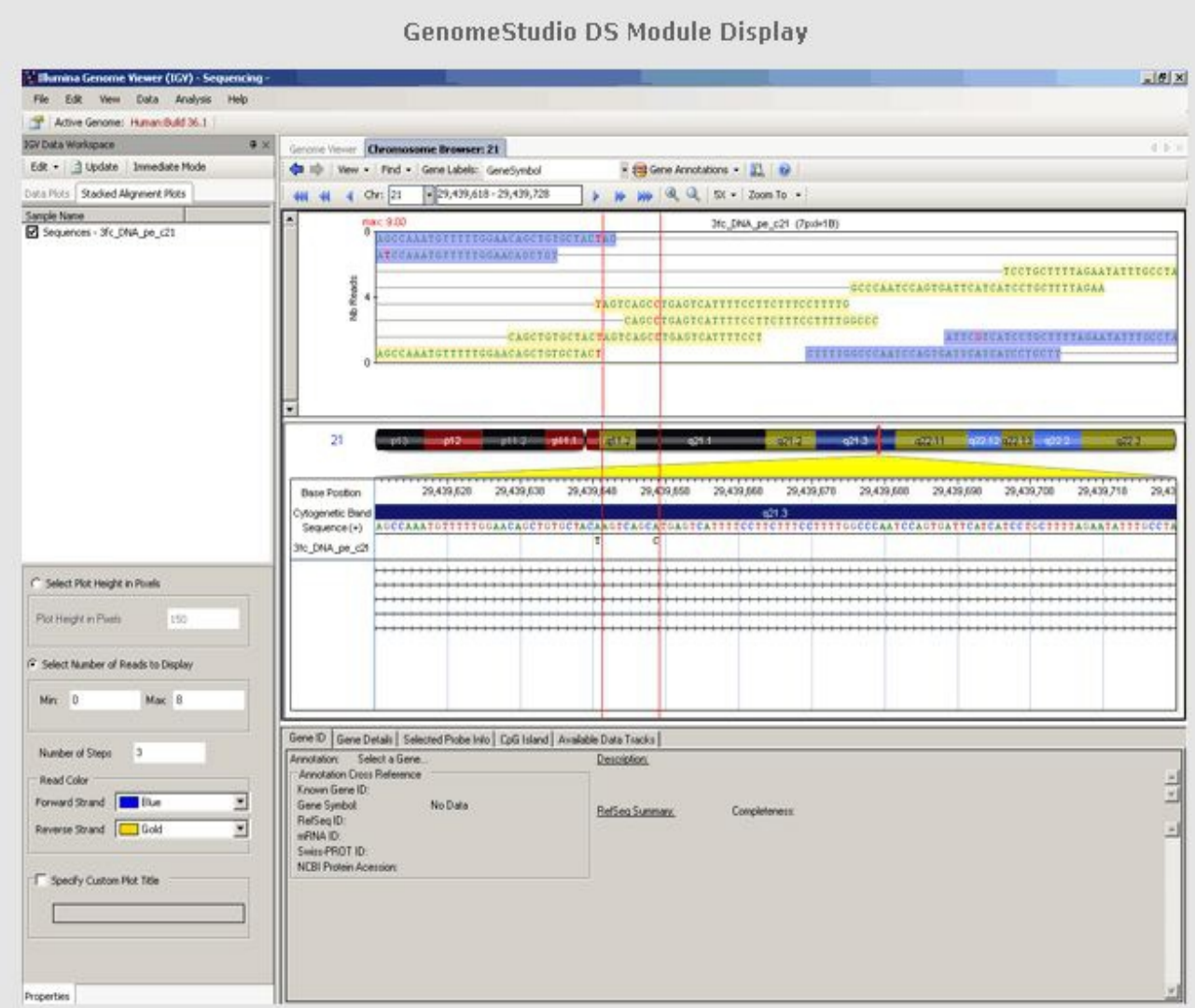
© Doug Brutlag 2015

Illumina Solexa Sequencing Technology



© Doug Brutlag 2015

Illumina Solexa Sequencing Technology

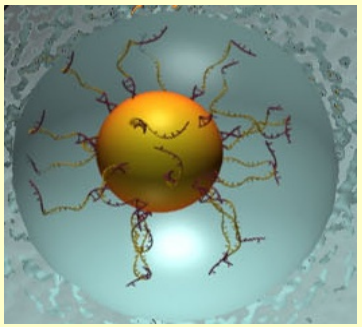


GenomeStudio displays SNPs identified by CASAVA based on alignment of reads against a reference sequence using the Illumina Chromosome Browser.

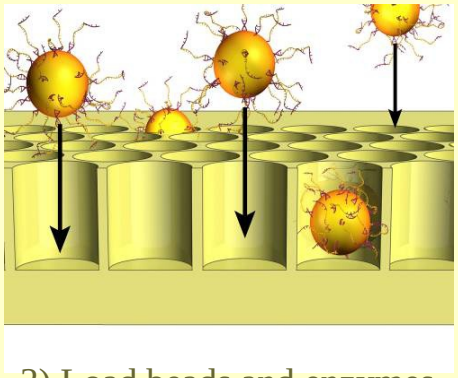
© Doug Brutlag 2015

Life Sciences 454 Process Overview

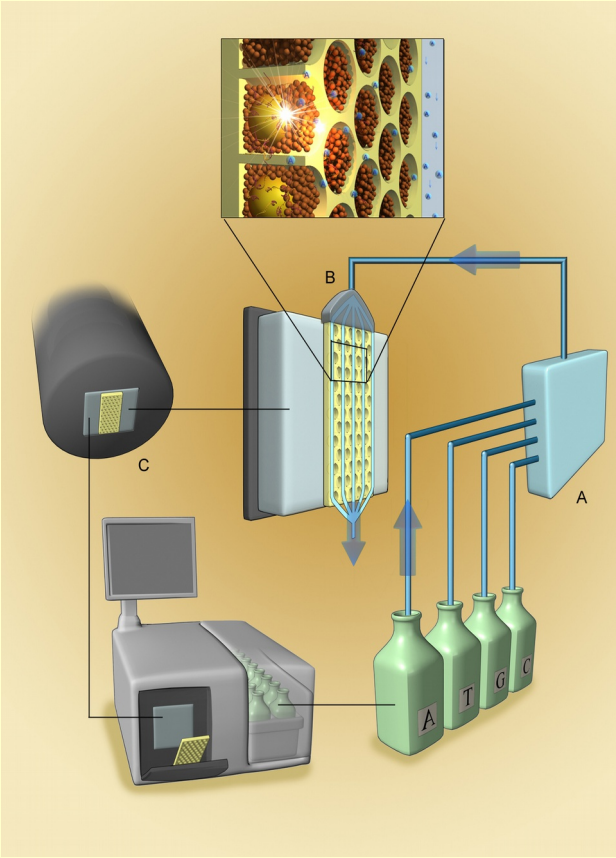
1) Prepare Adapter Ligated ssDNA Library



