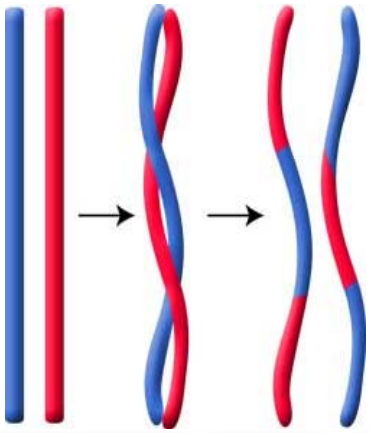
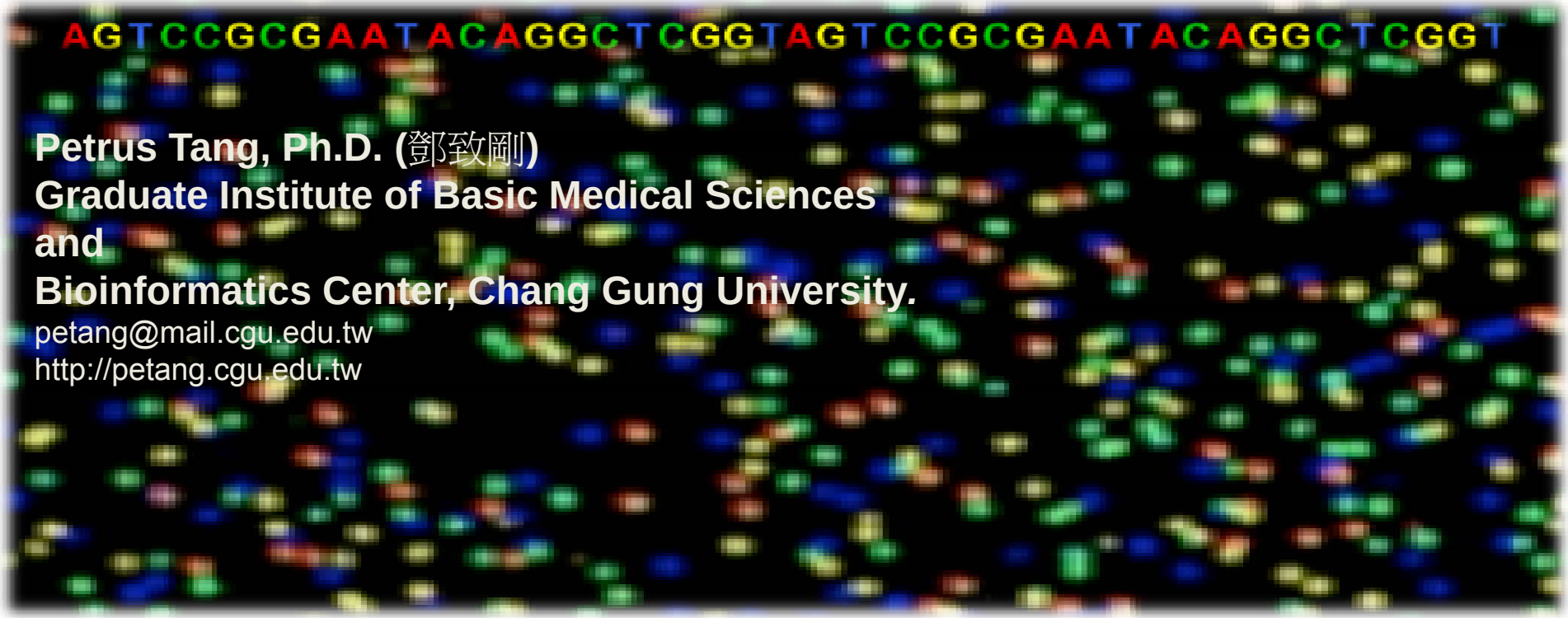
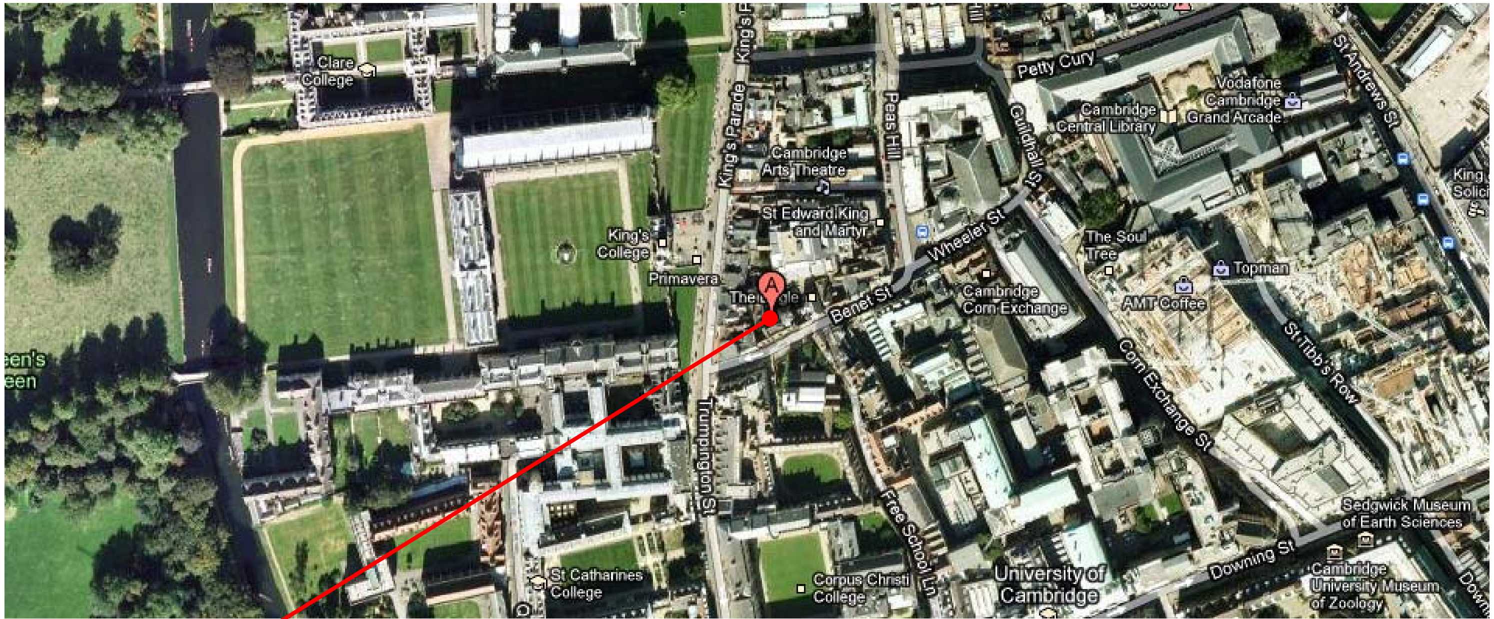


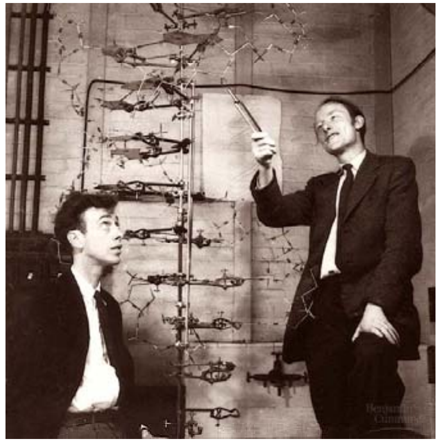
NEXT GENERATION SEQUENCING
Technologies, Applications & Analysis



Petrus Tang, Ph.D. (鄧致剛)
Graduate Institute of Basic Medical Sciences
and
Bioinformatics Center, Chang Gung University.
petang@mail.cgu.edu.tw
http://petang.cgu.edu.tw



The Eagle, Cambridge: the place where Francis Crick interrupted patrons' lunchtime on 28 February 1953 to announce that he and James Watson had "discovered the secret of life" after they had come up with their proposal for the structure of DNA



The Nobel Prize in Chemistry 1980

"for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA"

"for their contributions concerning the determination of base sequences in nucleic acids"

Paul Berg

◑ 1/2 of the prize

USA

Stanford University
Stanford, CA, USA

Walter Gilbert

◔ 1/4 of the prize

USA

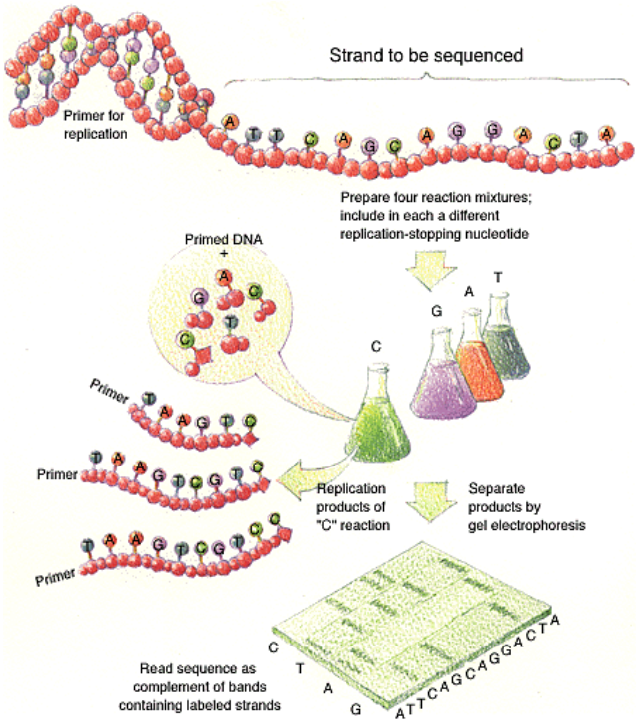
Harvard University,
Biological Laboratories
Cambridge, MA, USA

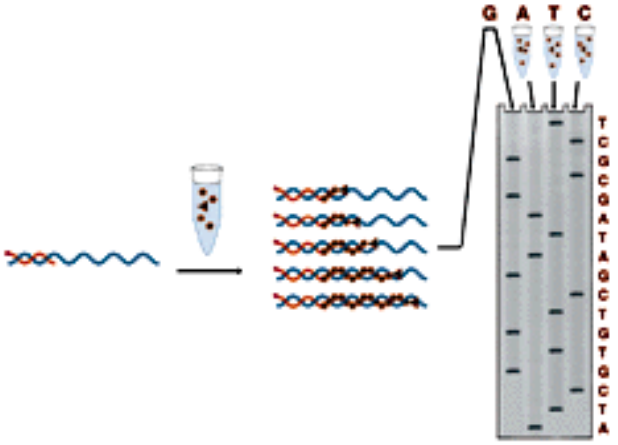
Frederick Sanger

◔ 1/4 of the prize

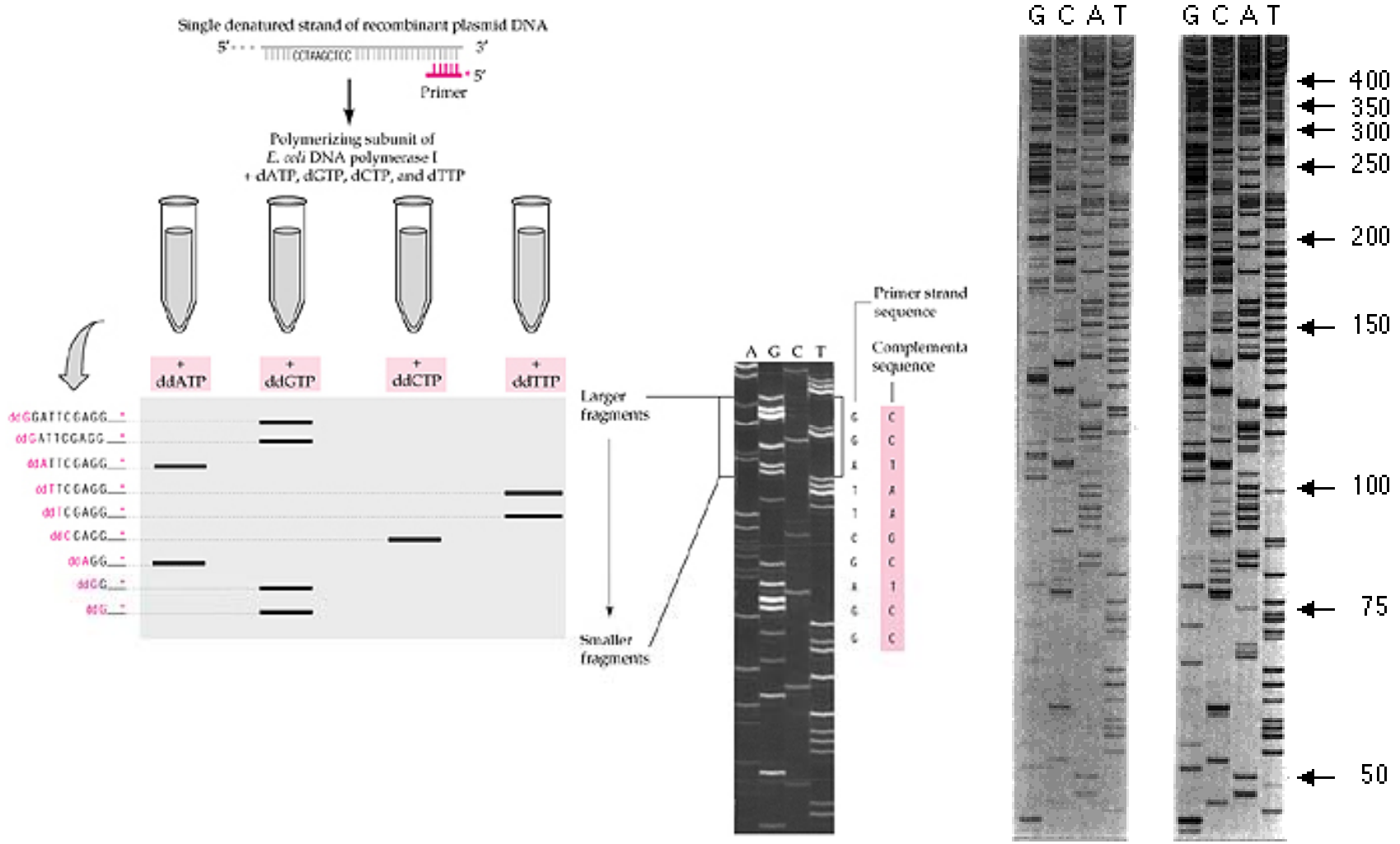
United Kingdom

MRC Laboratory of
Molecular Biology
Cambridge, United





Sanger Dideoxy Sequencing

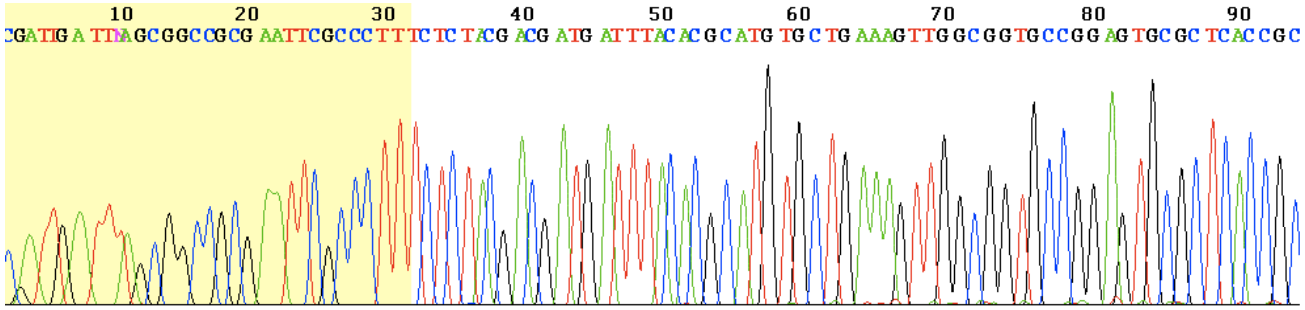


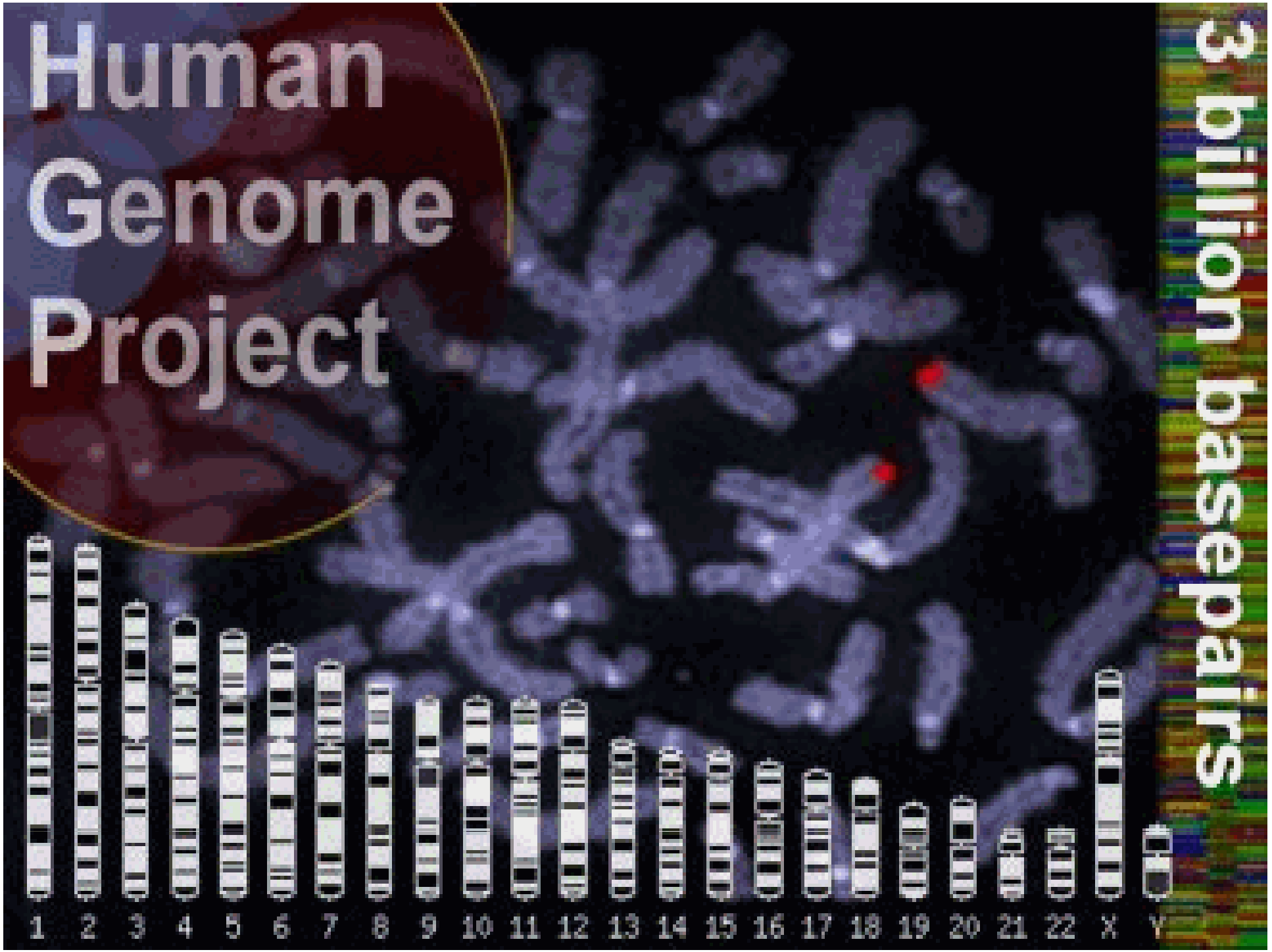
Sanger, F. et al. Nature 24, 687–695 (1977).

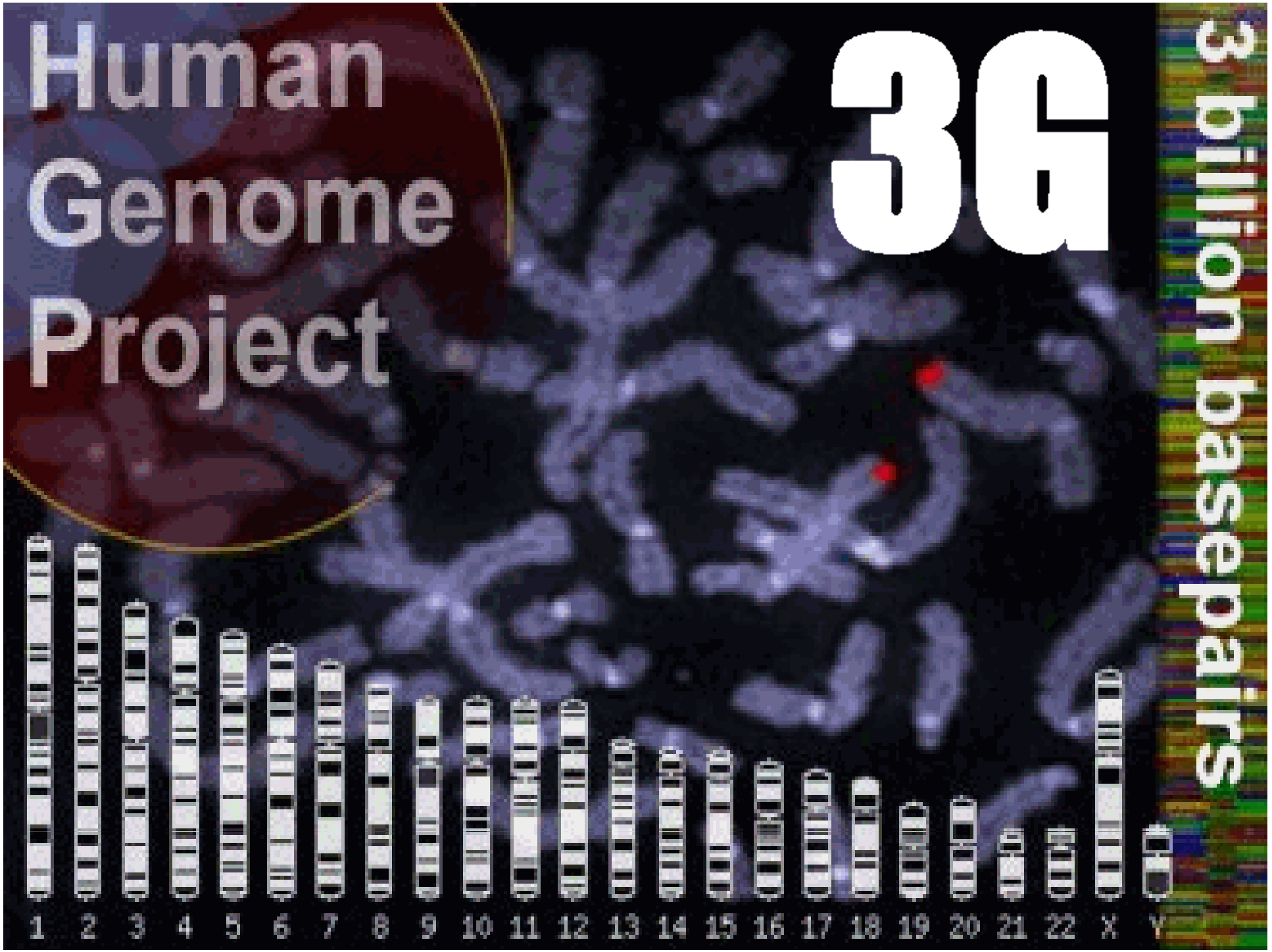
Sanger, F., Nicklen, S. & Coulson, A.R. Proc. Natl. Acad. Sci. USA 74, 5463–5467 (1977).

Basics of the “Old” Technology

- Clone the DNA.
- Generate a ladder of labeled (colored) molecules that are different by 1 nucleotide.
- Separate mixture on some matrix.
- Detect fluorochrome by laser.
- Interpret peaks as string of DNA.
- Strings are 500 to 1,000 letters long
- 1 machine generates 57,000 nucleotides/run
- Assemble all strings into a “whole”.

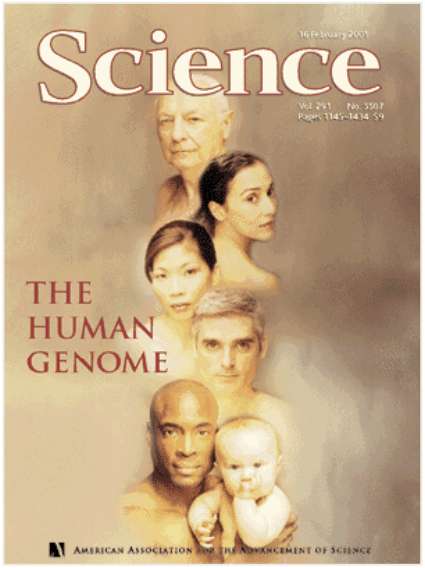



Human
Genome
Project
3 billion base pairs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y


Human
Genome
Project
36
3 billion base pairs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

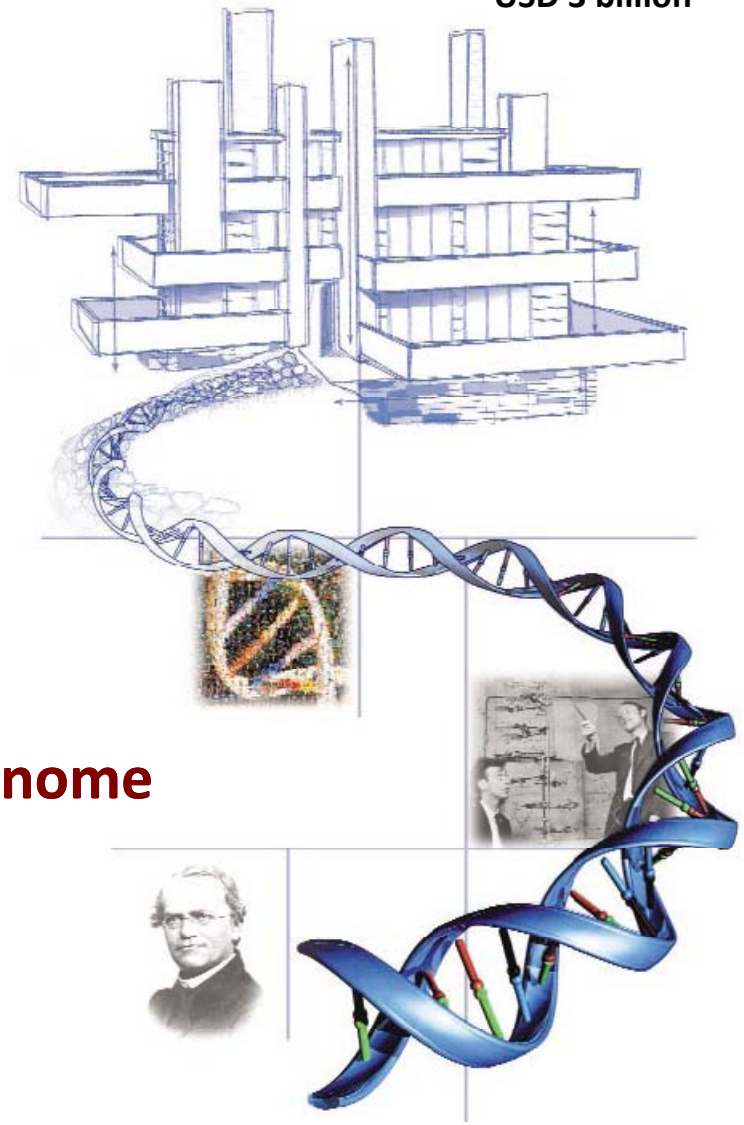
February 2001: Completion of the Draft Human Genome

>10 years to finish
USD 3 billion



Science, *16 February 2001*
Vol. 291, Pages 1145-1434

Nature, *15 February 2001*
Vol. 409, Pages 813-960



April 2003: High-Resolution Human Genome

A vision for the future of genomics research

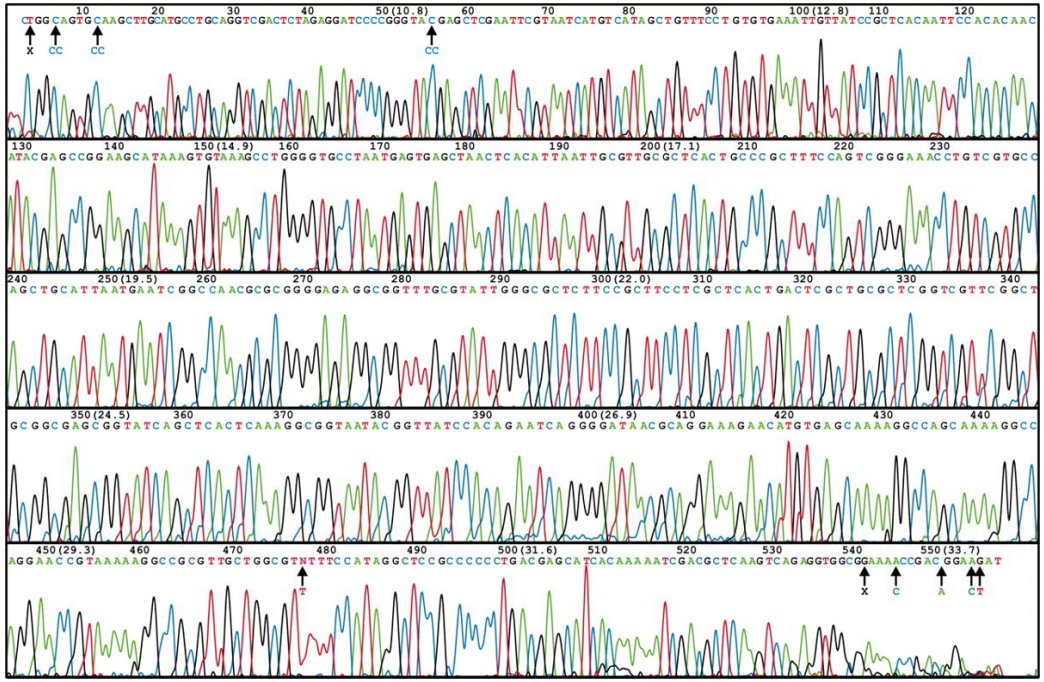
A blueprint for the genomic era.

Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer on behalf of the US National Human Genome Research Institute*

50 Years of DNA: *From Double Helix to Health*
A Celebration of the Genome

Nature, *23 April 2003*
Vol. 422, Pages 1-13

ABI 3730 XL DNA Sequencer



96/384 DNA sequencing in 2 hrs, approximately 600-1000 readable bps per run.

1-4 MB bps/day

*A human genome of **3GB** need **750 days** to finish **1X coverage***

273 authors

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1] Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1] George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5] Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9] Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1] Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1] Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1] Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1] Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1] Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1] Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1] Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1] Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1] Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12] Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1] Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1] Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1] Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1] Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1] Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1] Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1] Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1] Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1] Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1] Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1] Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1] Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1] Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1] Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1] Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1] Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1] Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1] Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1] Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1] Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1] Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1] Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1] Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1] Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1] Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1] Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1] Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1] Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1] Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1] Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

254 authors

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

Genome Sequencing Centres (Listed in order of total genomic sequence contributed, with a partial list of personnel. A full list of contributors at each centre is available as Supplementary Information.)

Whitehead Institute for Biomedical Research, Center for Genome Research: Eric S. Lander[1]*, Lauren M. Linton[1], Bruce Birren[1]*, Chad Nusbaum[1]*, Michael C. Zody[1]*, Jennifer Baldwin[1], Keri Devon[1], Ken Dewar[1], Michael Doyle[1], William FitzHugh[1]*, Roel Funke[1], Diane Gage[1], Katrina Harris[1], Andrew Heaford[1], John Howland[1], Lisa Kann[1], Jessica Lehoczky[1], Rosie LeVine[1], Paul McEwan[1], Kevin McKernan[1], James Meldrim[1], Jill P. Mesirov[1]*, Cher Miranda[1], William Morris[1], Jerome Naylor[1], Christina Raymond[1], Mark Rosetti[1], Ralph Santos[1], Andrew Sheridan[1], Carrie Sougnez[1], Nicole Stange-Thomann[1], Nikola Stojanovic[1], Aravind Subramanian[1] & Dudley Wyman[1]

The Sanger Centre: Jane Rogers[2], John Sulston[2]*, Rachael Ainscough[2], Stephan Beck[2], David Bentley[2], John Burton[2], Christopher Clee[2], Nigel Carter[2], Alan Coulson[2], Rebecca Deadman[2], Panos Deloukas[2], Andrew Dunham[2], Ian Dunham[2], Richard Durbin[2]*, Lisa French[2], Darren Grafham[2], Simon Gregory[2], Tim Hubbard[2]*, Sean Humphray[2], Adrienne Hunt[2], Matthew Jones[2], Christine Lloyd[2], Amanda McMurray[2], Lucy Matthews[2], Simon Mercer[2], Sarah Milne[2], James C. Mullikin[2]*, Andrew Mungall[2], Robert Plumb[2], Mark Ross[2], Ratna Shownkeen[2] & Sarah Sims[2]

Washington University Genome Sequencing Center: Robert H. Waterston[3]*, Richard K. Wilson[3], LaDeana W. Hillier[3]*, John D. McPherson[3], Marco A. Marra[3], Elaine R. Mardis[3], Lucinda A. Fulton[3], Asif T. Chinwalla[3]*, Kymberlie H. Pepin[3], Warren R. Gish[3], Stephanie L. Chissoe[3], Michael C. Wendl[3], Kim D. Delehaunty[3], Tracie L. Miner[3], Andrew Delehaunty[3], Jason B. Kramer[3], Lisa L. Cook[3], Robert S. Fulton[3], Douglas L. Johnson[3], Patrick J. Minx[3] & Sandra W. Clifton[3]

US DOE Joint Genome Institute: Trevor Hawkins[4], Elbert Branscomb[4], Paul Predki[4], Paul Richardson[4], Sarah Wenning[4], Tom Slezak[4], Norman Doggett[4], Jan-Fang Cheng[4], Anne Olsen[4], Susan Lucas[4], Christopher Elkin[4], Edward Uberbacher[4] & Marvin Frazier[4]

Baylor College of Medicine Human Genome Sequencing Center: Richard A. Gibbs[5]*, Donna M. Muzny[5], Steven E. Scherer[5], John B. Bouck[5]*, Erica J. Sodergren[5], Kim C. Worley[5]*, Catherine M. Rives[5], James H. Gorrell[6], Michael L. Metzker[5], Susan L. Naylor[6], Raju S. Kucherlapati[6], David L. Nelson, & George M. Weinstock[5]

RIKEN Genomic Sciences Center: Yoshiyuki Sakaki[9], Asao Fujiyama[9], Masahira Hattori[9], Tetsushi Yada[9], Atsushi Toyoda[9], Takehiko Itoh[9], Chiharu Kawagoe[9], Hidemi Watanabe[9], Yasushi Totoki[9] & Todd Taylor[9]

Genoscope and CNRS UMR-8030: Jean Weissenbach[10], Roland Heilig[10], William Saurin[10], Francois Artiguenave[10], Philippe Brottier[10], Thomas Bruls[10], Eric Pelletier[10], Catherine Robert[10] & Patrick Wincker[10]

GTC Sequencing Center: Douglas R. Smith[11], Lynn Doucette-Stamm[11], Marc Rubenfield[11], Keith Weinstock[11], Hong Mei Lee[11] & JoAnn Dubois[11]

Department of Genome Analysis, Institute of Molecular Biotechnology: André Rosenthal[12], Matthias Platzer[12], Gerald Nyakatura[12], Stefan Taudien[12] & Andreas Rump[12]

Beijing Genomics Institute/Human Genome Center: Huanming Yang[13], Jun Yu[13], Jian Wang[13], Guyang Huang[14] & Jun Gu[15]

Multimegabase Sequencing Center, The Institute for Systems Biology: Leroy Hood[16], Lee Rowen[16], Anup Madan[16] & Shizen Qin[16]

Stanford Genome Technology Center: Ronald W. Davis[17], Nancy A. Federspiel[17], A. Pia Abola[17] & Michael J. Proctor[17]

Stanford Human Genome Center: Richard M. Myers[18], Jeremy Schmutz[18], Mark Dickson[18], Jane Grimwood[18] & David R. Cox[18]

University of Washington Genome Center: Maynard V. Olson[19], Rajinder Kaul[19] & Christopher Raymond[19]

Department of Molecular Biology, Keio University School of Medicine: Nobuyoshi Shimizu[20], Kazuhiko Kawasaki[20] & Shinsei Minoshima[20]

University of Texas Southwestern Medical Center at Dallas: Glen A. Evans[21]†, Maria Athanasiou[21] & Roger Schultz[21]

University of Oklahoma's Advanced Center for Genome Technology: Bruce A. Roe[22], Feng Chen[22] & Huaqin Pan[22]

Max Planck Institute for Molecular Genetics: Juliane Ramser[23], Hans Lehrach[23] & Richard Reinhardt[23]

Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center: W. Richard McCombie[24], Melissa de la Bastide[24] & Neilay Dedhia[24]

GBF—German Research Centre for Biotechnology: Helmut Blöcker[25], Klaus Homischer[25] & Gabriele Nordsiek[25]

* **Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings):** Richa Agarwala[26], L. Aravind[26], Jeffrey A. Bailey[27], Alex Bateman[2], Serafim Batzoglou[1], Ewan Birney[28], Peer Bork[29,30], Daniel G. Brown[1], Christopher B. Burge[31], Lorenzo Cerutti[28], Hsiu-Chuan Chen[26], Deanna Church[26], Michele Clamp[2], Richard R. Copley[30], Tobias Doerks[29,30], Sean R. Eddy[32], Evan E. Eichler[27], Terrence S. Furey[33], James Galagan[1], James G. R. Gilbert[2], Cyrus Harmon[34], Yoshihide Hayashizaki[35], David Haussler[36], Henning Hermjakob[28], Karsten Hokamp[37], Wonhee Jang[26], L. Steven Johnson[32], Thomas A. Jones[32], Simon Kasif[38], Arek Kasprzyk[28], Scot Kennedy[39], W. James Kent[40], Paul Kitts[26], Eugene V. Koonin[26], Ian Korf[1], David Kulp[34], Doron Lancet[41], Todd M. Lowe[42], Aoife McLysaght[37], Tarjei Mikkelsen[38], John V. Moran[43], Nicola Mulder[28], Victor J. Pollara[1], Chris P. Ponting[44], Greg Schuler[26], Jörg Schultz[30], Guy Slater[28], Arian F. A. Smit[45], Elia Stupka[28], Joseph Szustakowki[38], Danielle Thierry-Mieg[26], Jean Thierry-Mieg[26], Lukas Wagner[26], John Wallis[3], Raymond Wheeler[34], Alan Williams[34], Yuri I. Wolf[26], Kenneth H. Wolfe[37], Shiaw-Pyng Yang[3] & Ru-Fang Yeh[31]

Scientific management: National Human Genome Research Institute, US National Institutes of Health: Francis Collins[46], Mark S. Guyer[46], Jane Peterson[46], Adam Felsenfeld[46], & Kris A. Wetterstrand[46]; **Office of Science, US Department of Energy:** Aristides Patrinos[47]; **The Wellcome Trust:** Michael J. Morgan[48]

Tong et al. Genome Biology 2010, 11:R91
http://genomebiology.com/2010/11/9/R91

Genome Biology

RESEARCH Open Access

Sequencing and analysis of an Irish human genome

Pin Tong[1†], James GD Prendergast[2†], Amanda J Lohan[1], Susan M Farrington[2,3], Simon Cronin[4], Nial Friel[5], Dan G Bradley[6], Orla Hardiman[7], Alex Evans[8], James F Wilson[9], Brendan Loftus[1*]

Abstract

Background: Recent studies generating complete human sequences from Asian, African and European subgroups have revealed population-specific variation and disease susceptibility loci. Here, choosing a DNA sample from a population of interest due to its relative geographical isolation and genetic impact on further populations, we extend the above studies through the generation of 11-fold coverage of the first Irish human genome sequence.

Results: Using sequence data from a branch of the European ancestral tree as yet unsequenced, we identify variants that may be specific to this population. Through comparisons with HapMap and previous genetic association studies, we identified novel disease-associated variants, including a novel nonsense variant putatively associated with inflammatory bowel disease. We describe a novel method for improving SNP calling accuracy at low genome coverage using haplotype information. This analysis has implications for future re-sequencing studies and validates the imputation of Irish haplotypes using data from the current Human Genome Diversity Cell Line Panel (HGDP-CEPH). Finally, we identify gene duplication events as constituting significant targets of recent positive selection in the human lineage.

Conclusions: Our findings show that there remains utility in generating whole genome sequences to illustrate both general principles and reveal specific instances of human biology. With increasing access to low cost sequencing we would predict that even armed with the resources of a small research group a number of similar initiatives geared towards answering specific biological questions will emerge.

nature genetics

Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing

Akihiro Fujimoto, Hidewaki Nakagawa, Naoya Hosono, Kaoru Nakano, Tetsuo Abe, Keith A Boroevich, Masao Nagasaki, Rui Yamaguchi, Tetsuo Shibuya, Michiaki Kubo, Satoru Miyano, Yusuke Nakamura & Tatsuhiko Tsunoda

Affiliations | **Contributions** | **Corresponding authors**

Nature Genetics **42**, 931–936 (2010) | doi:10.1038/ng.691
Received 18 February 2010 | Accepted 10 September 2010 | Published online 24 October 2010

Abstract

Abstract • **Author information** • **Supplementary information**

We report the analysis of a Japanese male using high-throughput sequencing to ×40 coverage. More than 99% of the sequence reads were mapped to the reference human genome. Using a Bayesian decision method, we identified 3,132,608 single nucleotide variations (SNVs). Comparison with six previously reported genomes revealed an excess of singleton nonsense and nonsynonymous SNVs, as well as singleton SNVs in conserved non-coding regions. We also identified 5,319 deletions smaller than 10 kb with high accuracy, in addition to copy number variations and rearrangements. *De novo* assembly of the unmapped sequence reads generated around 3 Mb of novel sequence, which showed high similarity to non-reference human genomes and the human herpesvirus 4 genome. Our analysis suggests that considerable variation remains undiscovered in the human genome and that whole-genome sequencing is an invaluable tool for obtaining a complete understanding of human genetic variation.