2) Clonal Amplification on 28 µ beads



3) Load beads and enzymes in PicoTiter Plate™



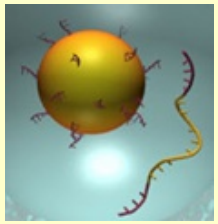
4) Perform Sequencing by synthesis on the 454 Instrument

© Doug Brutlag 2015

Emulsion Based Clonal Amplification

Single test tube generation of millions of clonally amplified sequencing templates
No cloning and colony picking

Adapter carrying library DNA



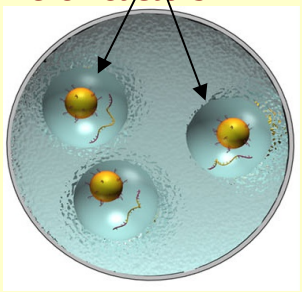
Anneal DNA template to capture beads

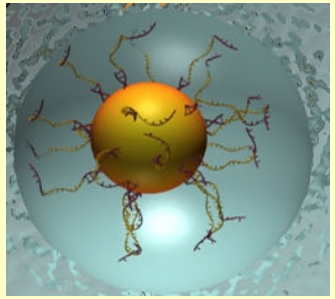
+ PCR Reagents

+ Emulsion Oil

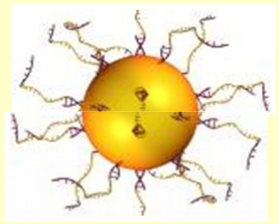
"Water-in-oil" emulsion

Micro-reactors





Perform emulsion PCR

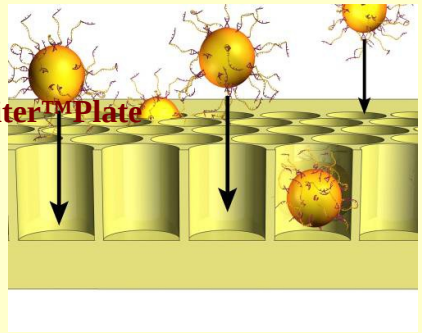


Break micro-reactors
Isolate DNA containing beads

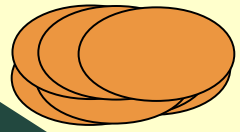
© Doug Brutlag 2015

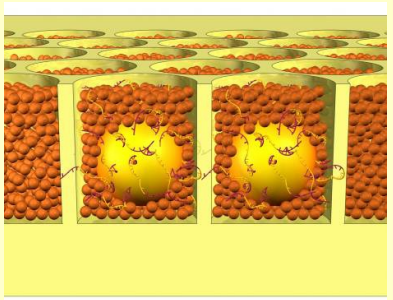
Depositing DNA Beads into the PicoTiter™Plate

Load beads into PicoTiter™Plate

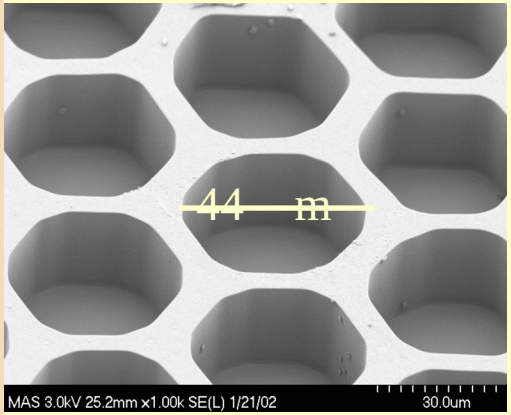


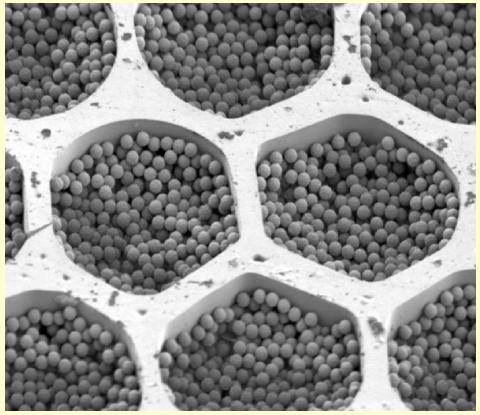
Load Enzyme Beads





Centrifuge Step



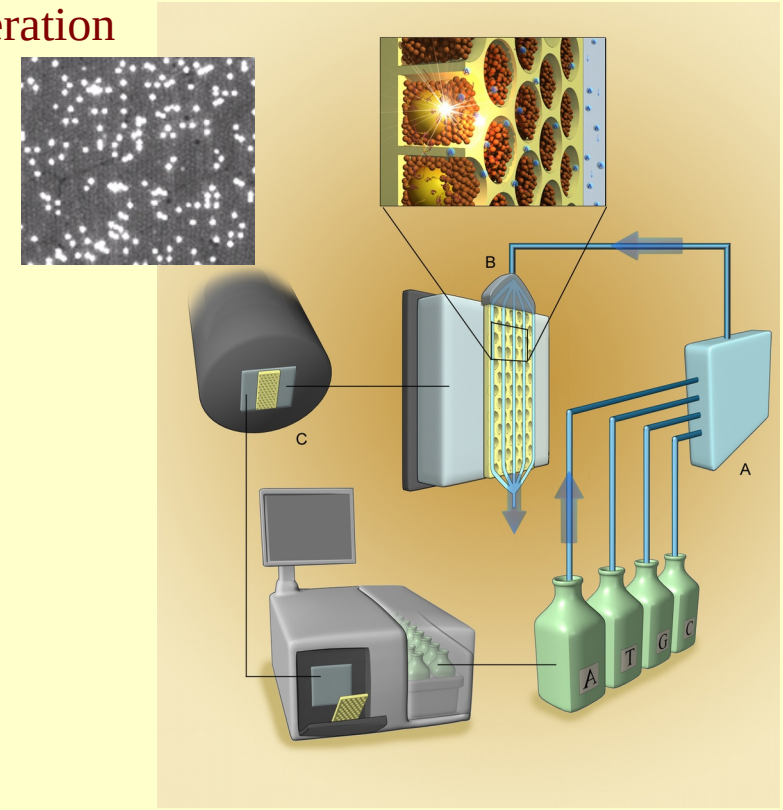


- 70x75mm array of fused optical f bers with etched wells
- 1.6 million wells monitored optically through f ber