Next Generation Sequencing Technology

Massively **P**arallel **S**ignature **S**equencing (MPSS)

NATURE METHODS | VOL.5 NO.1 | JANUARY 2008

***Nature Methods*' Method of the Year 2007 goes to next-generation sequencing. This series of articles showcase how these novel sequencing methods came into their own in 2007 and the incredible impact they promise to have in a variety of research applications. The Methods to Watch feature provide a glimpse and a wish list for future Methods of the Year.**

Roche **454 GS FLX**

http://www.454.com/

illumina **SOLEXA**

http://www.illumina.com/pages.ilmn?ID=250

applied biosystems **SOLID**

http://marketing.appliedbiosystems.com/

Throughput of NGS machines (2007-2009)

Vendor:	Roche			Illumina			ABI	
Technology:	454			Solexa			SOLiD	
Platform:	GS 20	FLX	Ti	GA	GA II		1	2
Reads:	500 k	500 k	1 M	28 M	80 M		40 M	115 M
Fragment								
Read length:	100	200	350	35	50	75	25	35
Run time:	6 hr	7 hr	9 hr	3 d	3 d	4 d	6 d	5 d
Yield:	50 Mb	100 Mb	400 Mb	1 Gb	4 Gb	6 Gb	1 Gb	4 Gb
Images:	11 GB	13 GB	27 GB	500 GB	1.1 TB	1.7 TB	1.8 TB	2.5 TB
PA Disk:	3 GB	3 GB	15 GB	175 GB	300 GB	350 GB	300 GB	750 GB
PA CPU:	10 hr	140 hr	220 hr	100 hr	70 hr	100 hr	NA	NA
SRA:	500 MB	1 GB	4 GB	30 GB	50 GB	75 GB	100 GB	140 GB
Paired-end								
Read length:		200		2×35	2×50	2×75	2×25	2×35
Insert:		3.5 kb		200 b	200 b	200 b	3 kb	3 kb
Run time:		7 hr		6 d	6 d	8 d	12 d	10 d
Yield:		100 Mb		2 Gb	8 Gb	11 Gb	2 Gb	8 Gb
Images:		13 GB		1 TB	2.2 TB	3.4 TB	3.6 TB	5 TB
PA Disk:		3 GB		350 GB	500 GB	600 GB	600 GB	1.5 TB
PA CPU:		140 hr		160 hr	120 hr	170 hr	NA	NA
SRA:		1 GB		60 GB	100 GB	150 GB	200 GB	280 GB

Throughput of NGS machines (2010)

Vendor:	Roche			Illumina			ABI	
Technology:	454			Solexa			SOLiD	
Platform:	GS 20	FLX	Ti	GA	GA II		1	2
Reads:	500 k	500 k	1 M	28 M	80 M		40 M	115 M
Fragment								
Read length:	100	200	350	35	50	75	25	35
Run time:	6 hr	7 hr	9 hr	3 d	3 d	4 d	6 d	5 d
Yield:	50 Mb	100 Mb	400 Mb	1 Gb	4 Gb	6 Gb	1 Gb	4 Gb
Images:	11 GB	13 GB	27 GB	500 GB	1.1 TB	1.7 TB	1.8 TB	2.5 TB
PA Disk:	5 GB	5 GB	15 GB	175 GB	300 GB	350 GB	300 GB	750 GB
PA CPU:	10 hr	140 hr	220 hr	100 hr	70 hr	100 hr	NA	NA
SRA:	500 MB	1 GB	4 GB	30 GB	50 GB	75 GB	100 GB	140 GB
Paired-end								
Read length:		200		2×35	2×50	2×75	2×25	2×35
Insert:		3.5 kb		200 b	200 b	200 b	3 kb	3 kb
Run time:		7 hr		6 d	6 d	8 d	12 d	10 d
Yield:		100 Mb		2 Gb	8 Gb	11 Gb	2 Gb	8 Gb
Images:		13 GB		1 TB	2.2 TB	3.4 TB	3.6 TB	5 TB
PA Disk:		3 GB		350 GB	500 GB	600 GB	600 GB	1.5 TB
PA CPU:		140 hr		160 hr	120 hr	170 hr	NA	NA
SRA:		1 GB		60 GB	100 GB	150 GB	200 GB	280 GB

illumina®

HiSeq2000 (launched in 2010)

Read Length	Run Time	Output
1 x 35 bp	~1.5 days	26-35 Gb
2 x 50 bp	~4 days	75-100 Gb
2 x 100 bp	~8 days	150-200 Gp

Up to 25GB per day for a 2 x100 bp run

applied biosystems

SOLiD 4 (launched in 2009)

NEXT GENERATION SEQUENCING

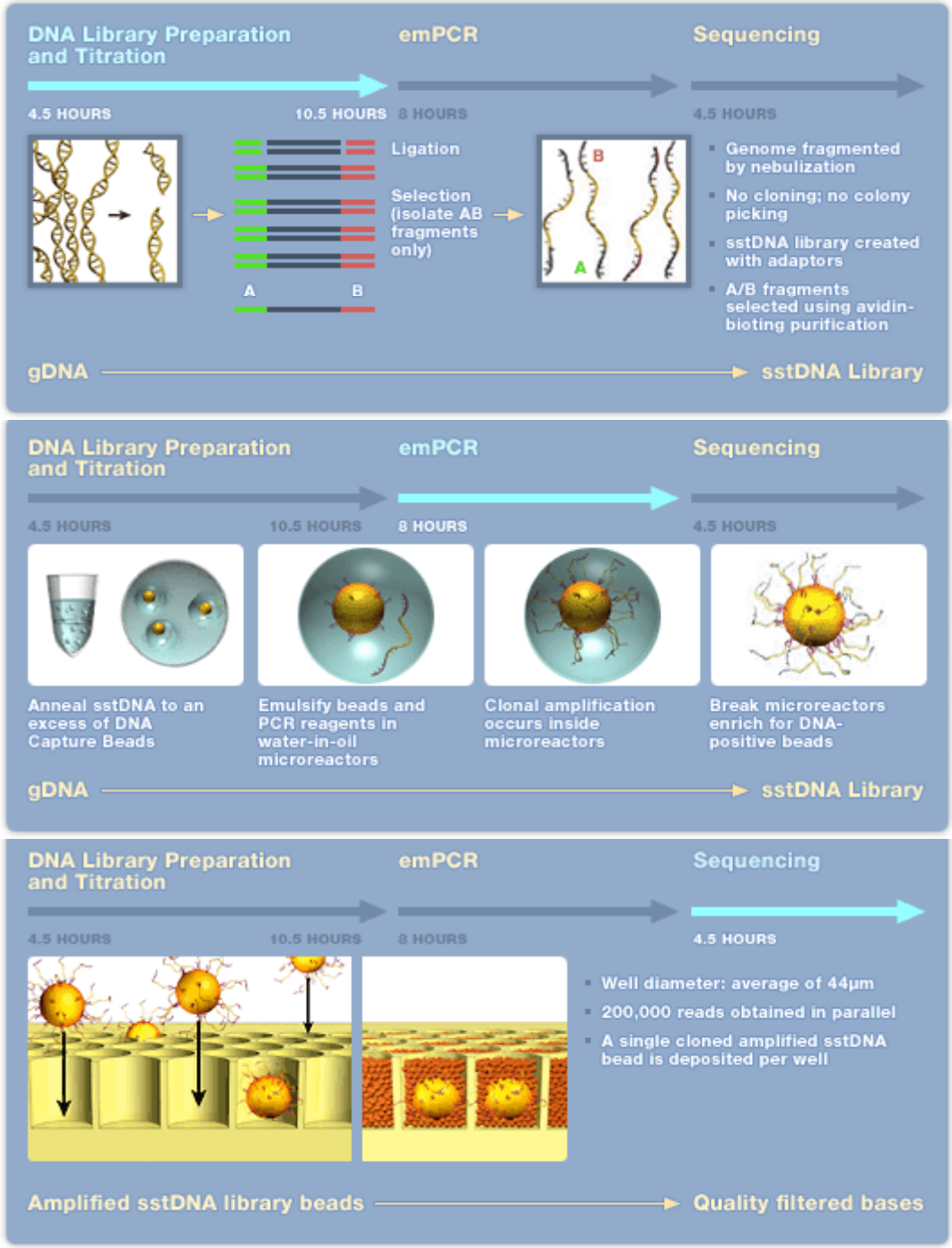
Technologies, Applications & Analysis

NGS

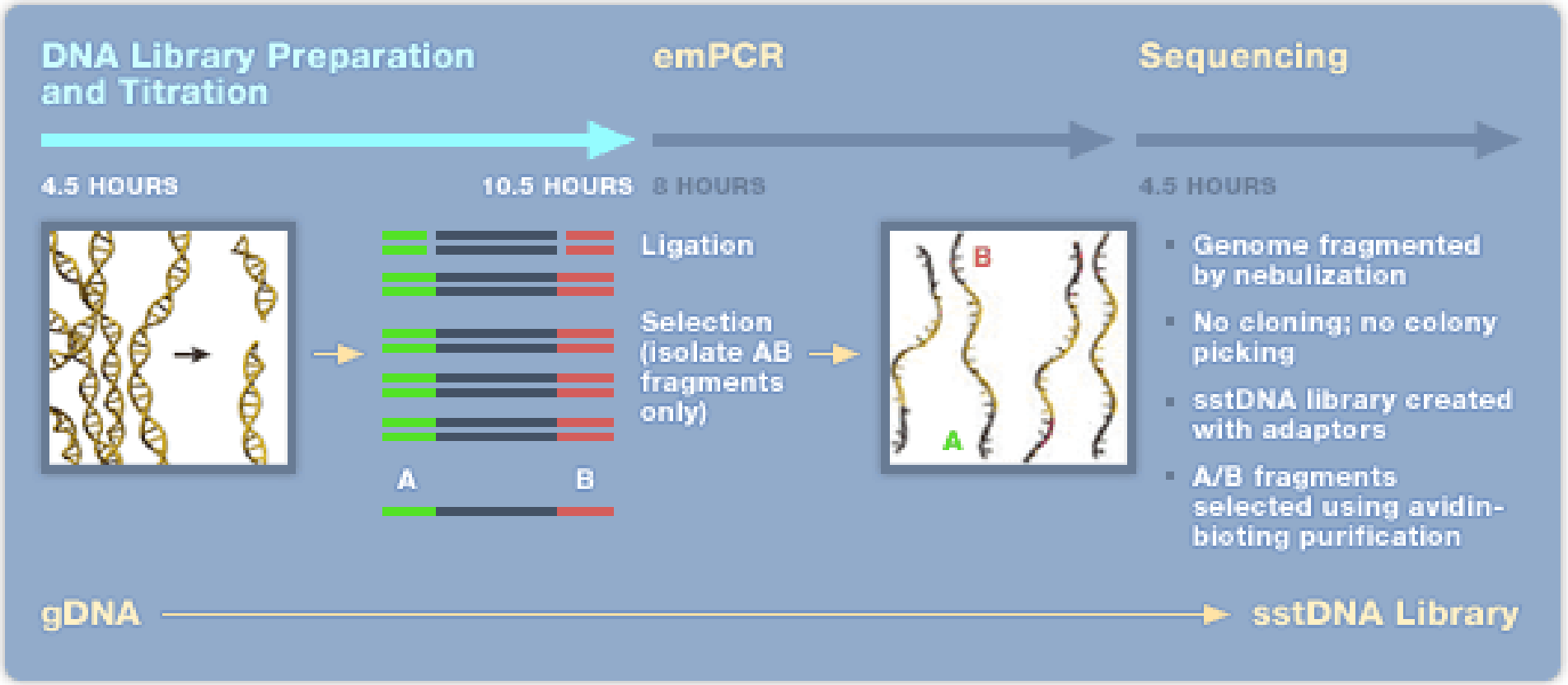
	Feature generation	Sequencing by synthesis
454	Emulsion PCR	Polymerase (pyrosequencing)
Solexa	Bridge PCR	Polymerase (reversible terminators)
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)

Roche 454 SEQUENCING

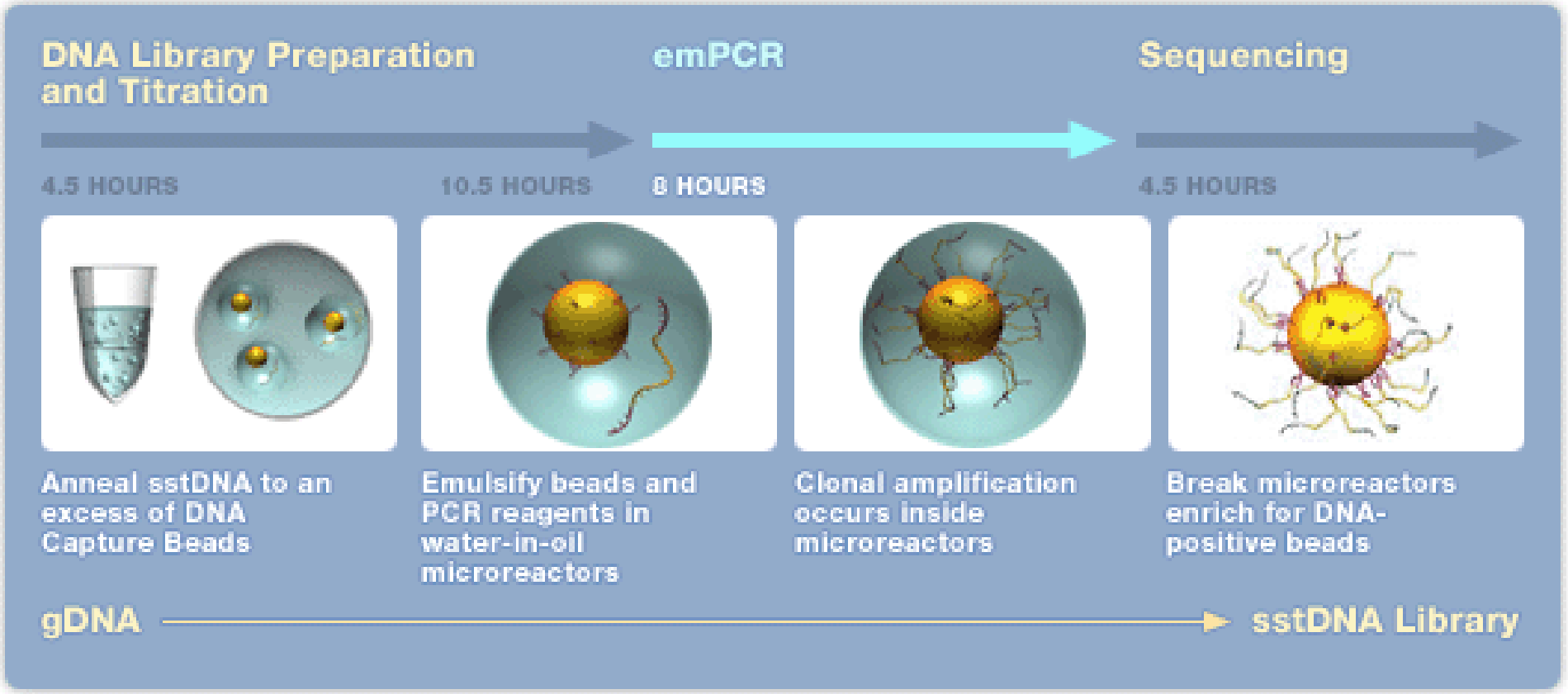
June 2000	454 Life Sciences is founded
October 2005	Release of the Genome Sequencer 20, the first next-generation sequencing system on the market
October 2005	Collaboration agreement signed with Roche Diagnostics
December 2005	454 Life Sciences Awarded the Wall Street Journal's Gold Medal for Innovation
November 2006	454 Life Sciences, in collaboration with Svante Paabo, describes in *Nature* the first million base pairs of the Neanderthal genome and initiates the Neanderthal Genome Project.
January 2007	Release of the Genome Sequencer FLX System
March 2007	Roche Diagnostics completes integration with 454 Life Sciences
May 2007	Complete sequence of Jim Watson published in Nature. First genome to be sequenced for less than $1 million.
November 2007	Announcement of the 100th peer-reviewed publication enabled by 454 Sequencing
June 2008	454 Joins the 1000 Genome Project, an international effort to build the most detailed map to date of human genetic variation as a tool for medical research
September 2008	Announcement of the 250th peer-reviewed publication enabled by 454 Sequencing
October 2008	Release of Genome Sequencer FLX Titanium Series reagents, featuring 1 million reads at 400 base pairs in length

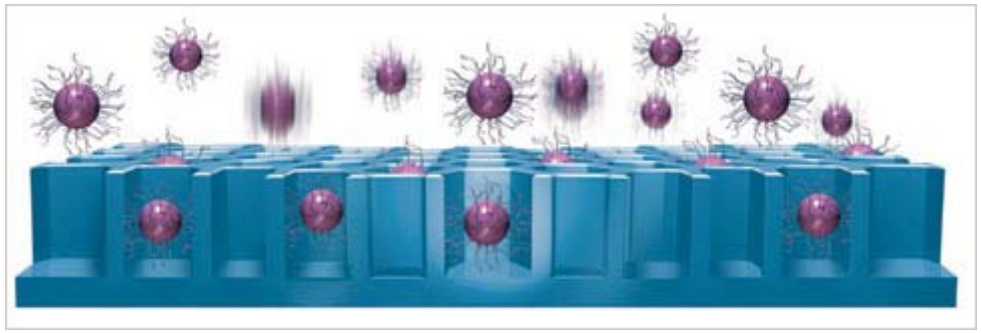

DNA Library Preparation and Titration
emPCR
Sequencing
4.5 HOURS
10.5 HOURS
8 HOURS
4.5 HOURS
Ligation
Selection (isolate AB fragments only)
A
B
B
A
Genome fragmented by nebulization
No cloning; no colony picking
sstDNA library created with adaptors
A/B fragments selected using avidin-bioting purification
gDNA
sstDNA Library
DNA Library Preparation and Titration
emPCR
Sequencing
4.5 HOURS
10.5 HOURS
8 HOURS
4.5 HOURS
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
gDNA
sstDNA Library
DNA Library Preparation and Titration
emPCR
Sequencing
4.5 HOURS
10.5 HOURS
8 HOURS
4.5 HOURS
Well diameter: average of 44µm
200,000 reads obtained in parallel
A single cloned amplified sstDNA bead is deposited per well
Amplified sstDNA library beads
Quality filtered bases

Step 1. DNA Library Preparation

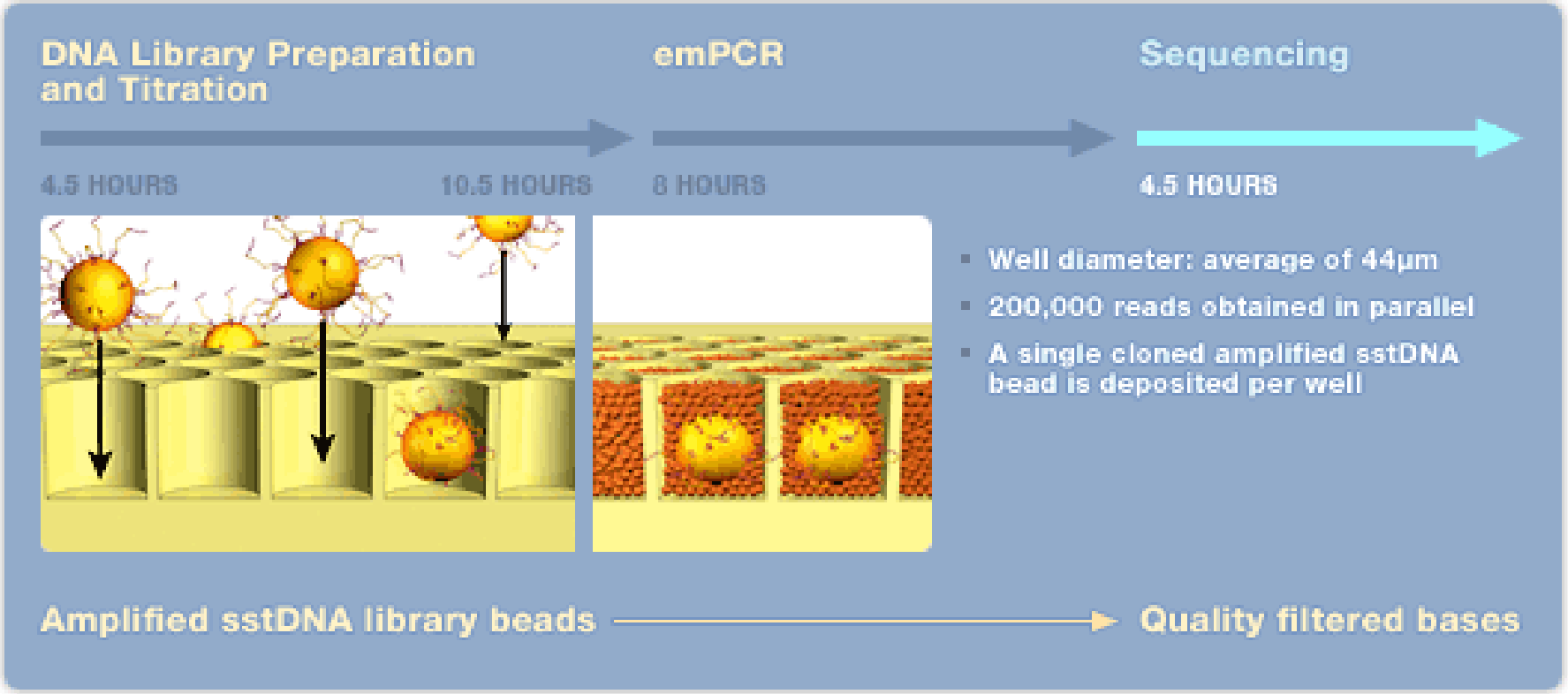


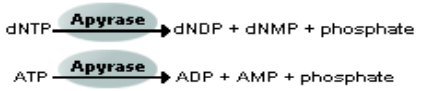
Step 2. Emulsion PCR (emPCR)