© Doug Brutlag 2015

Sequencing-By-Synthesis

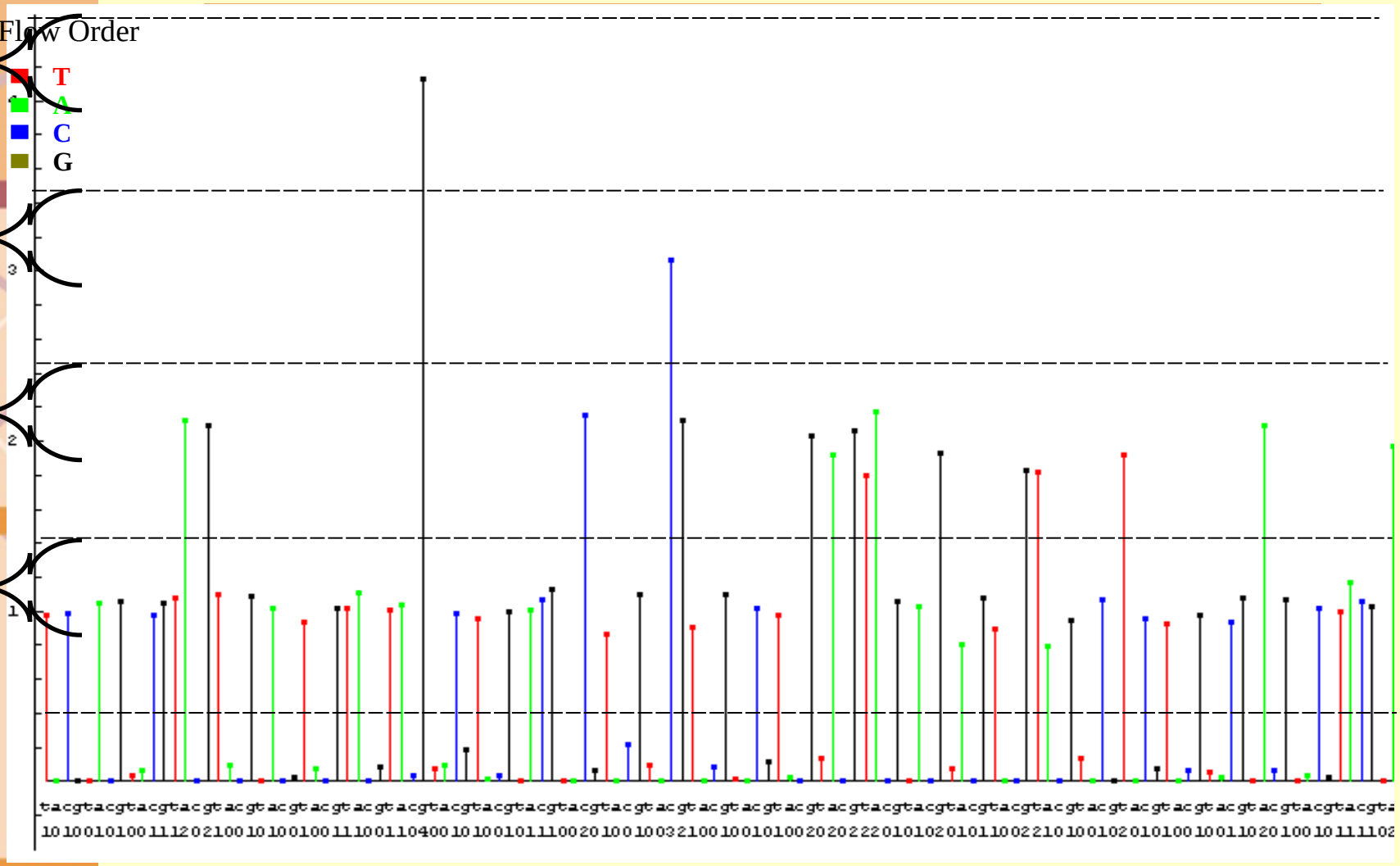
- Simultaneous sequencing of the entire genome in hundreds of thousands of picoliter-size wells
- Pyrophosphate signal generation



© Doug Brutlag 2015

454 LIFE SCIENCES

Flowgrams and BaseCalling



© Doug Brutlag 2015

Pacific Biosciences SMRT Sequencing

© Doug Brutlag 2015

PACIFIC BIOSCIENCES

Pacific Biosciences SMRT Sequencing

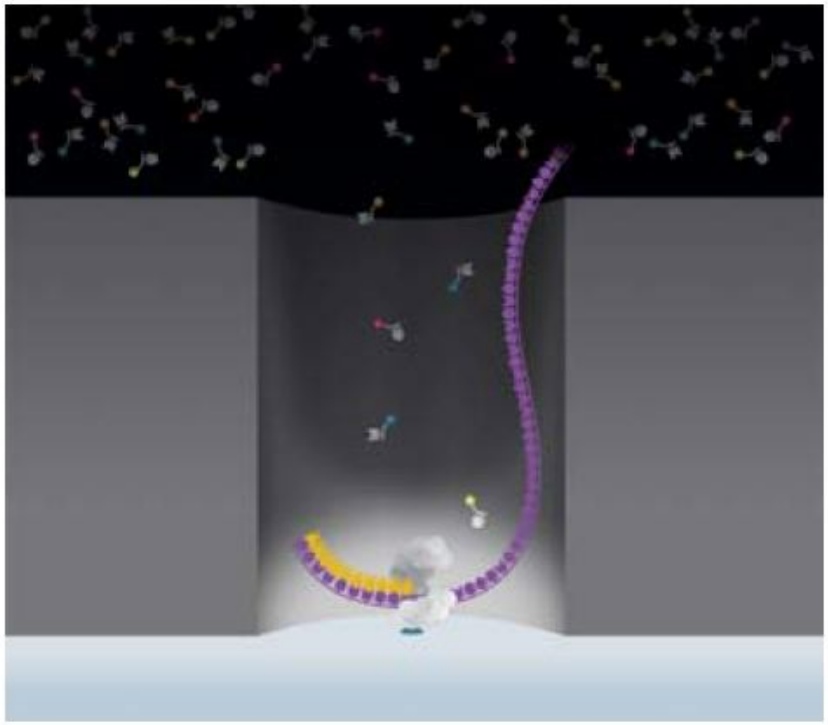


Figure 6. ZMW with DNA polymerase and phospholinked nucleotides

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

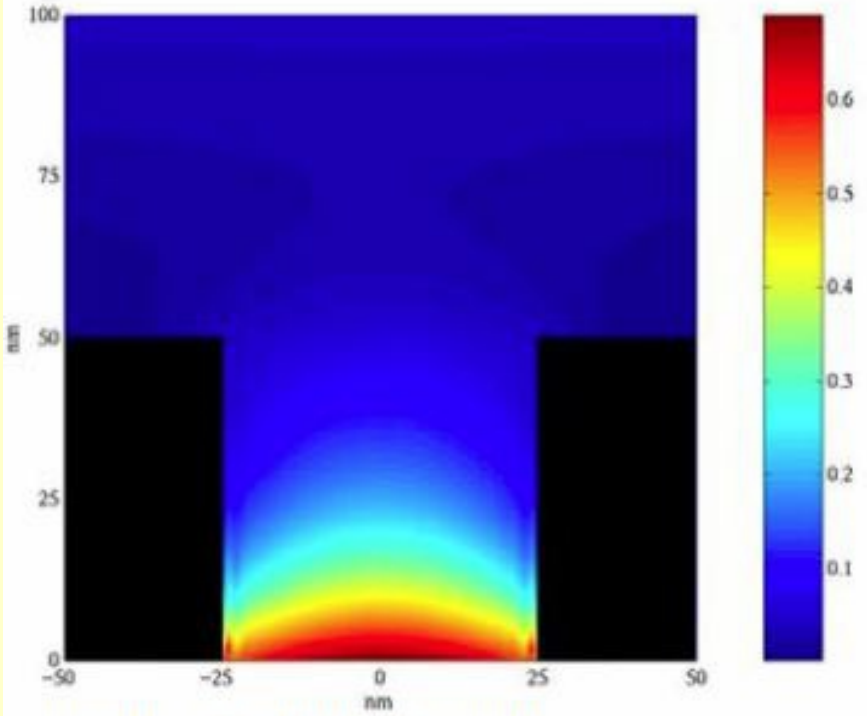


Figure 4. Detection volume

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).