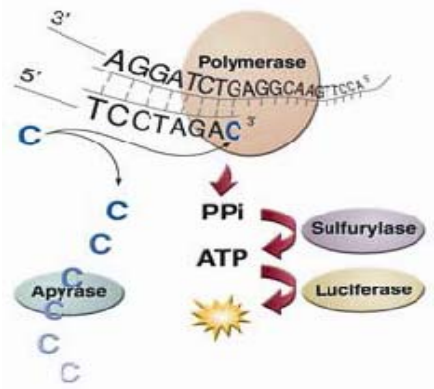




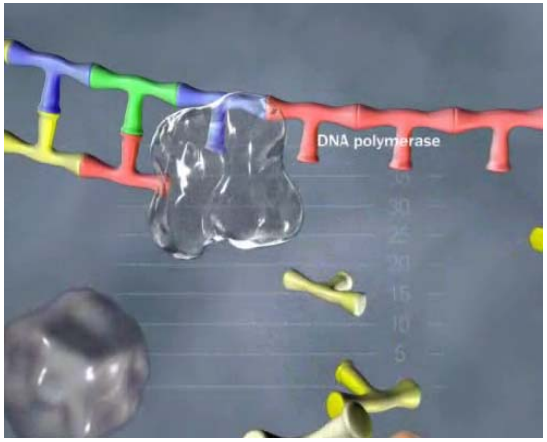
Step 3. Pyrosequencing



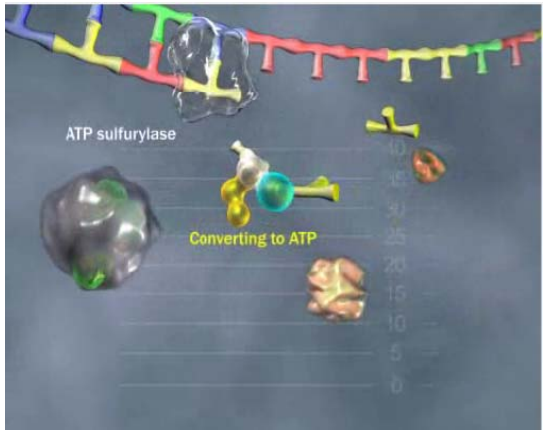




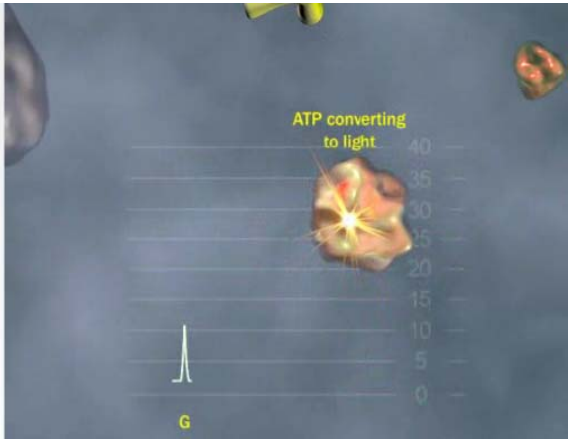
result→read the signal of light



DNA polymerase→ add the A.T.C.G

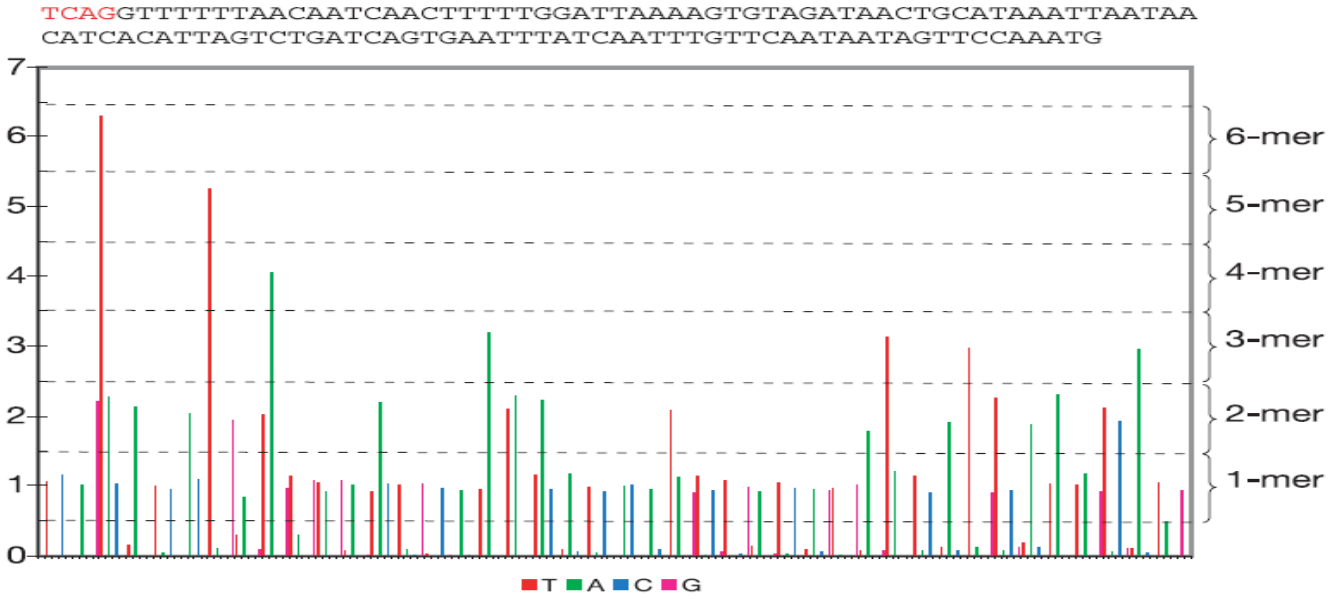


ATP sulfurylase→ convert pyrophosphate to ATP



luciferase→react the ATP with luciferin to generate light

apyrase→ degrade unincorporated dNTPs and excess ATP



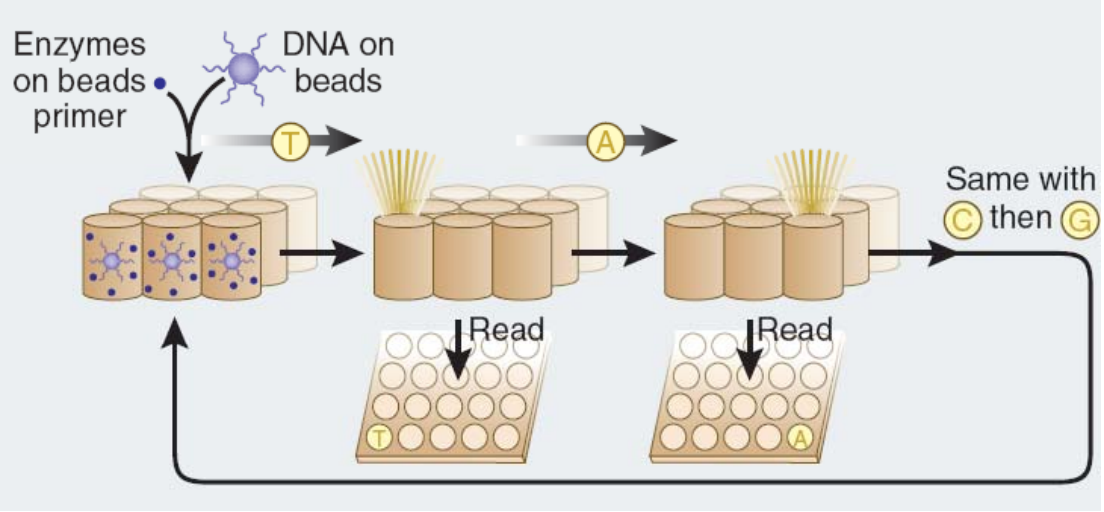
454 technology (Pyrosequencing)

Sample preparation.

Fragments of DNA are ligated to adapters that facilitate their capture on beads (one fragment per bead). A water-in-oil emulsion containing PCR reagents and one bead per droplet is created to amplify each fragment individually in its droplet. After amplification, the emulsion is broken, DNA is denatured and the beads, containing one amplified DNA fragment each, are distributed into the wells of a fiber-optic slide.

Pyrosequencing.

The wells are loaded with sequencing enzymes and primer (complementary to the adapter on the fragment ends), then exposed to a flow of one unlabeled nucleotide at a time, allowing synthesis of the complementary strand of DNA to proceed. When a nucleotide is incorporated, pyrophosphate is released and converted to ATP, which fuels the luciferase-driven conversion of luciferin to oxyluciferin and light. As a result, the well lights up. The read length is between 100 and 150 nucleotides.



454 SEQUENCING

Home | About 454 | Careers | Contact Us | Roche Home | my454 Home

Roche

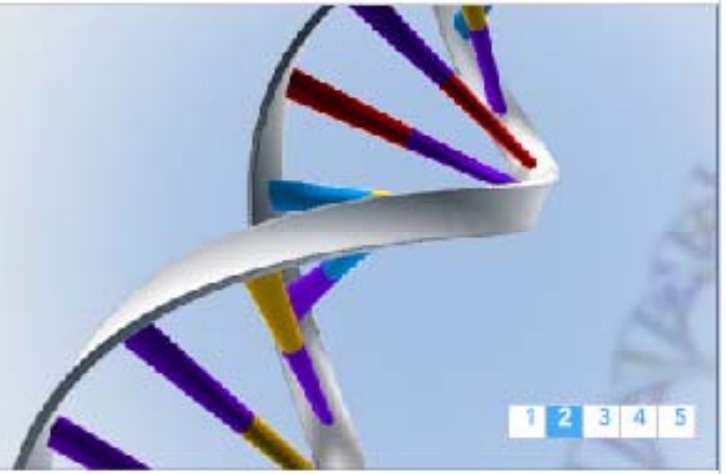
Products & Solutions | Fields of Biology | Applications | Publications & Resources

Search

First look

Raising the standards in sequencing. What's coming next from 454.

See the Future of 454 Sequencing



1 2 3 4 5

News

August 5, 2009
454 Life Sciences and Roche NimbleGen Announce Collaboration with Eli Lilly and SeqWright to Sequence Genomic Regions Associated with Psychiatric Disease.

July 29, 2009
Roche and Google.org Start Initiative for Early Discovery of New Diseases

July 16, 2009
Researchers to Decode Antarctic Ice Metagenome with the 454 Sequencing System, to Explore the Effects of Climate Change

» Read more news

Publications | 525 and counting

> The novel polysaccharide deacetylase homolog Pdi contributes to virulence of the aquatic pathogen Streptococcus iniae. Microbiology.

> Genomic Diversity and Evolution of Mycobacterium ulcerans Revealed by Next-Generation Sequencing. PLoS Pathogens.

> Recognition and coupling of A-to-I edited sites are determined by the tertiary structure of the RNA. Nucleic Acids Research.

» Read more publications

Stimulus Funding Grant Support
» Read more

Contact Us / Support

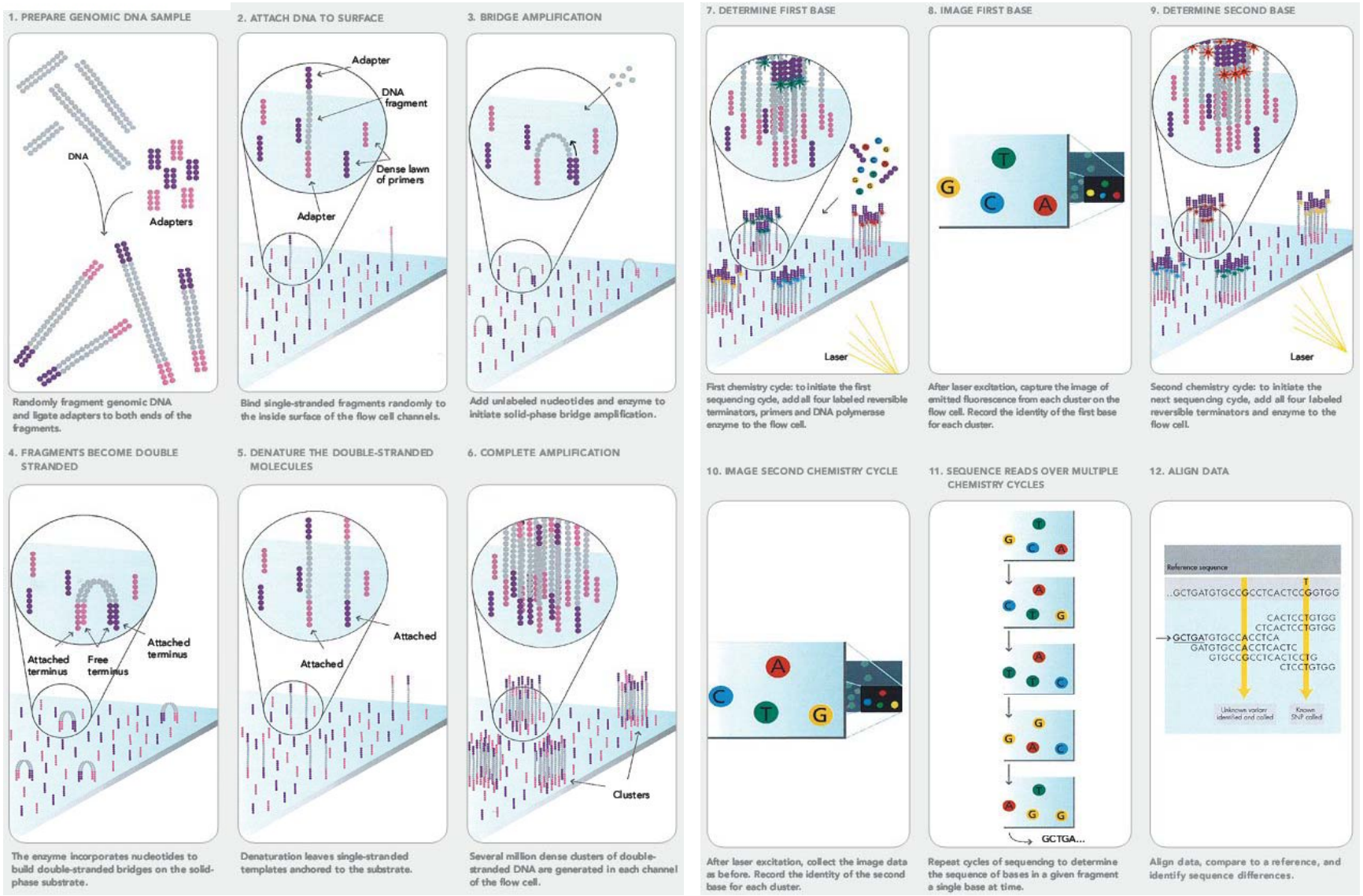
Get in touch with us directly! Select your area of interest:

Select ....