Phospholinked Fluorophores

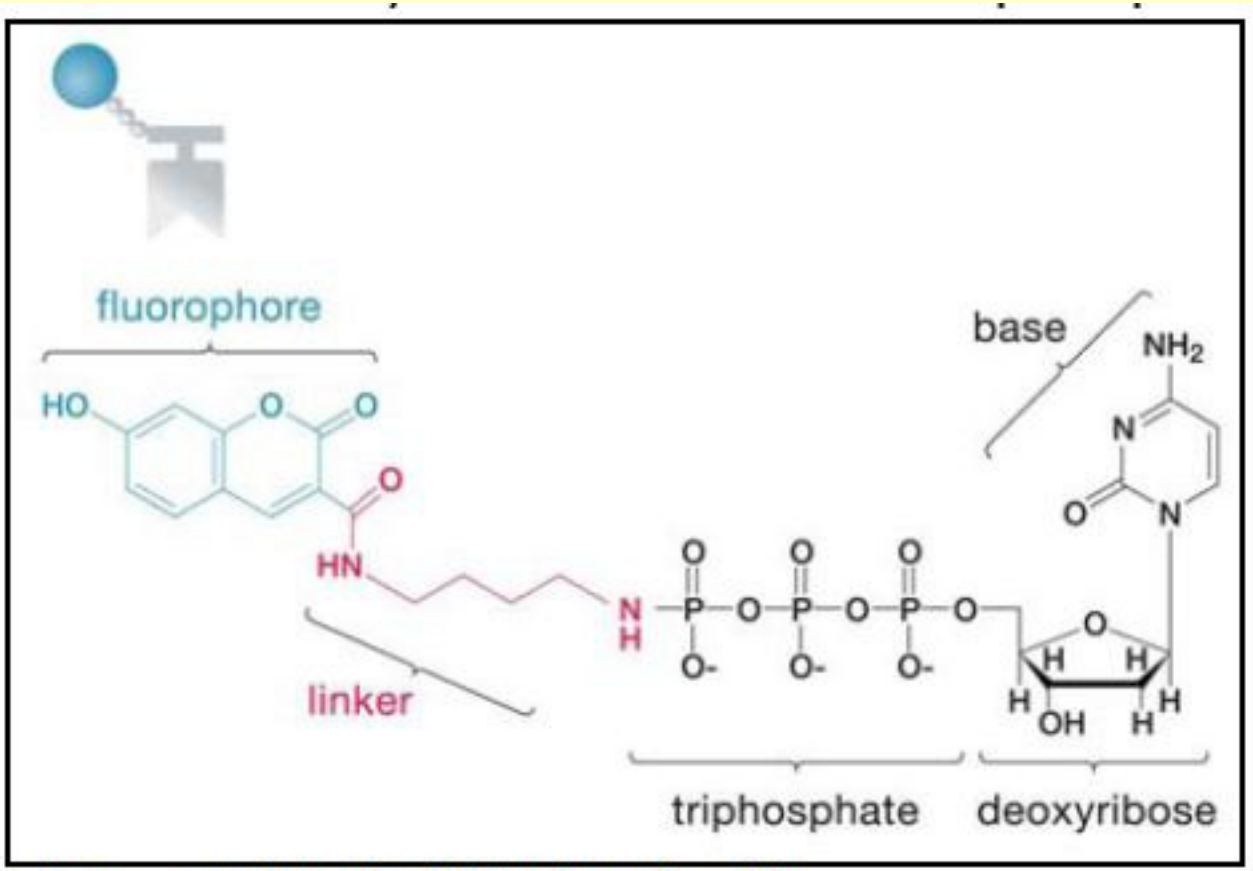


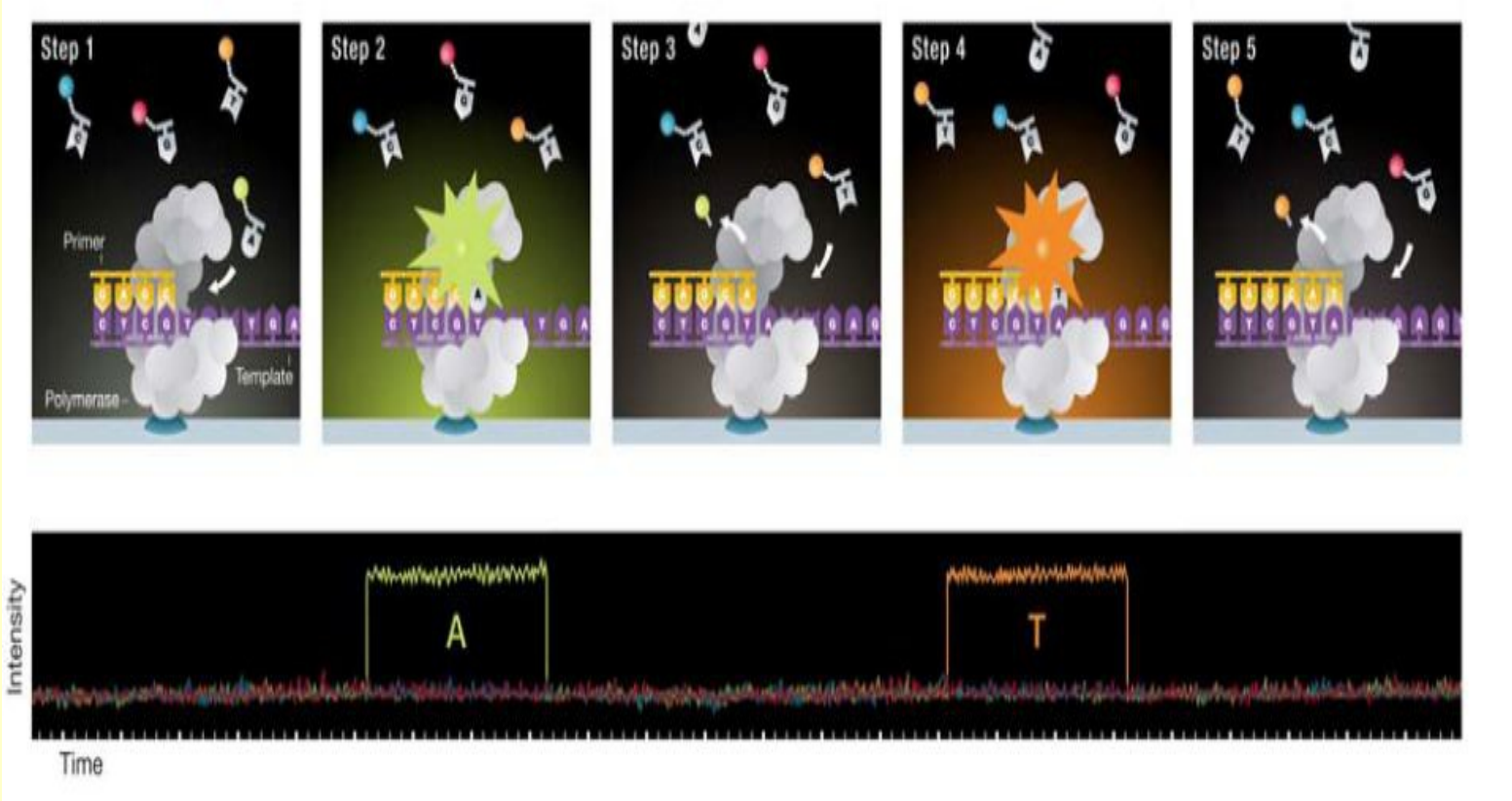
Figure 9. Phospholinked nucleotides

Phospholinked nucleotides have fluorophores attached to the triphosphate chain, which is naturally cleaved when the nucleotide is incorporated.

© Doug Brutlag 2015

PACIFIC BIOSCIENCES

Processive Synthesis



© Doug Brutlag 2015

Synthesis of Long Duplex DNA

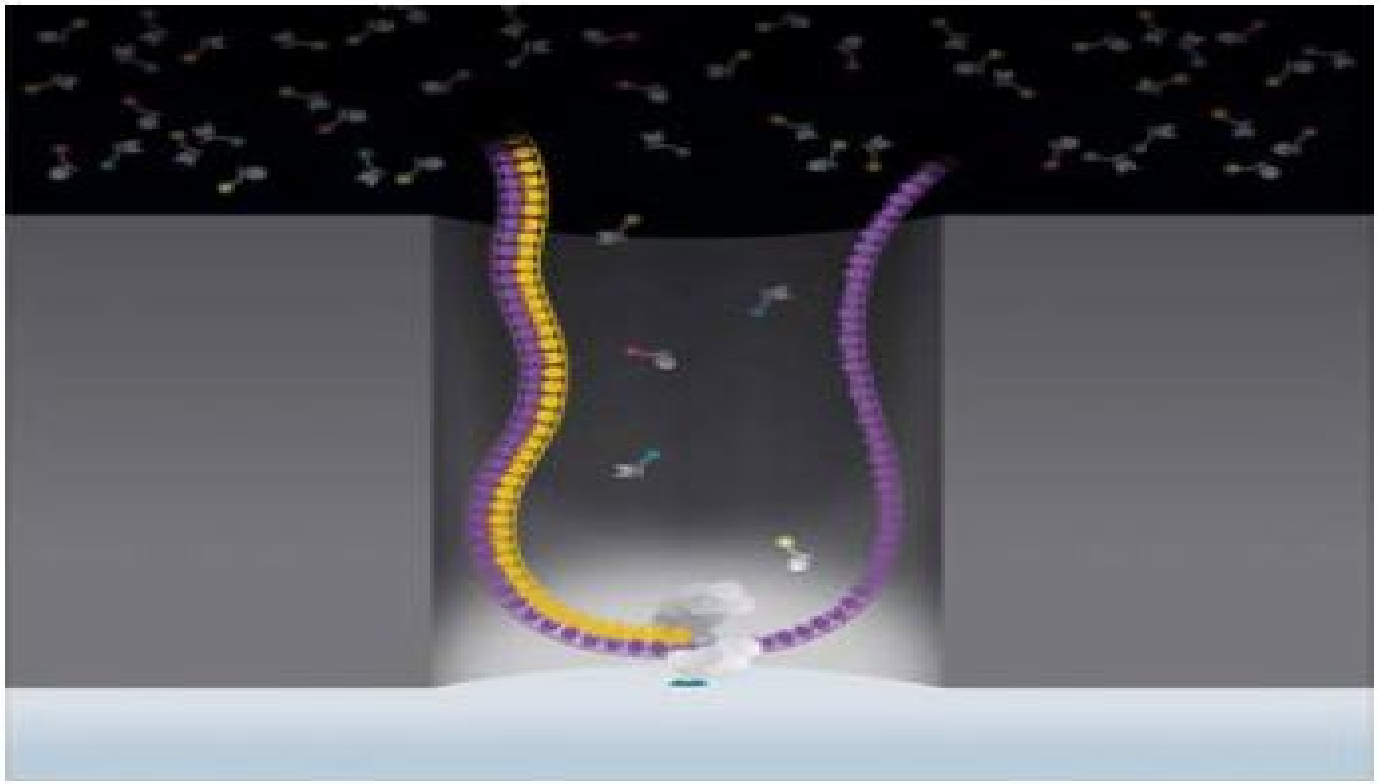
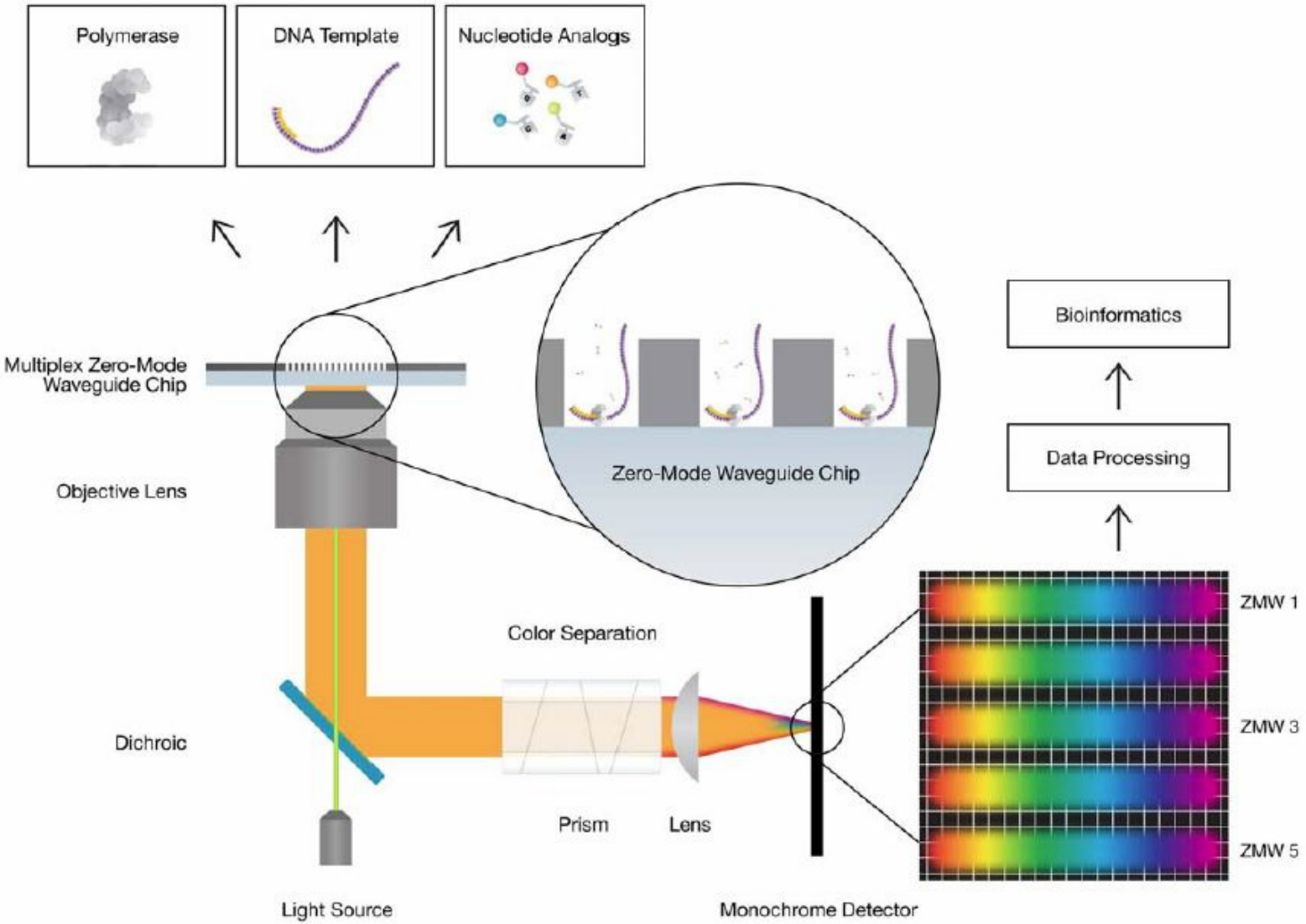


Figure 11. Synthesis of long DNA.

DNA polymerase processively incorporates nucleotides producing long, natural DNA.