Solexa Genome Analyzer

illumina
Solexa

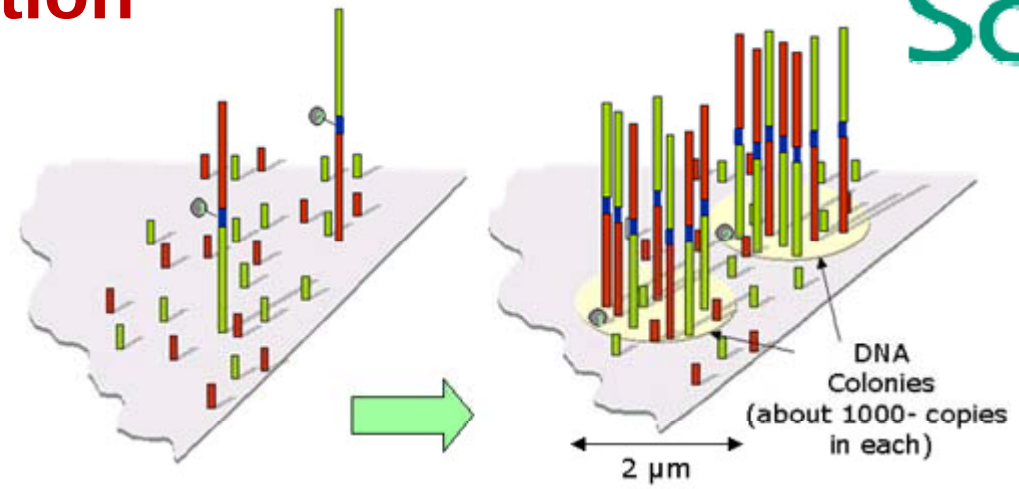
Solexa Genome Analyzer



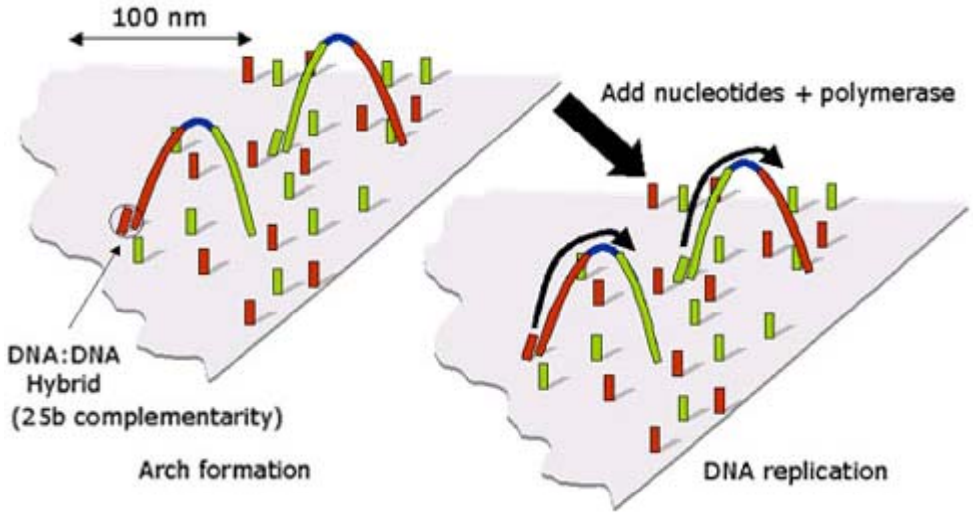
Cluster Generation

Solexa™

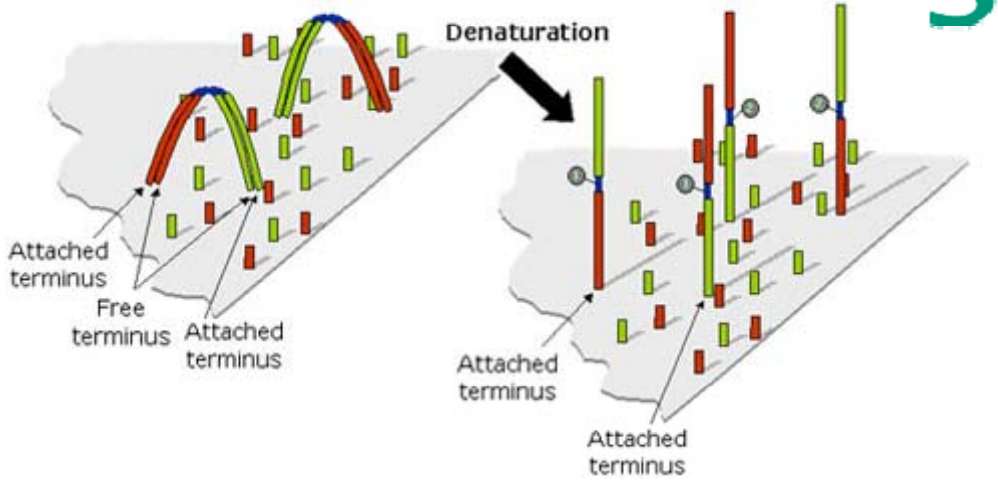
Clusters Array



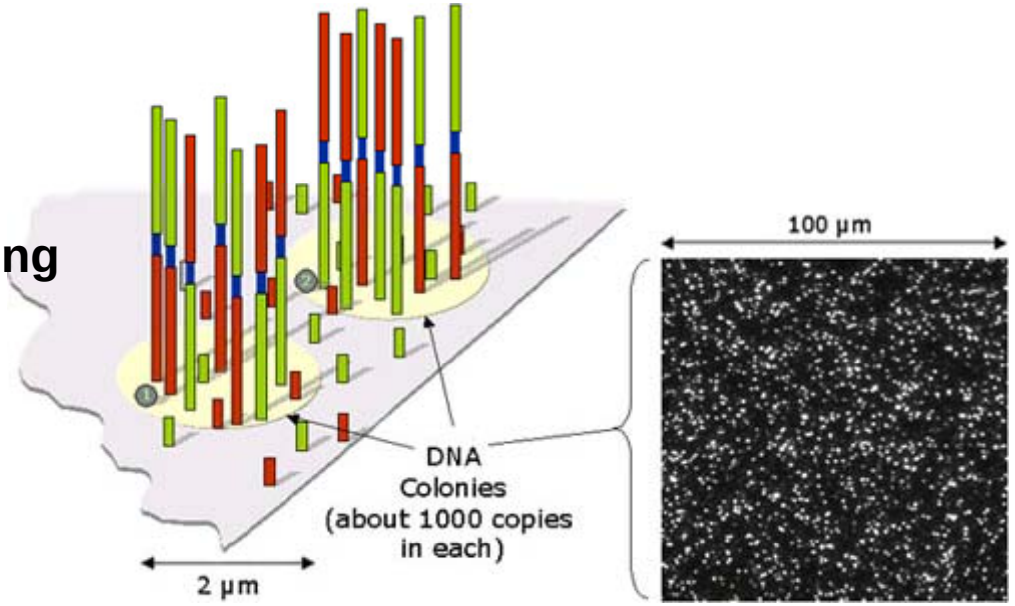
Template Bridging



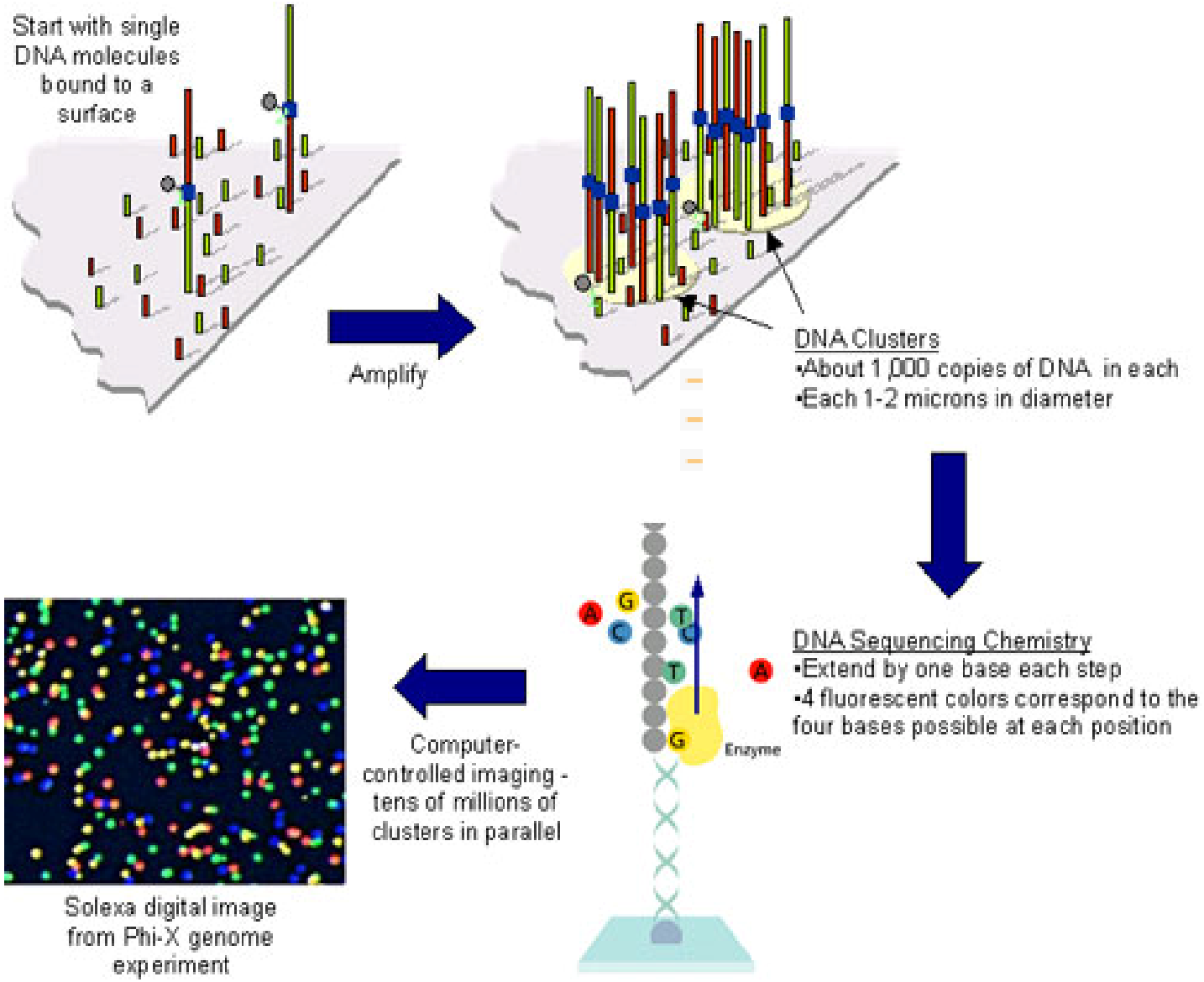
Denaturation



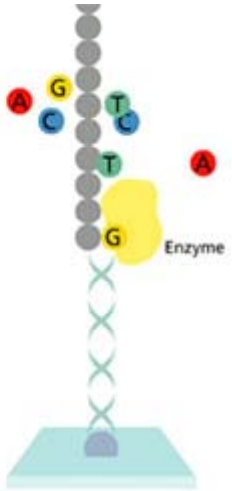
Ready For Sequencing



Sequencing-By-Synthesis

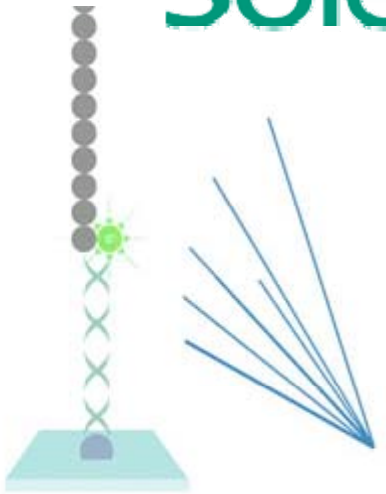


Cycle 1



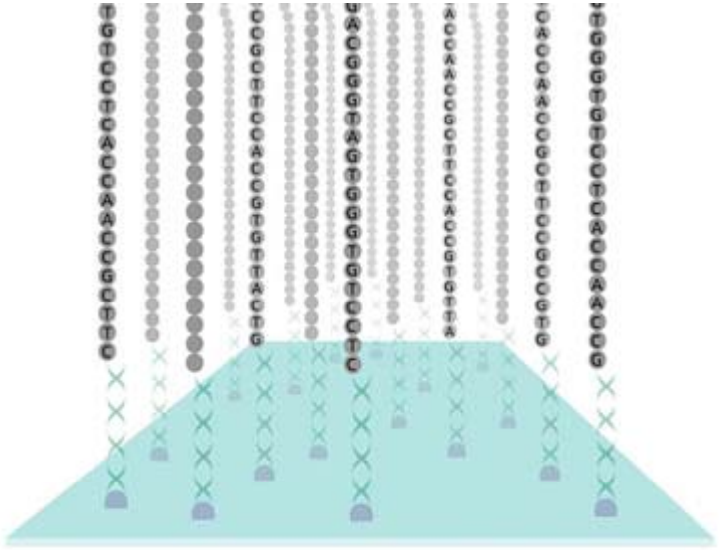
Incorporation

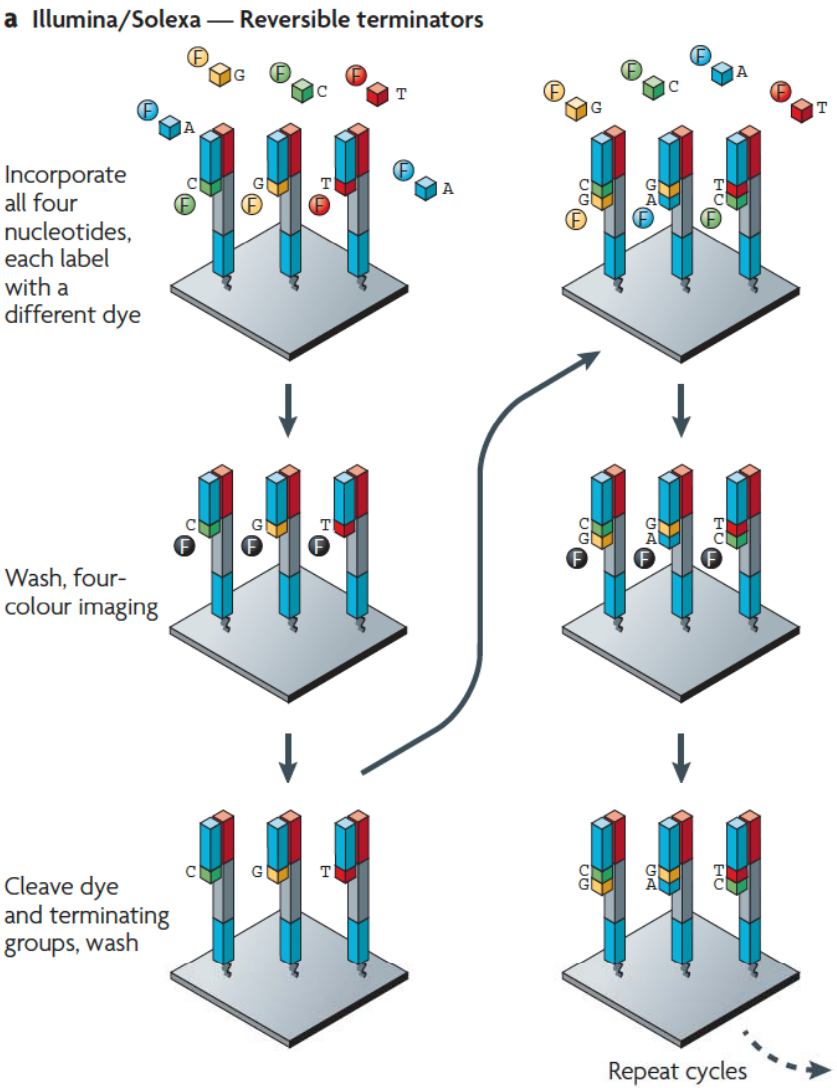
Cycle 1

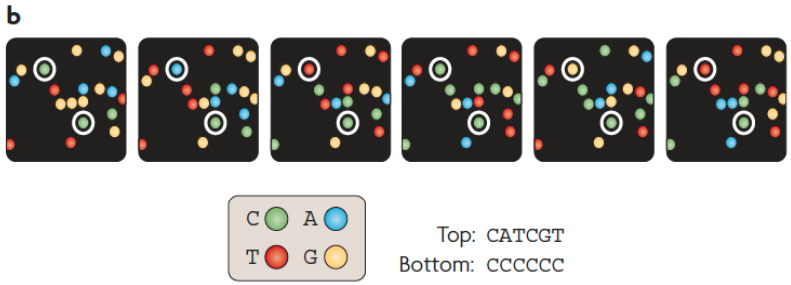


Fluorescent Detection

Sequence Generated At Every Site On The Array

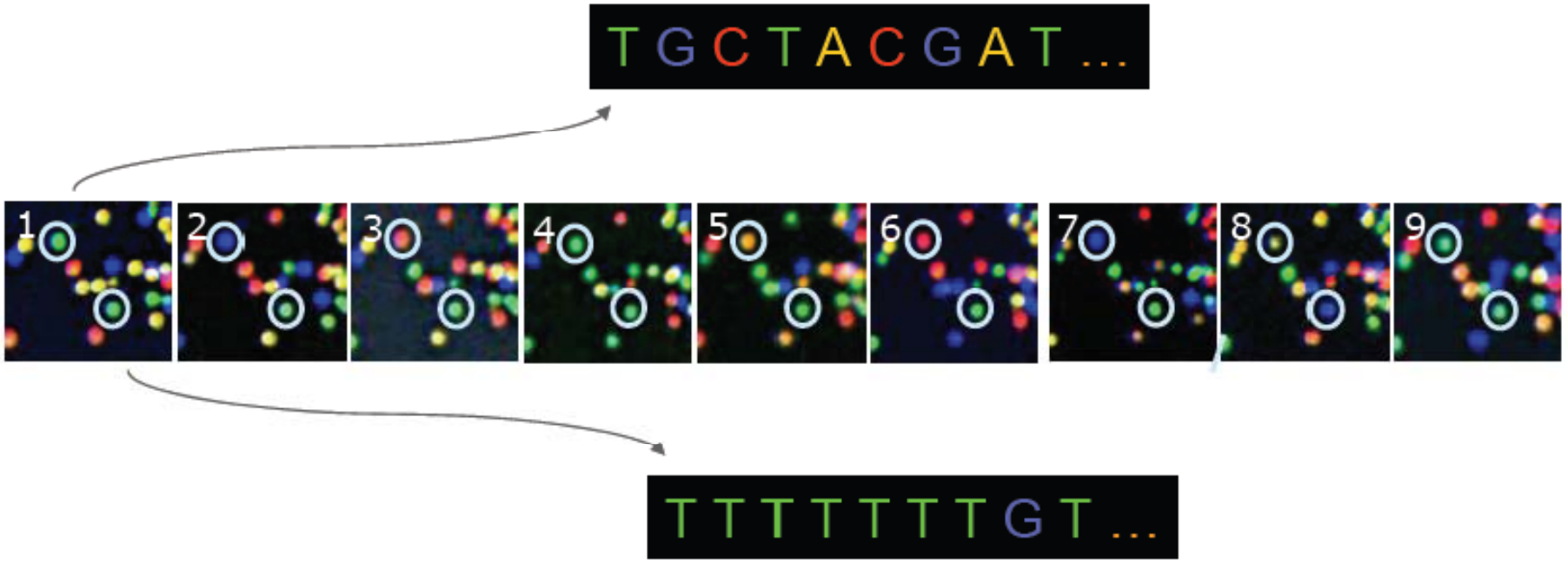






From Debbie Nickerson, Department of Genome Sciences, University of Washington, http://tinyurl.com/6zbzh4

Base calling from raw data



The identity of each base of a cluster is read off from sequential images

From Debbie Nickerson, Department of Genome Sciences, University of Washington, http://tinyurl.com/6zbzh4

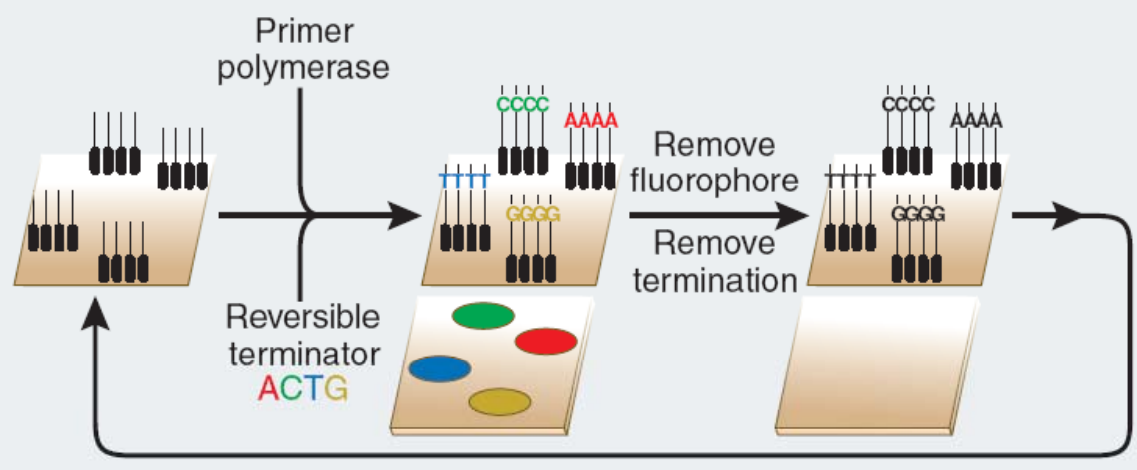
Solexa technology (sequencing-by-synthesis)

Sample preparation.

Fragments of DNA are ligated to end adapters, denatured and bound at one end to a solid surface already coated with a dense layer of the adapters. Each single stranded fragment is immobilized at one end, while its free end ‘bends over’ and hybridizes to a complementary adapter on the surface, which initiates the synthesis of the complementary strand in the presence of amplification reagents. Multiple cycles of this solid-phase amplification followed by denaturation create clusters of ~1,000 copies of single-stranded DNA molecules distributed randomly on the surface.

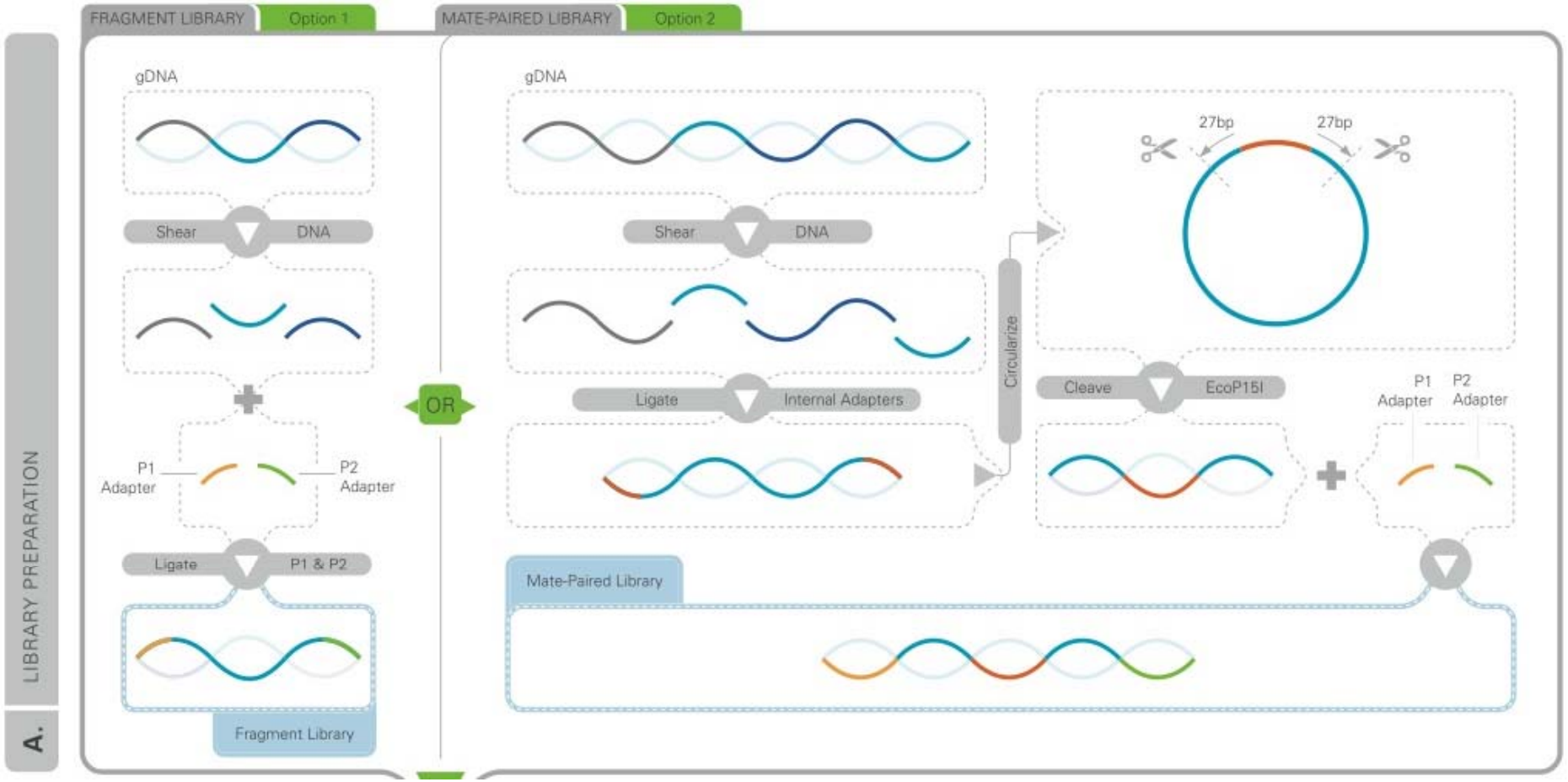
Sequencing with reversible terminators.

Synthesis reagents, added to the flow cell, consist of primers, DNA polymerase and four differently labeled, reversible terminator nucleotides. After incorporation of a nucleotide, which is identified by its color, the 3′ terminator on the base and the fluorophore are removed, and the cycle is repeated for a read length of 30–35 nucleotides.