© Doug Brutlag 2015

Highly Parallel Optics System



© Doug Brutlag 2015

PACIFIC BIOSCIENCES

Circular Templates Gives Redundant Sequencing and Accuracy

© Doug Brutlag 2015

Circular Templates Gives Redundant Sequencing and Accuracy

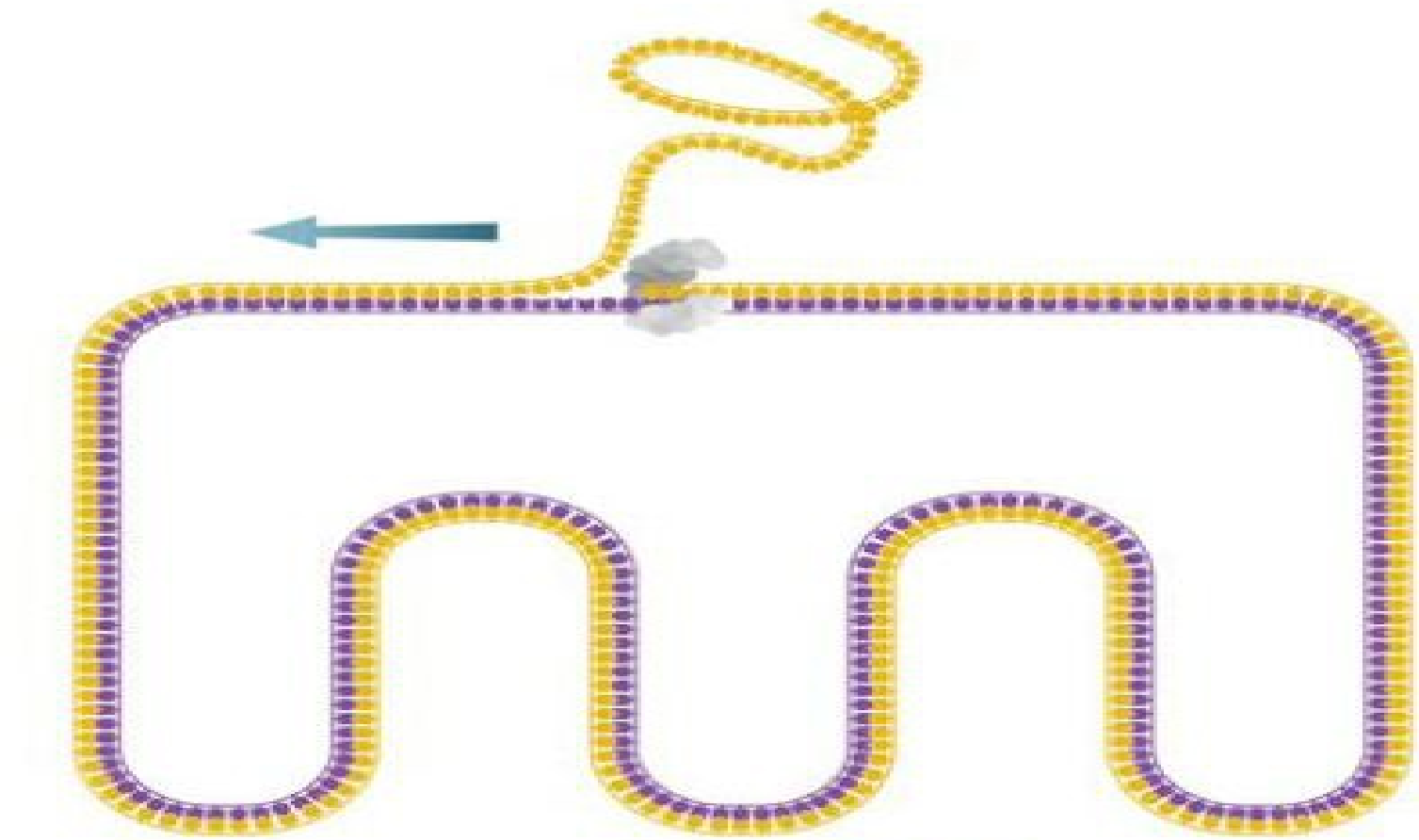


Figure 16. Molecular redundant sequencing

© Doug Brutlag 2015

Ion Torrent Sequencing

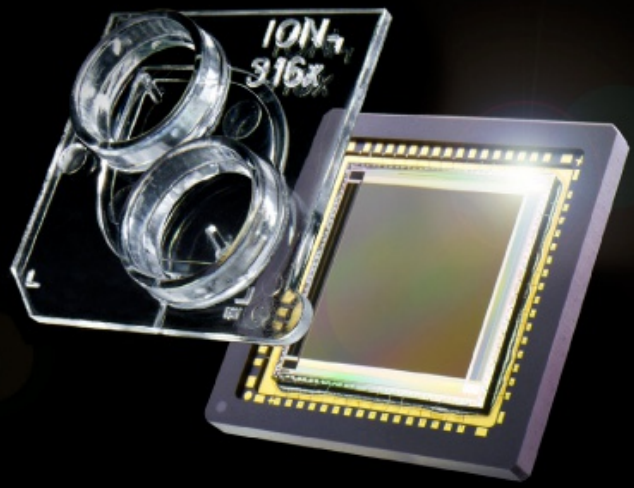
life technologies | ion torrent

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 2015

Ion Torrent Sequencing