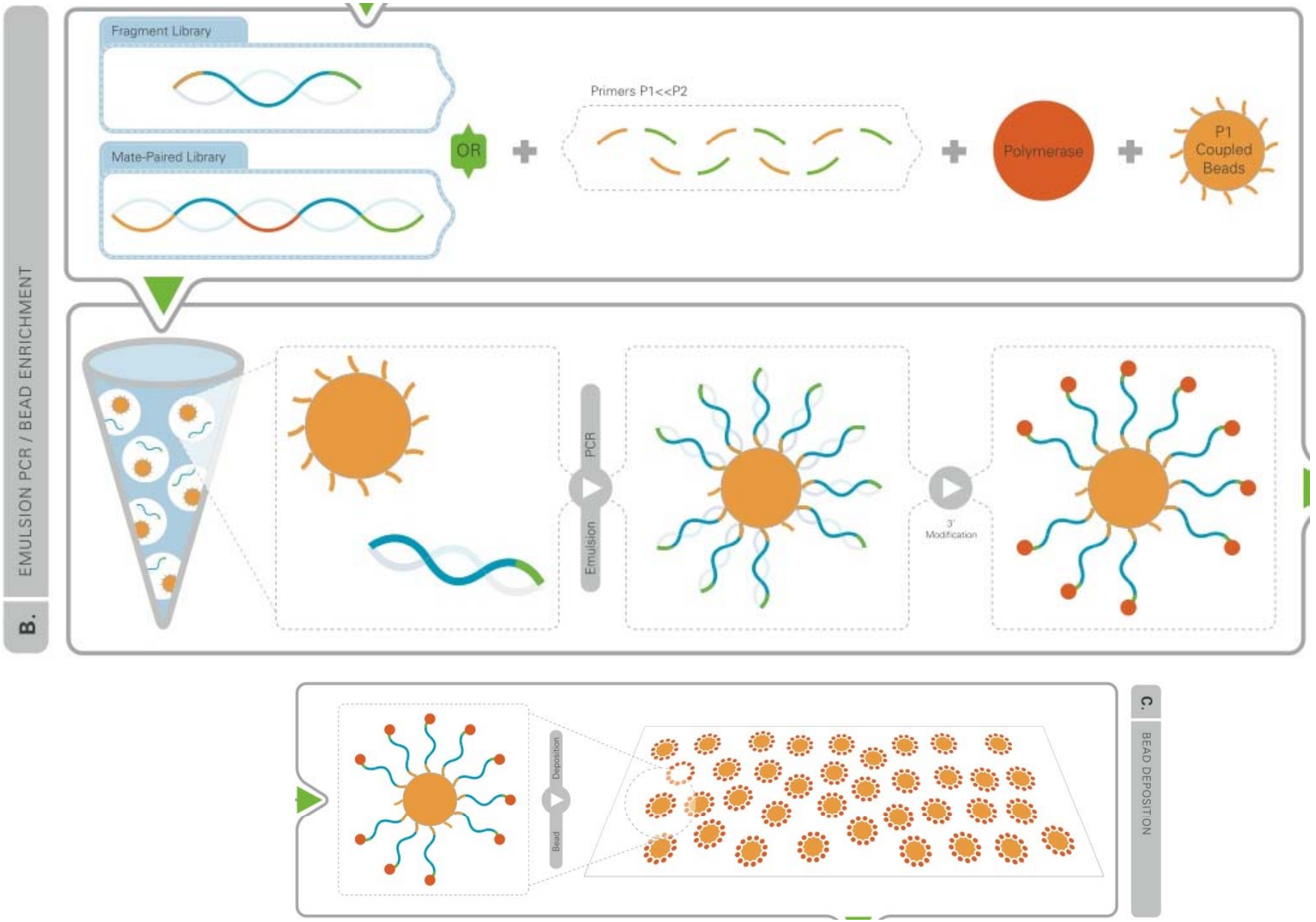


Applied Biosystems SOLiD™ System 2.0

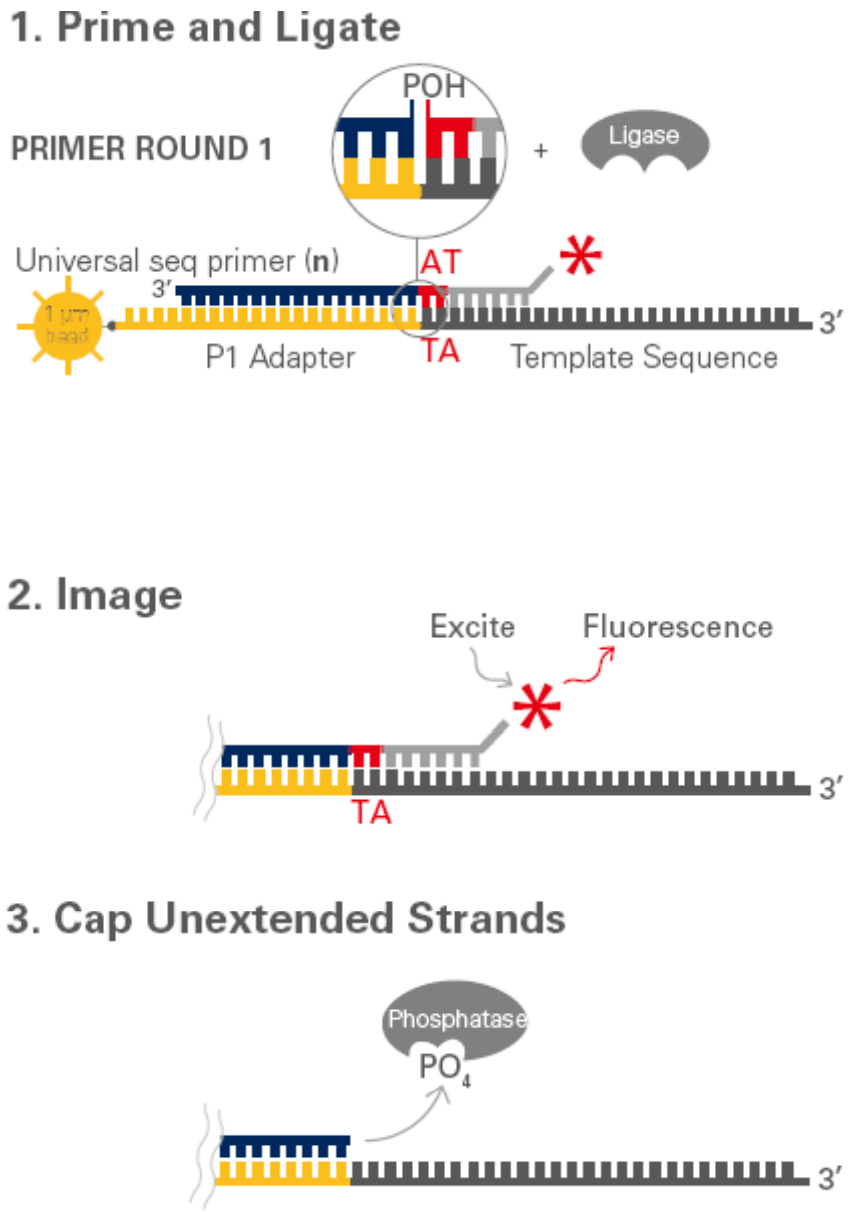
Step 1. Library preparation



Step 2. Emulsion PCR

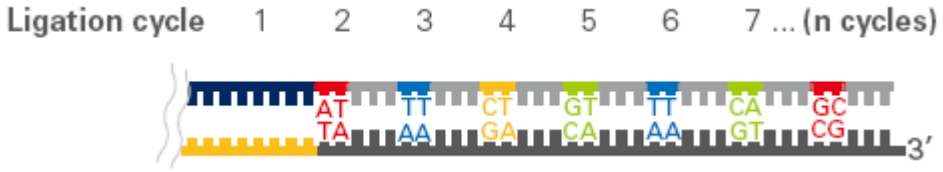


Step 3. Sequencing-by-Ligation

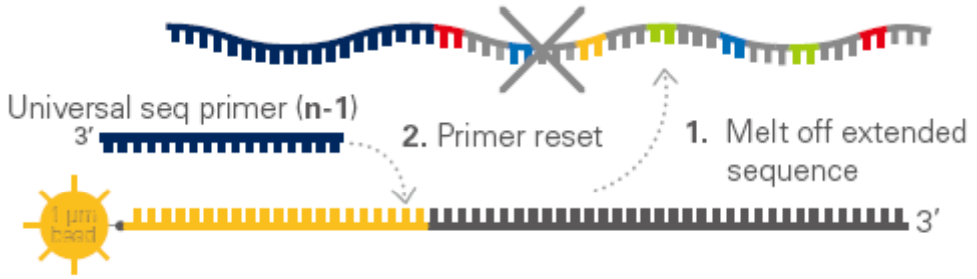


4. Cleave off Fluor

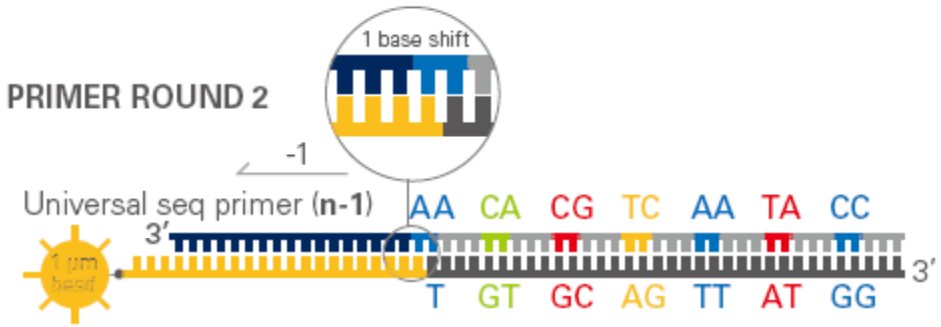
5. Repeat steps 1-4 to Extend Sequence



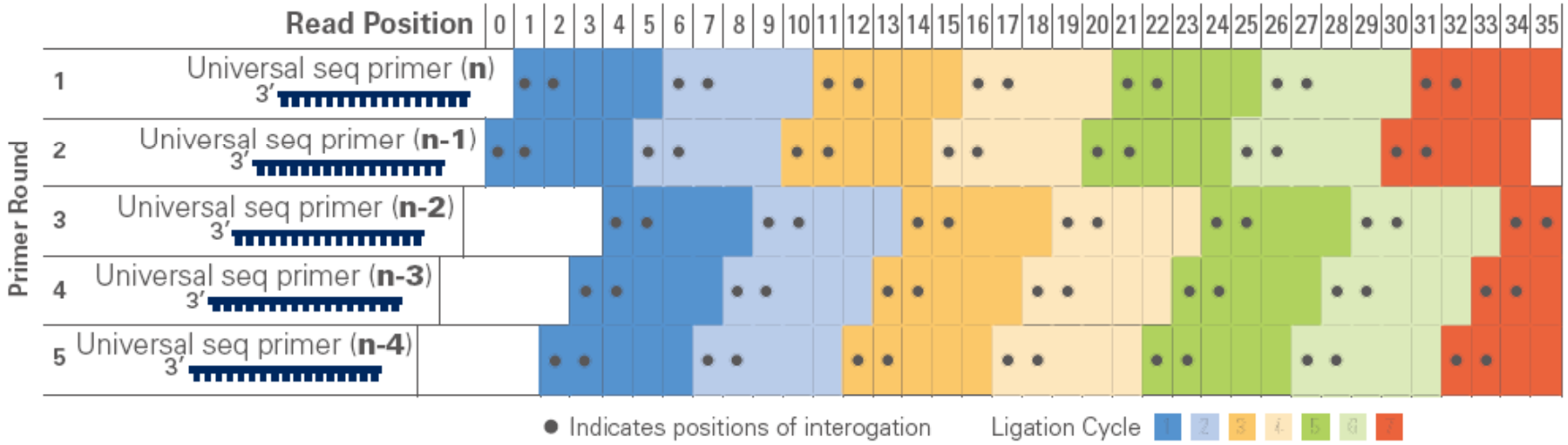
6. Primer Reset



7. Repeat steps 1-5 with new primer



8. Repeat Reset with , n-2, n-3, n-4 primers



Its format is
>TAG_ID
Color_space

```
>1_88_1830_R3
G32113123201300232320
>1_89_1562_R3
G23133131233333101320
```

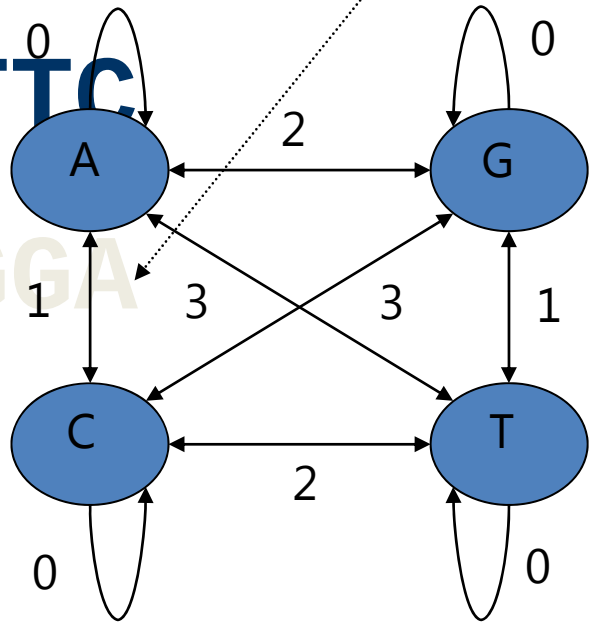
AB SOLiD: Dibase Sequencing

AB SOLiD reads look like this:

T012233102 **T012033102**
TGAGCGTTC **TGAATAGGA**

	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

TGAGCGTTC
TGAATAGGA



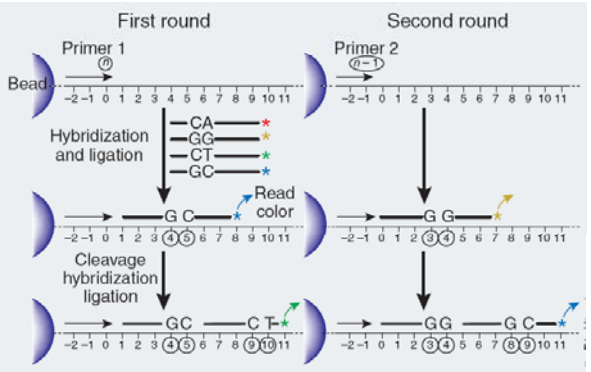
SOLiD technology (sequencing-by-ligation)

Sample preparation.

Fragments of DNA are ligated to adapters and amplified on beads by emulsion PCR. The DNA is denatured and the beads deposited onto a glass slide.

Sequencing by ligation.

A sequencing primer is hybridized to the adapter and its 5' end is available for ligation to an oligonucleotide hybridizing to the adjacent sequence. A mixture of octamer oligonucleotides compete for ligation to the primer (the bases in fourth and fifth position on these oligos are encoded by one of four color labels). After its color has been recorded, the ligated oligonucleotide is cleaved between position 5 and 6, which removes the label, and the cycle of ligation-cleavage is repeated. In the first round, the process determines possible identities of bases in positions 4, 5, 9, 10, 14, 15, etc. The entire process is repeated, offset by one base by using a shorter sequencing primer, to determine positions 3, 4, 8, 9, 13, 14, etc., until the first base in the sequencing primer (position 0) is reached. Since the identity of this base is known, the color is used to decode its neighboring base at position 1, which in turn decodes the base at position 2, etc., until all sequence pairs are identified. The current read length is between 30 and 35 nucleotides



Evolution of Sequencing Technology

Sanger dideoxy-sequencing

ABI 3730XL

Massive parallel sequencing

Roche 454 FLX, Illumina Genome Analyzer, Life Technologies SOLiD

Bead-based em-PCR and sequencing by ligation

Dover Systems' Polonator

Massive parallel sequencing and single molecule sequencing

Pacific Biosciences (single-molecule real-time DNA sequencing (SMRT) technology)
Helicos (true single-molecule-sequencing (tSMS) technology)
VisiGen Biotechnologies (real-time, single-molecule sequencing fluorescence resonance energy transfer (FRET) technology)

Single molecule sequencing and nanopore technology?

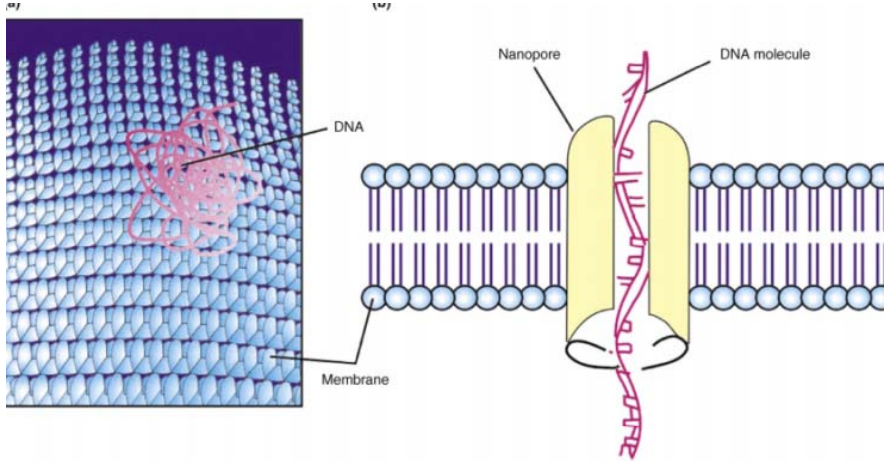
Oxford Nanopore Technologies (label-free, single-molecule sequencing (BASE) technology), Affymetrix, Reveo, Base4innovation, Genome Corp, and Complete Genomics.

Nanopore Sequencing

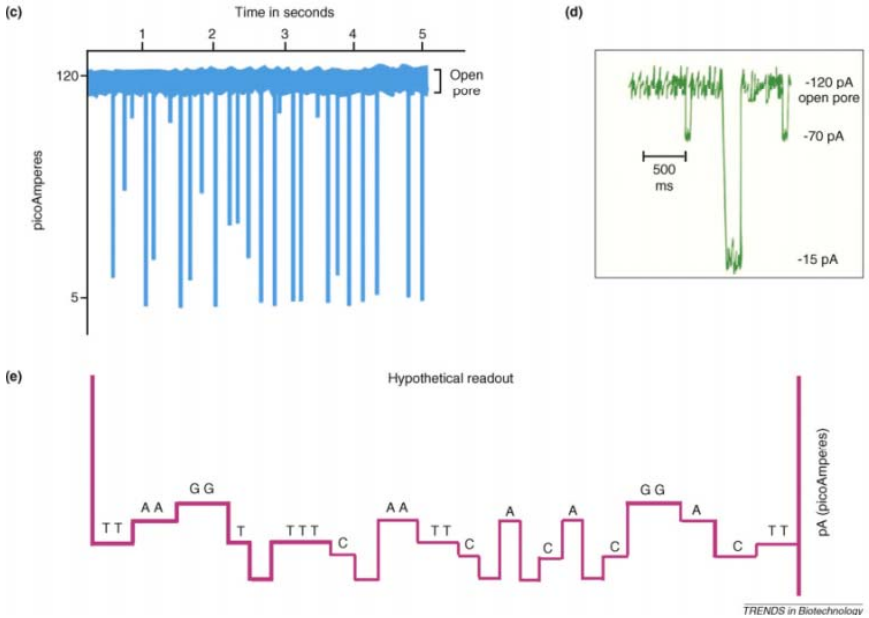
Ion Torrent

Oxford Nanopore

Nucleic acids driven through a nanopore.



Differences in conductance of pore provide readout.

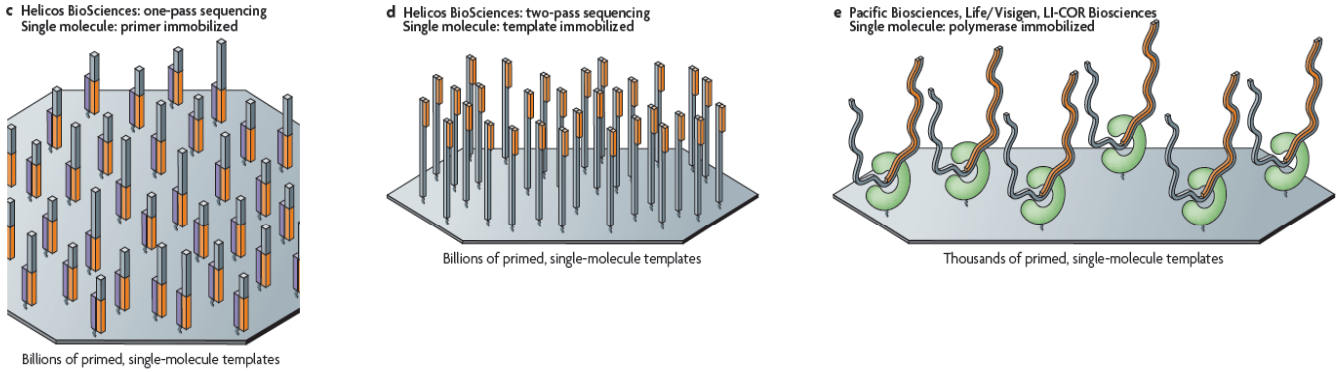


Single Molecule Real-Time Sequencing

Real-time monitoring of PCR activity

Read-out by fluorescence resonance energy transfer between polymerase and nucleotides or

Waveguides allow direct observation of polymerase and fluorescently labeled nucleotides



Helicos Biosciences

VisGen Biotechnologies

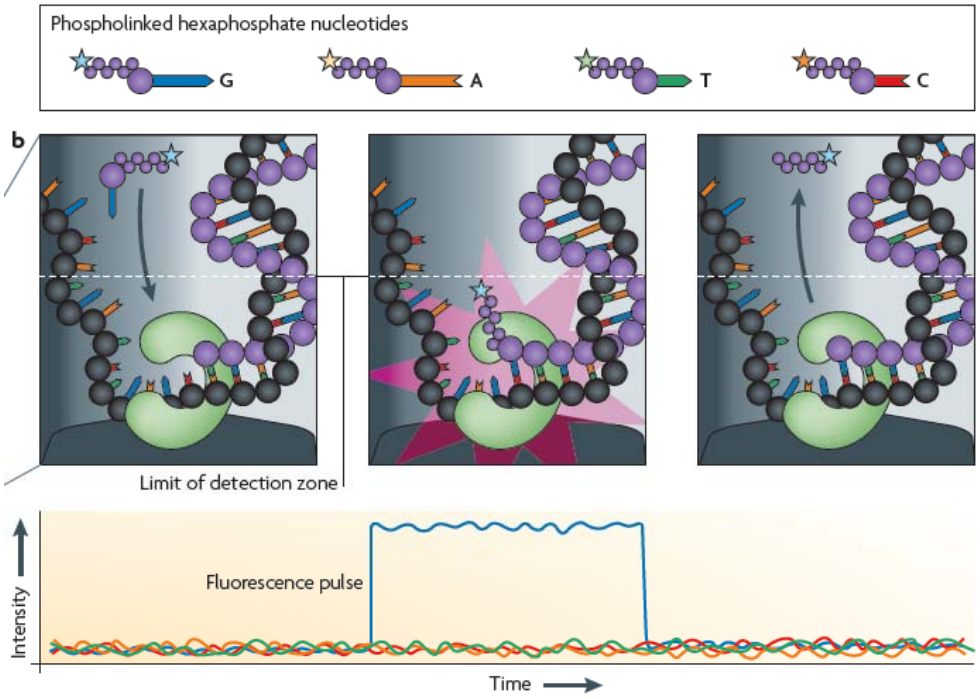
Pacific Biosciences

Single Molecule Real-Time (SMRT) Sequencing

www.sciencemag.org SCIENCE VOL 323 2 JANUARY 2009 133

Real-Time DNA Sequencing from Single Polymerase Molecules

John Eid,* Adrian Fehr,* Jeremy Gray,* Khai Luong,* John Lyle,* Geoff Otto,* Paul Peluso,* David Rank,* Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach,† Stephen Turner†



Pacific Biosciences — Real-time sequencing

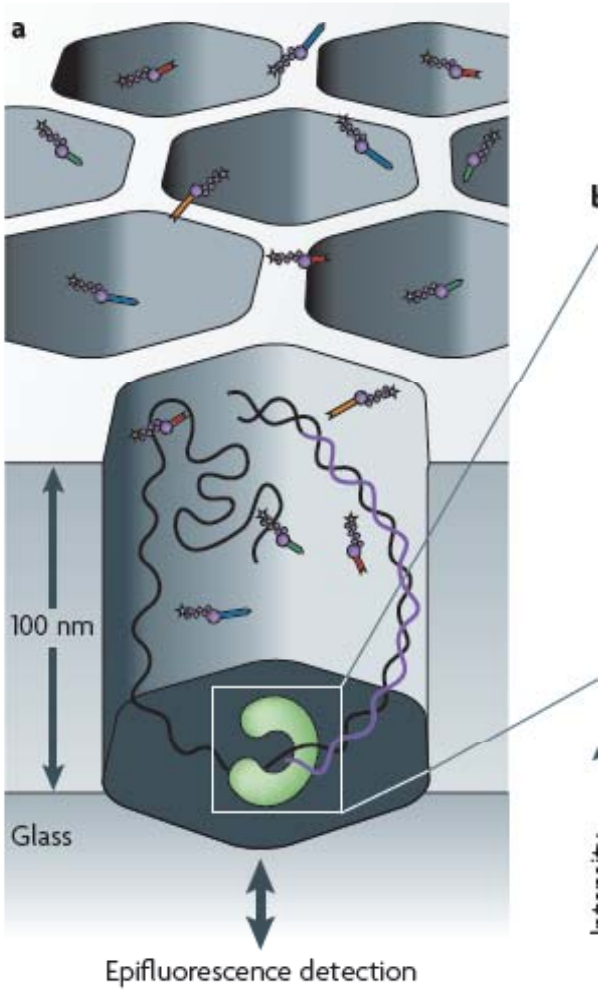
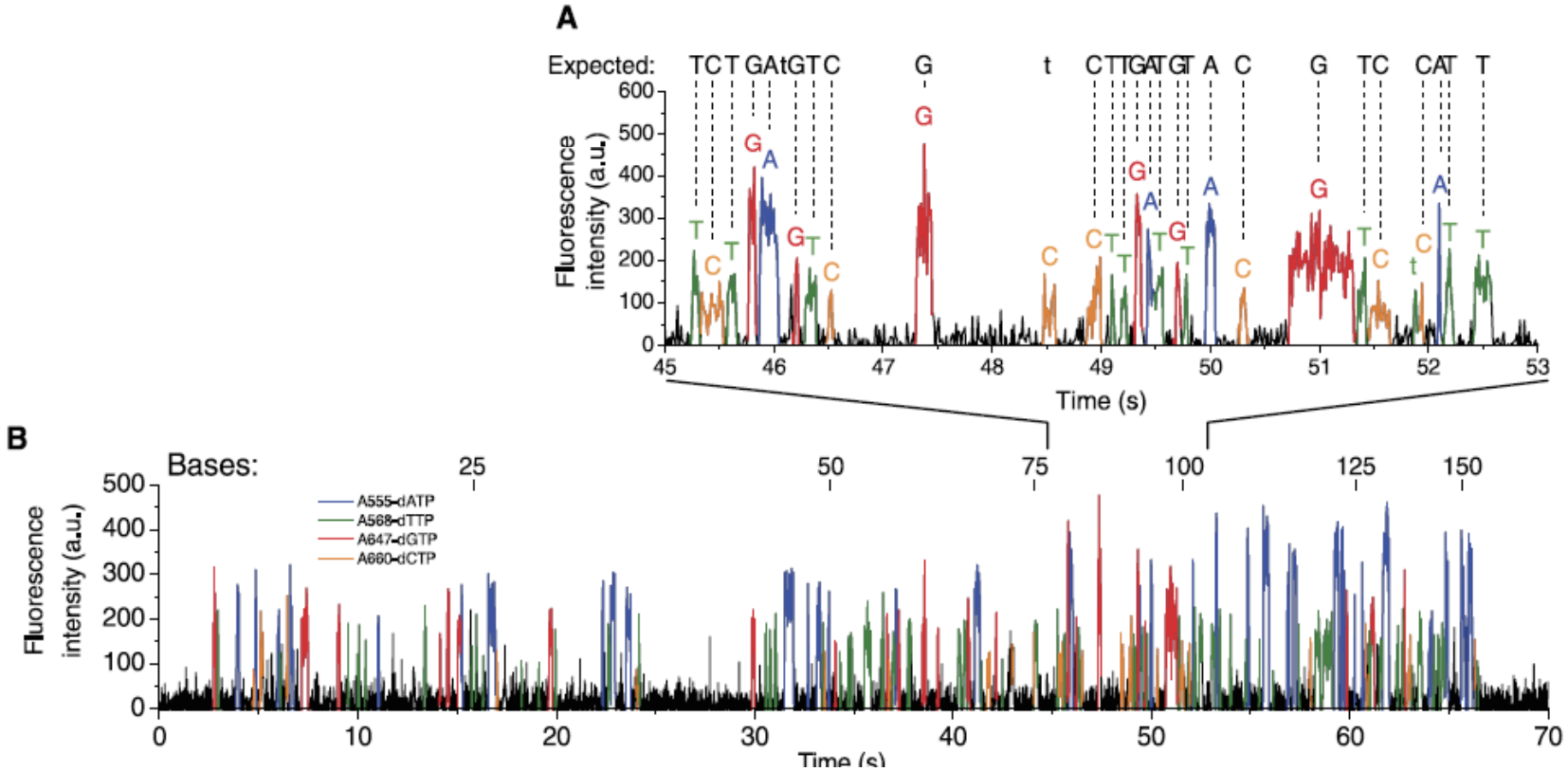


Figure 10. Processive Synthesis with Phospholinked Nucleotides.

Step 1: Fluorescent phospholinked labeled nucleotides are introduced into the ZMW.
Step 2: The base being incorporated is held in the detection volume for tens of milliseconds, producing a bright flash of light.
Step 3: The phosphate chain is cleaved, releasing the attached dye molecule.
Step 4-5: The process repeats.



What makes the differences?

easier library preparation, shorter run time, longer read length, higher throughput, lower cost

Summary of second-generation sequencers

Sequencing platform	Sample requirements	Length of library prep/feature generation (days)	Method of feature generation	Sequencing chemistry	Read length (bases)	Run time	Throughput/run (Gb)	Throughput/day (Gb)
Roche 454 (FLX-Titanium)	1 µg for shotgun library, 5 µg for paired end	3–4	Bead-based/emulsion PCR	Pyrosequencing	400–500	10 h	0.4–0.5	~1
Illumina Genome Analyzer (GAII)	<1 µg for single or paired-end libraries	2	Isothermal 'bridge amplification' on flowcell surface	Reversible terminator SBS	35–75	2 days for 36-cycle single-end run, 4 days for 36-cycle paired-end run	3–6	1.5
ABI SOLiD	<2 µg for shotgun library, 5–20 µg for paired end	2–4.5	Bead-based/emulsion PCR	Ligation	25–75	6–7 days for fragment libraries, 8 days for 2 × 25 base paired-end libraries	10–20	1.7–2

Pacific Biosciences

1 day

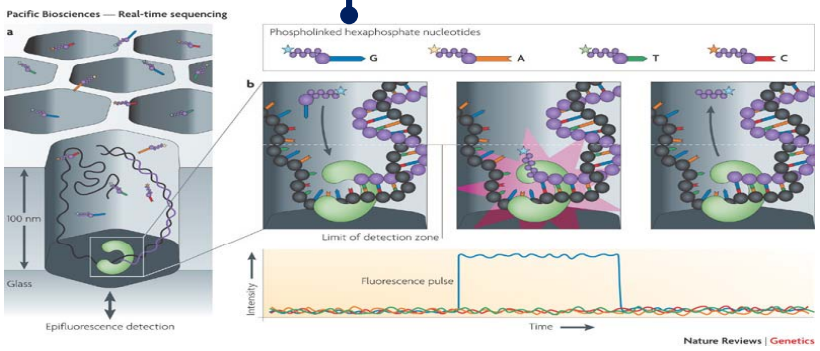
zero-mode waveguides array

1000-2000

30 min
Cost: US$100

0.8 Gb

38.4 GB



Single-molecule real-time DNA sequencing (SMRT)

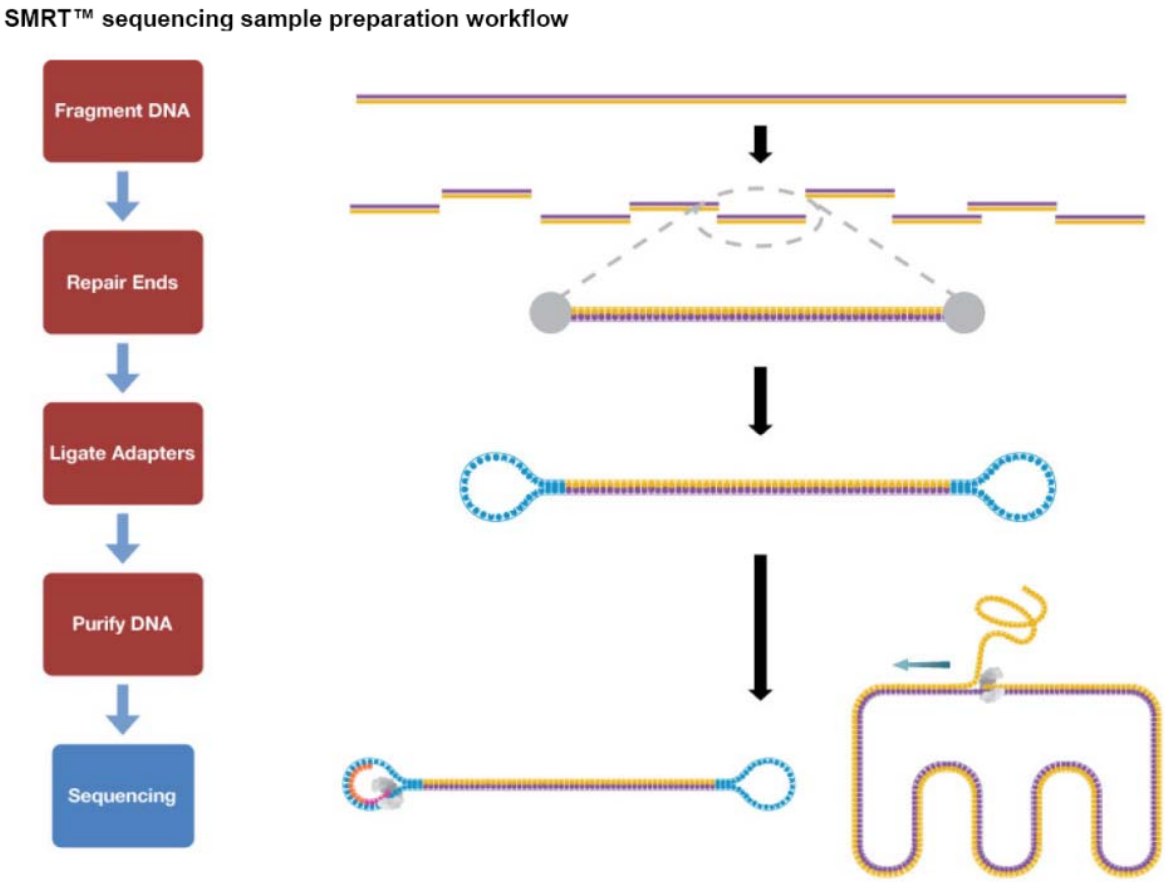
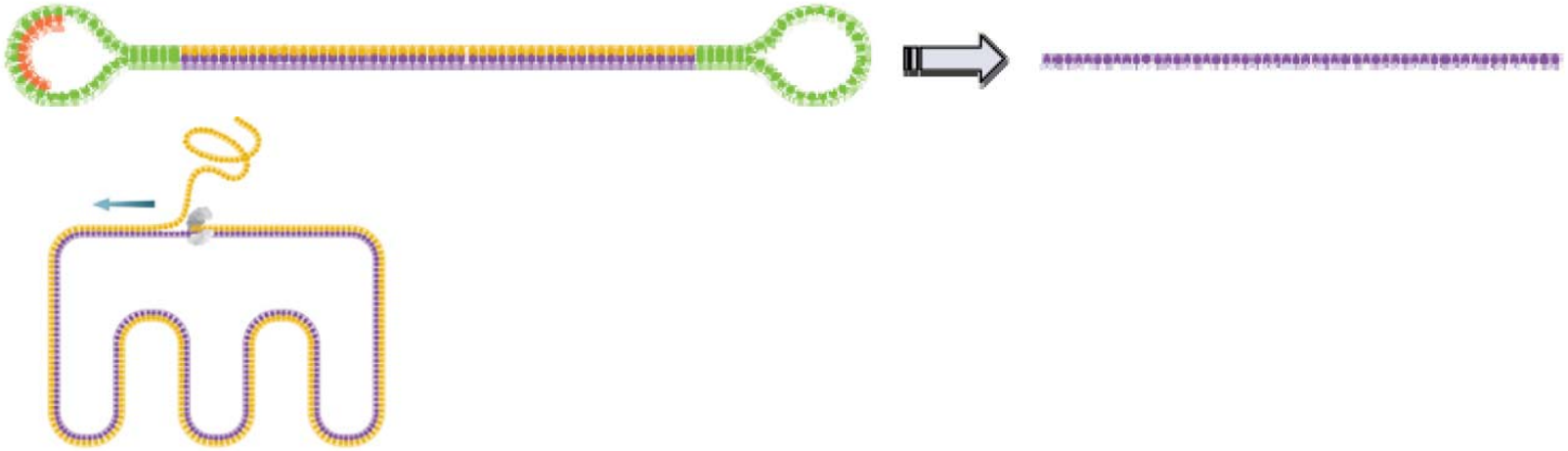


Figure 17. Sample Prep Workflow.
The input sample is first fragmented to the desired size. The ends are repaired and the hairpin structures are ligated to the ends of each fragment. A size selection and purification step selects those fragments with the adaptors attached to both ends. The SMRTbell templates then can go through the sequencing reaction. A strand displacing polymerase enzyme opens the SMRTbell into a circular template and can generate independent reads, both forward and reverse of the same DNA molecule. The quality score increases linearly with the number of times the molecule is sequenced.

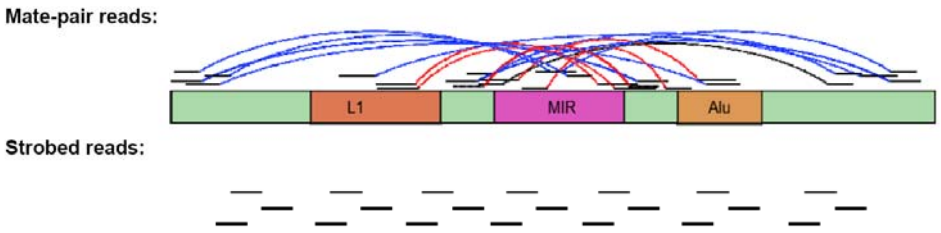
Standard sequencing protocol.

Enzyme processivity enables long readlengths while the speed of synthesis drives fast time to results.



Strobe sequencing protocol.

Strobe sequencing offers greater flexibility and eliminates the need to create multiple libraries of different sizes.



ABI 3730XL

- up to 1100 bases/read
- 96 reads/run
- approx. 1MB/day and machine

LONG READS

First choice for finishing projects; full length cDNA sequencing; single sample sequencing

Roche GS FLX

- in average 250 bases/read
- up to 400 000 reads/run
- up to 100MB/run/7.5 hours

EXTREME SPEED

Optimal for *de novo* sequencing (procaryots & eucaryots) metagenomes, transcriptomes/cDNA, BAC/fosmid pools

Illumina Genome Analyzer

- up to 50 bases/read
- up to 60 000 000 reads/run (paired-end)
- up to 2000MB/run/6.5 days
- Sequencing by synthesis

ABI SOLiD

- up to 35 bases/read
- up to 85 000 000 reads/run (paired-end)
- up to 3000MB/run/6 days
- Sequencing by ligation

HUGE DATA QUANTITY

Highly attractive for resequencing projects of e.g. production strains; small RNAs, SAGE/CAGE and ChIP-Seq; ultra deep mutation/SNPs

VIDEOS

454 SEQUENCING

http://www.454.com/products-solutions/multimedia-presentations.asp

Genome Sequencer FLX Multimedia Presentation

Genome Sequencer FLX Standard Series Workflow Presentation

Genome Sequencer FLX Amplicon Sequencing Presentation

Solexa™ http://www.illumina.com/technology/sequencing_technology.ilmn

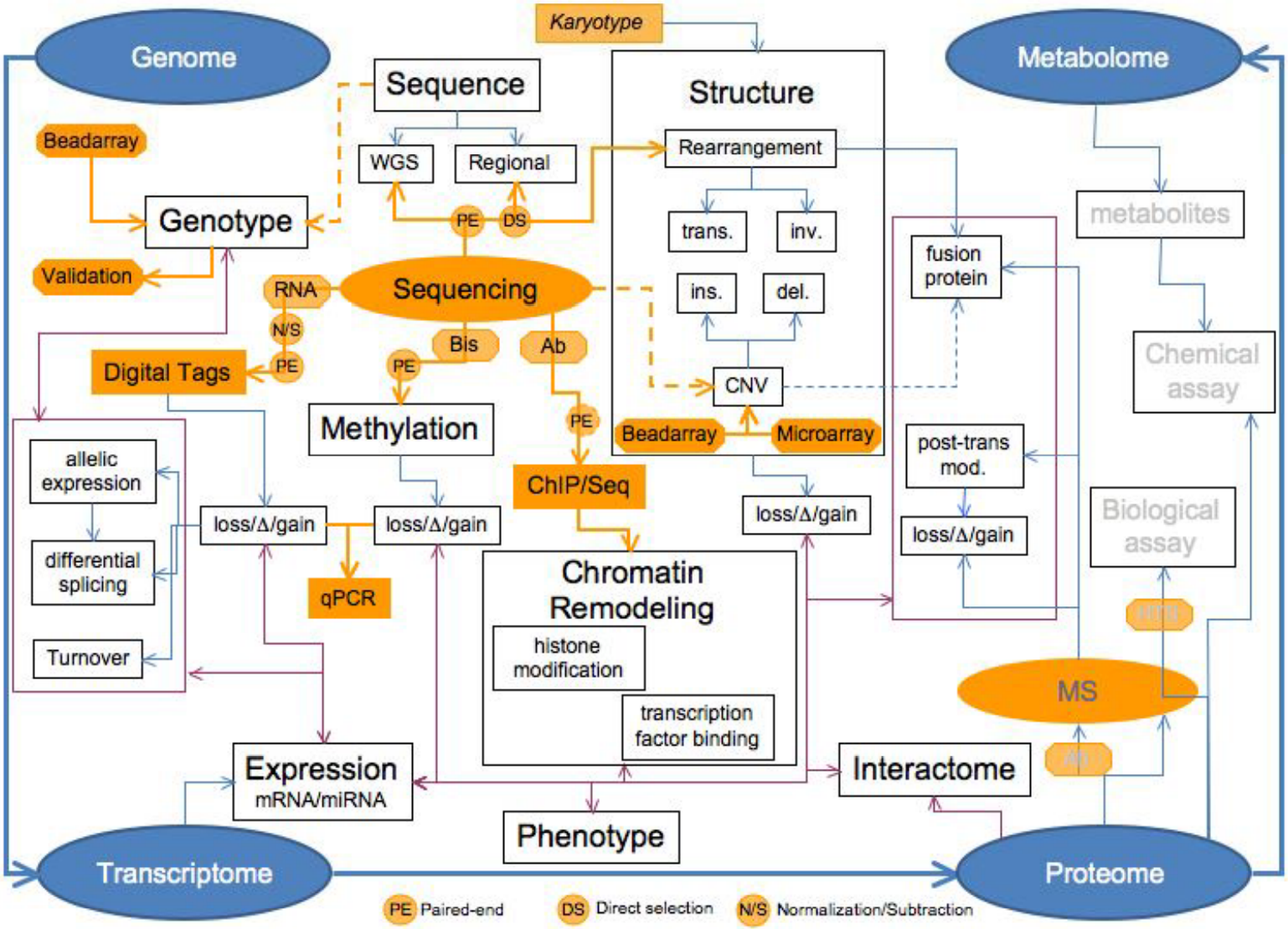
http://marketing.appliedbiosystems.com/images/Product/Solid_Knowledge/flash/102207/solid.html

Helicos BioSciences Corporation http://www.helicosbio.com/Technology/TrueSingleMoleculeSequencing/tSMStradeHowItWorks/tabid/162/Default.aspx

VISIGEN BIOTECHNOLOGIES, INC http://visigenbio.com/technology_movie_streaming.html

PACIFIC BIOSCIENCES™ *http://www.pacificbiosciences.com/video_lg.html*

John McPherson, OICR, Director of the Cancer Genomics platform



"The scaffold has been handed down to us from our ancestors, and through it we are connected to all other life on earth."

- Svante Pääbo

Revolution Through Competition.

TAKE ACTION

ARCHON X PRIZE FOR GENOMICS

A $10 MILLION PRIZE
FOR THE FIRST TEAM TO SUCCESSFULLY SEQUENCE
100 HUMAN GENOMES IN 10 DAYS

ARCHON X PRIZE FOR GENOMICS

Scientists know that a map of our genome holds boundless potential, ranging from identifying our susceptibility to disease to discovering cures for cancer. But since 1953, when James Watson and Francis Crick concluded that DNA contained the "stuff of life," only a handful of human genomes have been mapped. In fact, it still takes many months and millions of dollars to sequence a single genome.

Understanding our genomes may help delay or even prevent disease. For those suffering from genetic illnesses, personal genetic information can determine which medicines will drive their disease into remission without negative side-effects.

The Archon X PRIZE for Genomics challenges scientists and engineers to create better, cheaper and faster ways to sequence genomes. The knowledge gained by compiling and comparing a library of human genomes will create a new era of preventive and personalized medicine — and transform medical care from reactive to proactive.

The X PRIZE Foundation and scientists the world over dream of the day when we fully understand the human genetic blueprint — enabling us to make informed decisions about our own health and create a brighter future for generations to come.

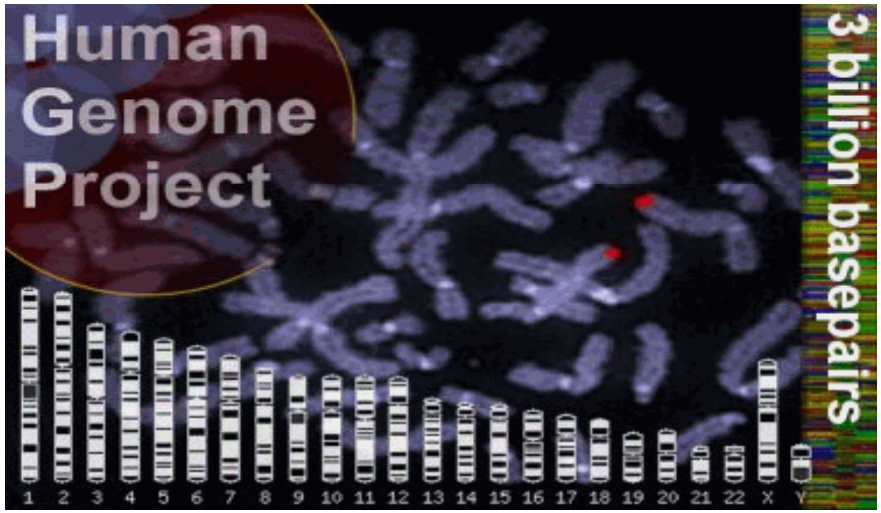
Please join the X PRIZE Foundation in our challenge to create a breakthrough in genomics that will benefit all of humanity.

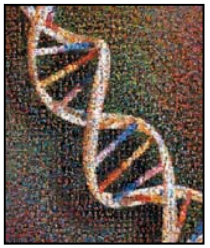
Join the Revolution

"Many thought that the unbelievable feat in 2001 of deciphering the human genome was the end of the story. Not so! We now firmly believe that it was only the beginning. The true challenge now appears to be in deciphering individual variations within the human genome. It is there that the true importance of genomics resides. Only by unearthing the blueprint details of each and every individual, will we be able to fully fathom health and disease, mental and bodily faculties, and the true secrets of human biology. The Archon X PRIZE for Genomics will greatly enhance scientists' ambition to develop methodologies for individual genomic "blue-printing". I believe the PRIZE will bring about bona-fide successes within 5 years, much earlier than otherwise feasible. It will thereby create the real genomic revolution."

Prof. Doron Lancet
Head, Crown Human Genome Center
Weizmann Institute of Science
Archon X PRIZE for Genomics
Scientific Advisory Board

How much does it cost for a Human Genome?

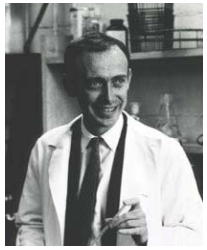




Human Reference Genome

April, 2003

>10 years to finish
USD 3 billion



James Watson's Personal Genome

June, 2007

1 year
USD 2 million

Craig Venter's Personal Genome

September, 2007

1 year
USD 1 million

YH Genome

November, 2008

1 year
~USD 0.5 million

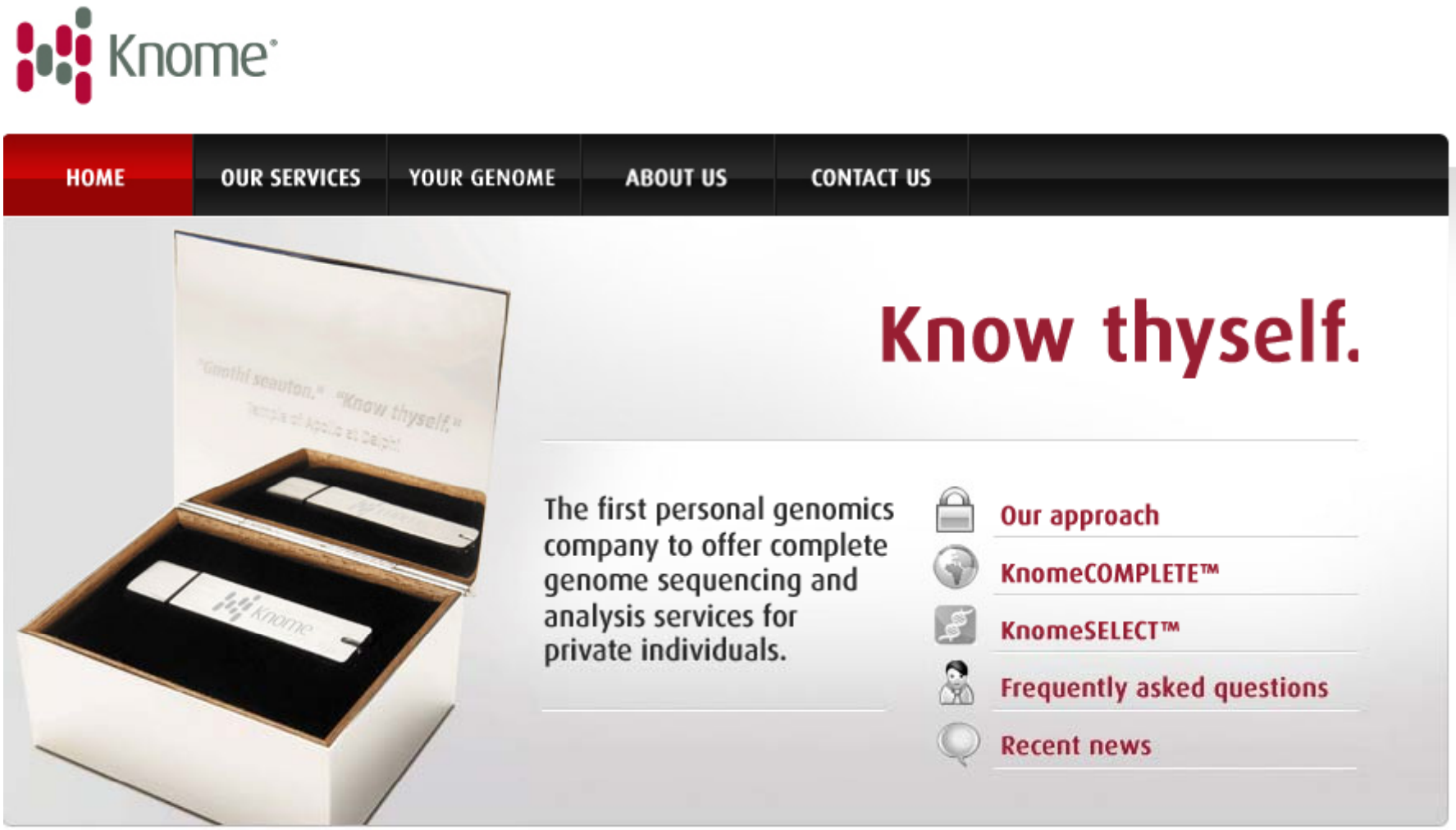
Personal Genome

When will it available?
How long will it take?
How much?