http://www.iontorrent.com/

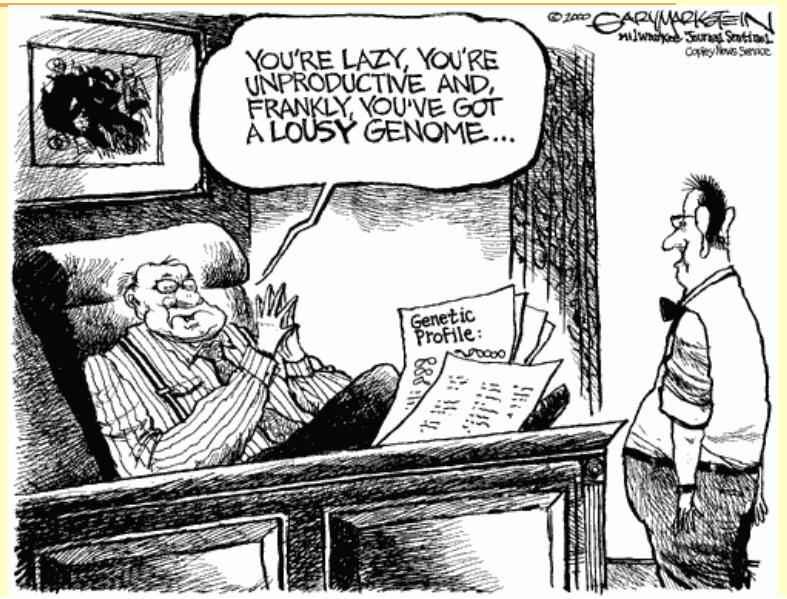
© Doug Brutlag 2015

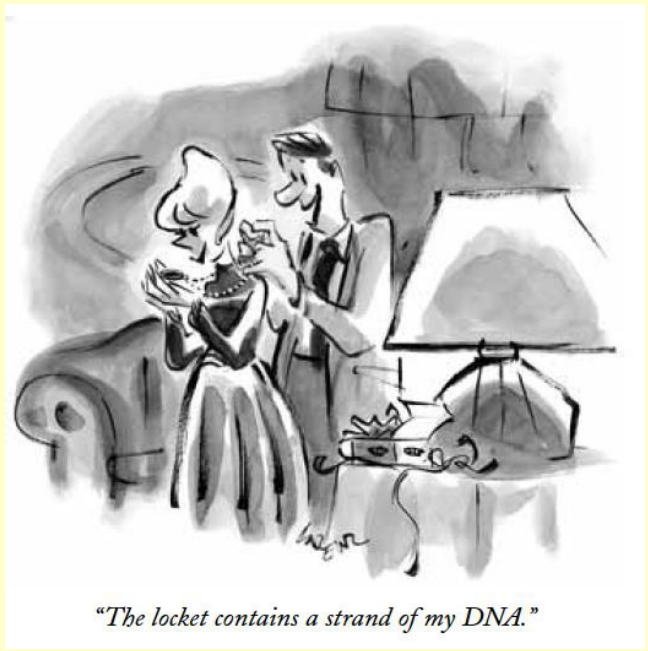
Ion Torrent Sequencing

© Doug Brutlag 2015

The Human Genome
How fast is the cost going down?

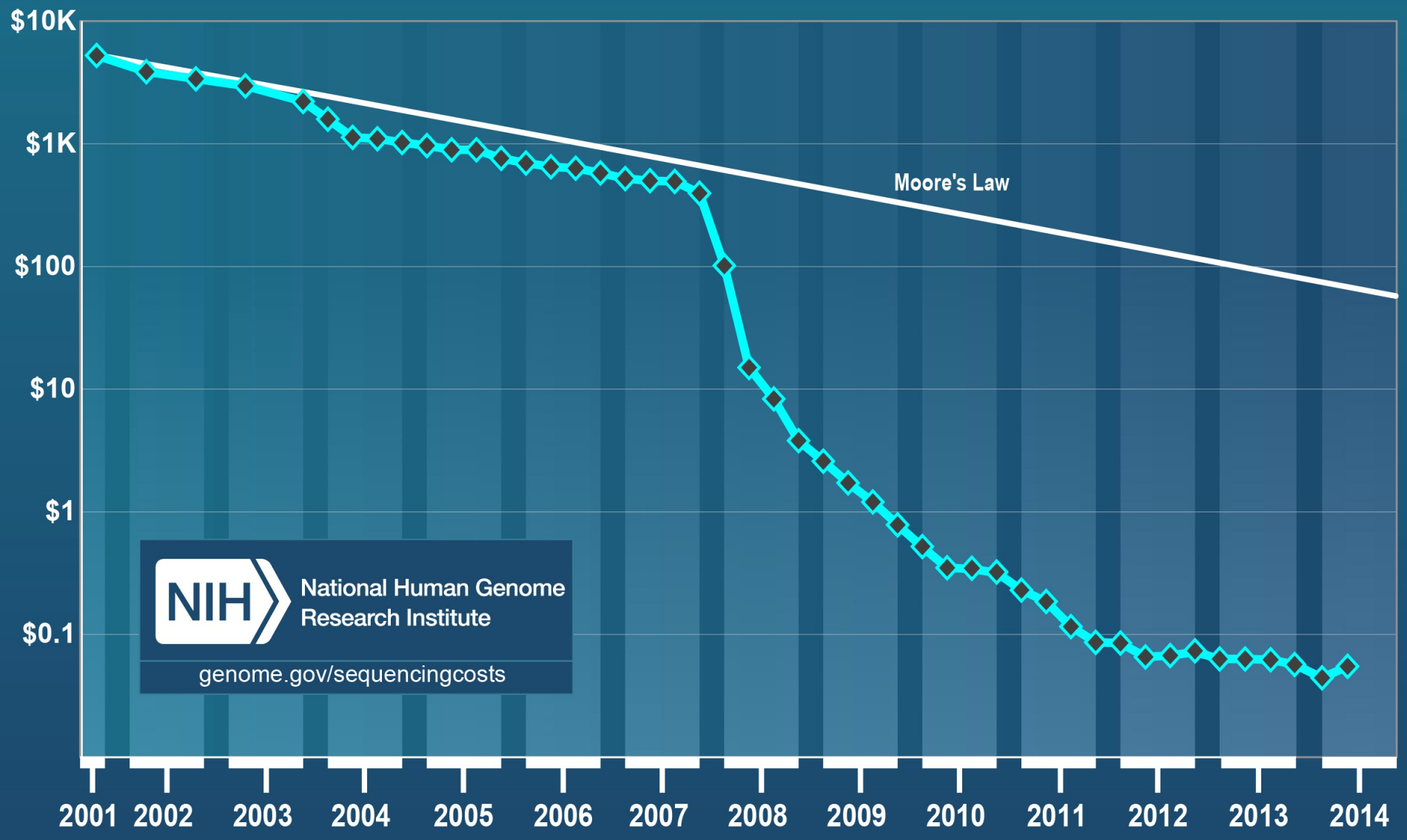
- 2006: $ 50 million
- 2008: $500,000
- 2009: $50,000
- 2010: $20,000
- 2011: $5,000
- 2012: $4,000
- 2013: $3,000
- 2014 $1,400




"The locket contains a strand of my DNA."

Thanks to Seraf ın Batzoglou

Cost per Raw Megabase of DNA Sequence



Cost per Genome