Sequenced Human Genomes

Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%)*	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2‡	250§	7.4	Aligned*	95‖	3.32 (BLAT)	234	1,000,000¶
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3,410‡	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000¶
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921‡	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000¶
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393‡	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000¶
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647‡	35, 74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000¶,#
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211‡	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000¶,**
		91% MP: 600–3,500 bp	2,075‡	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725‡	32§	28	Aligned*	90	2.81 (IndexDP)	4	48,000¶
AML female	Illumina/ Solexa	Frag: 150–200 bp‡‡	2,730‡,‡‡	32	32.7	Aligned*	91	3.81‡‡ (MAQ)	98	1,600,000‖
		Frag: 150–200 bp§§	1,081‡,§§	35	13.9		83	2.92§§ (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp‡‡	1,620‡,‡‡	35	23.3	Aligned*	98.5	3.46‡‡ (MAQ)	16.5	500,000‖‖
		MP: 200–250 bp§§	1,351‡,§§	50	21.3		97.4	3.45§§ (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238‡	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000¶,¶¶
		84% MP: 600–3,500 bp	1,211‡	25, 50						

USD 10000 per Genome in 2015?



US$ XX,000 for a personal genome

Complete genomics

Revolutionizing human genome discovery

Human Genome Sequencing & Analysis Service

Dedicated to **complete human genome sequencing and analysis** provided as an innovative, **end-to-end, outsourced service model**, Complete Genomics enables researchers to conduct **large-scale complete human genome studies**.

By **optimizing our sequencing platform for human DNA**, we are able to achieve **accuracy levels of** 99.999% at a total cost that is significantly less than the total cost of purchasing and using commercially available DNA sequencing instruments.

» read more

Receive Research-Ready Genomic Data

We offer our customers an end-to-end, outsourced solution that delivers **research-ready genomic data**.

Our CGA™ Service provides reliable access to assembled and **annotated sequence data** and our analytical tools enable our customers to **rapidly analyze and compare genomic data**.

» read more

WEBINAR

Pedigree Genome Sequencing

presented by:
Jared Roach, Ph.D., M.D.,
Institute for Systems Biology

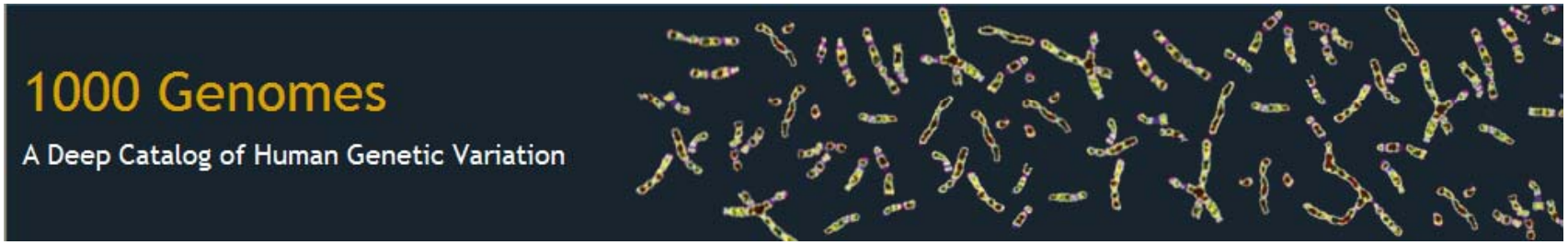
View

News and Events

» **Complete Genomics Announces Pricing of Initial Public Offering**

» **Complete Genomics Expands CGA Tools Software Suite**

» **Complete Genomics to Sequence**

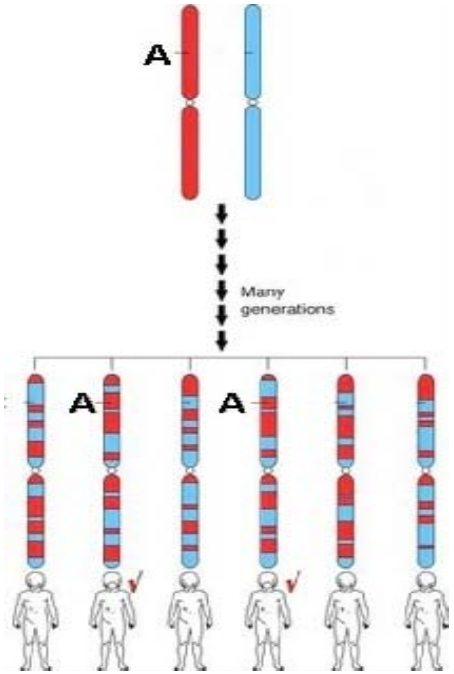


The 1000 Genomes Project is an international research consortium formed to create the most detailed and medically useful picture to date of human genetic variation. The project involves sequencing the genomes of approximately 1200 people from around the world and receives major support from the Wellcome Trust Sanger Institute in Hinxton, England, the Beijing Genomics Institute Shenzhen in China and the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health (NIH).

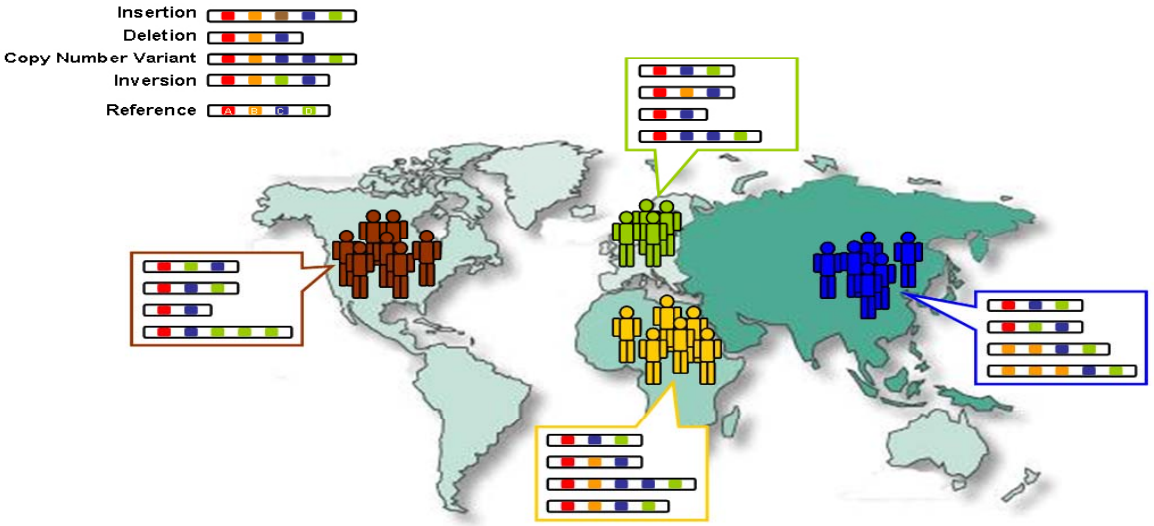
Drawing on the expertise of multidisciplinary research teams, the 1000 Genomes Project will develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.

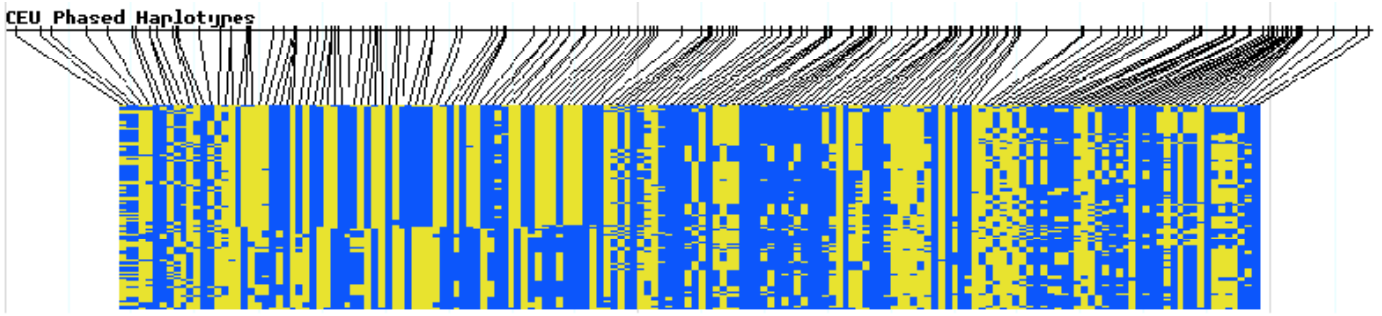
On 4 September 2008, the co-chairs of the analysis group and overall project co-chairs drafted a letter to the NIH Council about 1000 Genomes Project. This letter, available here, reviews the goals, describes the current status, and provide an update on the critical tasks the Analysis Group must accomplish in order to deliver a valuable community resource and achieve the Project's goals.

The Scientific Goals of the 1000 Genomes Project



To produce a catalog of variants that are present at 1 percent or greater frequency in the human population across most of the genome, and down to 0.5 percent or lower within genes.





A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium

Affiliations | Contributions | Corresponding author

Nature **467**, 1061–1073 (28 October 2010) | doi:10.1038/nature09534
Received 20 July 2010 | Accepted 30 September 2010 | Published online 27 October 2010

Abstract

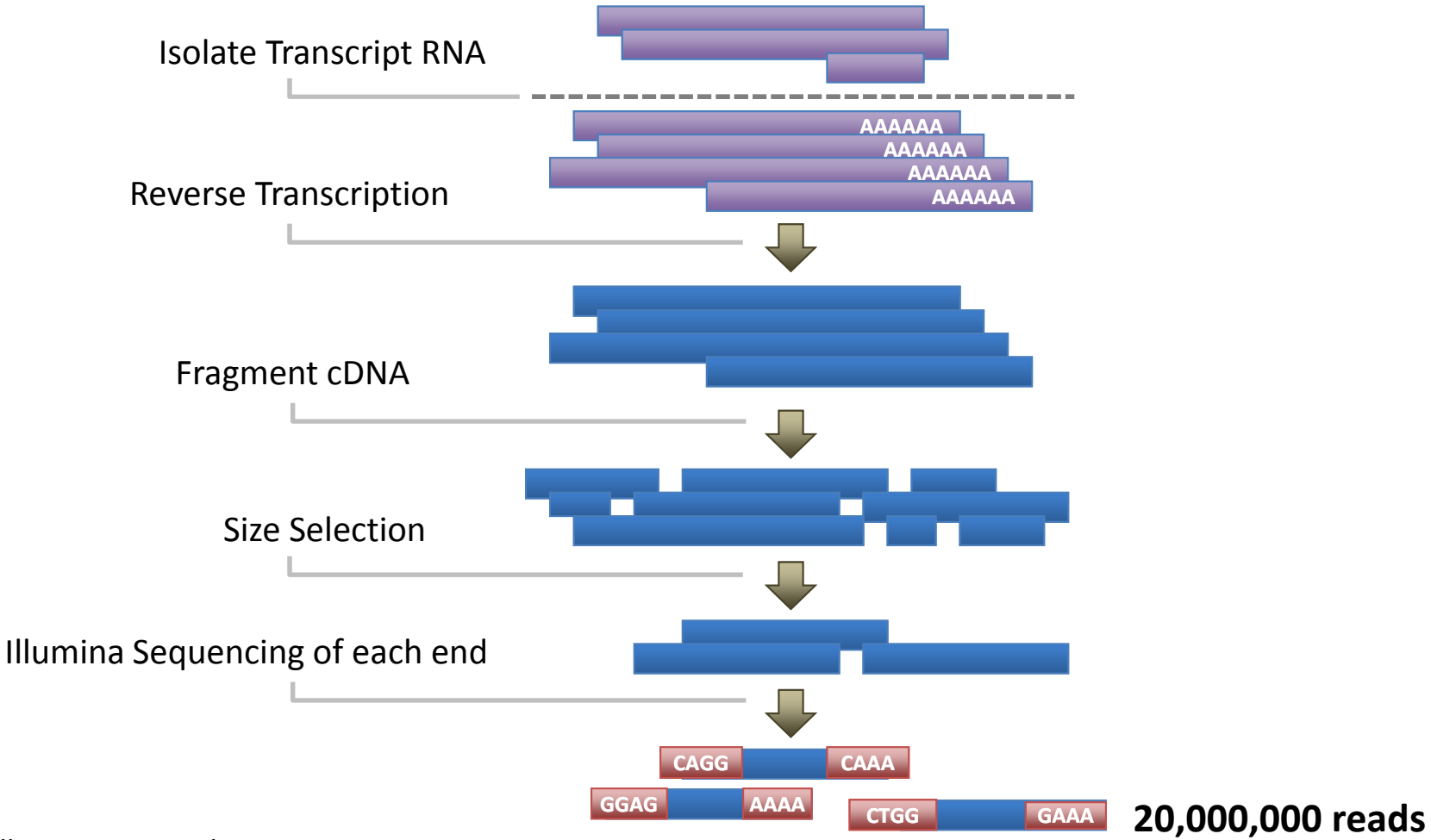
The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. We undertook three projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother–father–child trios; and exon-targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of approximately 15 million single nucleotide polymorphisms, 1 million short insertions and deletions, and 20,000 structural variants, most of which were previously undescribed. We show that, because we have catalogued the vast majority of common variation, over 95% of the currently accessible variants found in any individual are present in this data set. On average, each person is found to carry approximately 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrate how these results can be used to inform association and functional studies. From the two trios, we directly estimate the rate of *de novo* germline base substitution mutations to be approximately 10^{-8} per base pair per generation. We explore the data with regard to signatures of natural selection, and identify a marked reduction of genetic variation in the neighbourhood of genes, due to selection at linked sites. These methods and public data will support the next phase of human genetic research.

1000 Genomes Samples							05-Mar-10
Population	Status	Cell lines avail (all data approx)	Adult/child samples from trios avail	First set	Second set	Third set	Total
Utah residents (CEPH) with Northern and Western European ancestry (CEU)	Available - HapMap samples	At Coriell		100			100
Tuscan in Italia (TSI)	Available - HapMap samples	At Coriell		100			100
British from England and Scotland (GBR)	Available	At Coriell		100			100
Finnish from Finland (FIN)	Available	At Coriell		100			100
Iberian populations in Spain (IBS)	Collecting samples	Jun 2010	yes		100		100
TOTAL European ancestry				400	100		500
Han Chinese in Beijing, China (CHB)	Available - HapMap samples	At Coriell		100			100
Japanese in Tokyo, Japan (JPT)	Available - HapMap samples	At Coriell		100			100
Han Chinese South (CHS)	Available	At Coriell	yes	100			100
Chinese Dai in Xishuangbanna (CDX)	Awaiting govt approval	late 2010	yes		100		100
Kinh in Ho Chi Minh City, Vietnam (KHV)	Awaiting govt approval	late 2010	yes		100		100
Chinese in Denver, Colorado (CHD) (pilot 3only)	Available - HapMap samples	At Coriell		0	0		0
TOTAL East Asian ancestry				300	200		500
Yoruba in Ibadan, Nigeria (YRI)	Available - HapMap samples	At Coriell	yes	100			100
Luhya in Webuye, Kenya (LWK)	Available - HapMap samples	At Coriell		100			100
Gambian in Western Division, The Gambia (GWD)	Collecting samples	late 2010	yes		100		100
Ghanaian in Navrongo, Ghana (GHN)	Final IRB approval expected soon	late 2010	yes		100		100
Malawian in Blantyre, Malawi (MAB)	Discussing issues for participation	?	yes		100		100
TOTAL West African ancestry				200	300		500
African Ancestry in Southwest US (ASW)	Available - HapMap samples	At Coriell	yes	61			61
African American in Jackson, MS (AJM)	IRB approval received, recruit soon	late 2010	yes		80		80
African Caribbean in Barbados (ACB)	Collecting samples	late 2010	yes		79		79
Mexican Ancestry in Los Angeles, CA (MXL)	Available - HapMap samples	At Coriell	yes	70			70
Puerto Rican in Puerto Rico (PUR)	Available	Jun 2010	yes	70			70
Colombian in Medellin, Colombia (CLM)	Collecting samples	Jun 2010	yes		70		70
Peruvian in Lima, Peru (PEL)	Collecting samples	late 2010	yes		70		70
TOTAL Americas				201	299		500
Ahom in the State of Assam, India	Awaiting govt & IRB approval	2011	yes			100	100
Kayadtha in Calcutta, India	Awaiting govt & IRB approval	2011	yes			100	100
Reddy in Hyderabad, India	Awaiting govt & IRB approval	2011	yes			100	100
Maratha in Bombay, India	Awaiting govt & IRB approval	2011	yes			100	100
Punjabi in Lahore, Pakistan	Awaiting IRB approval	late 2010	yes			100	100
TOTAL South Asian ancestry						500	500
TOTAL				1101	899	500	2500

Applications on Biomedical Sciences

Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

Transcriptome Sequencing

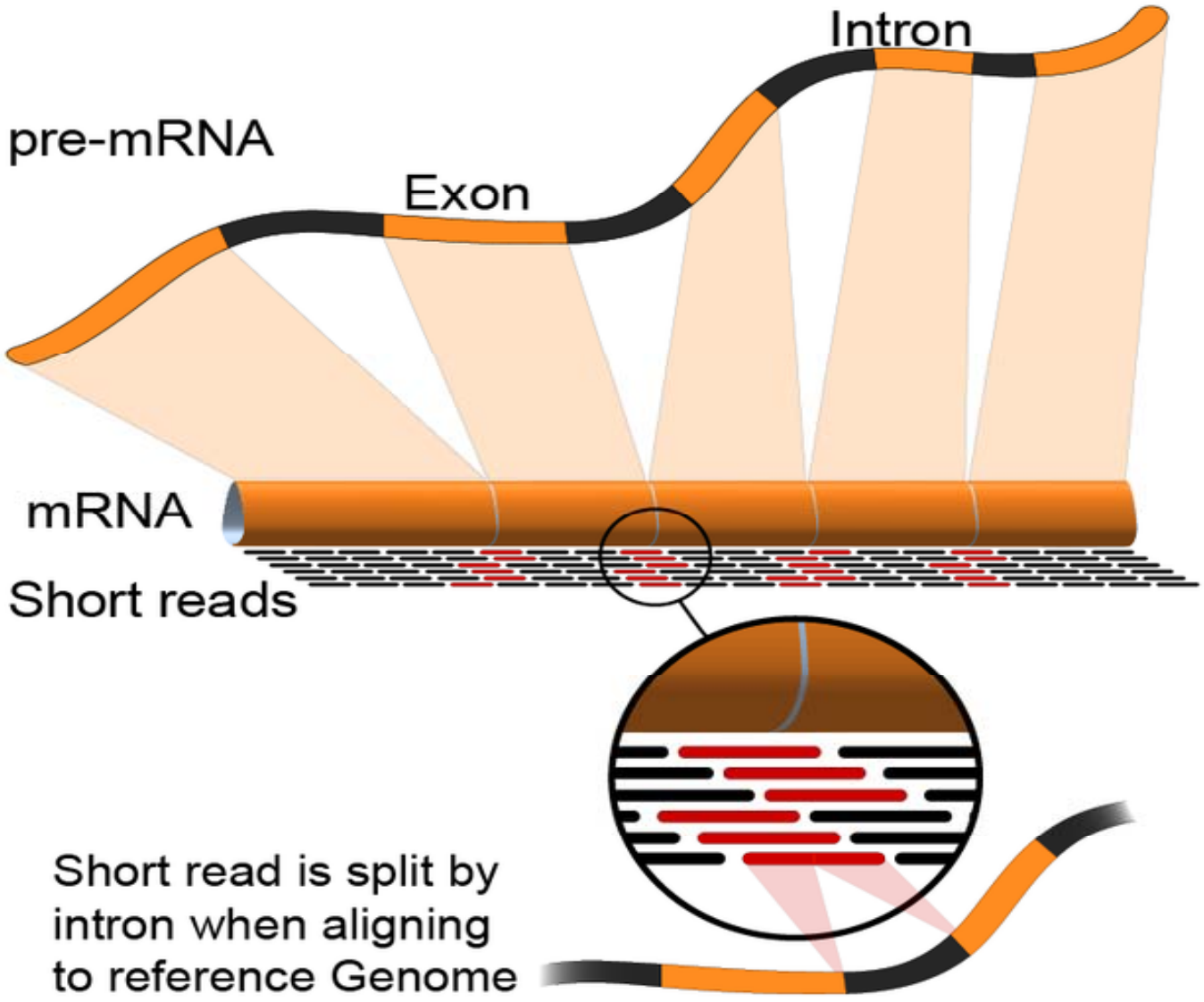


*based on Illumina approach

**strand-specific RNA-seq protocols exist for both Illumina and SOLiD

Slide complements of Andrew McPherson

Transcriptome Sequencing



Analysis Strategies: Reference Sequence Alignment *vs* De novo Assembly

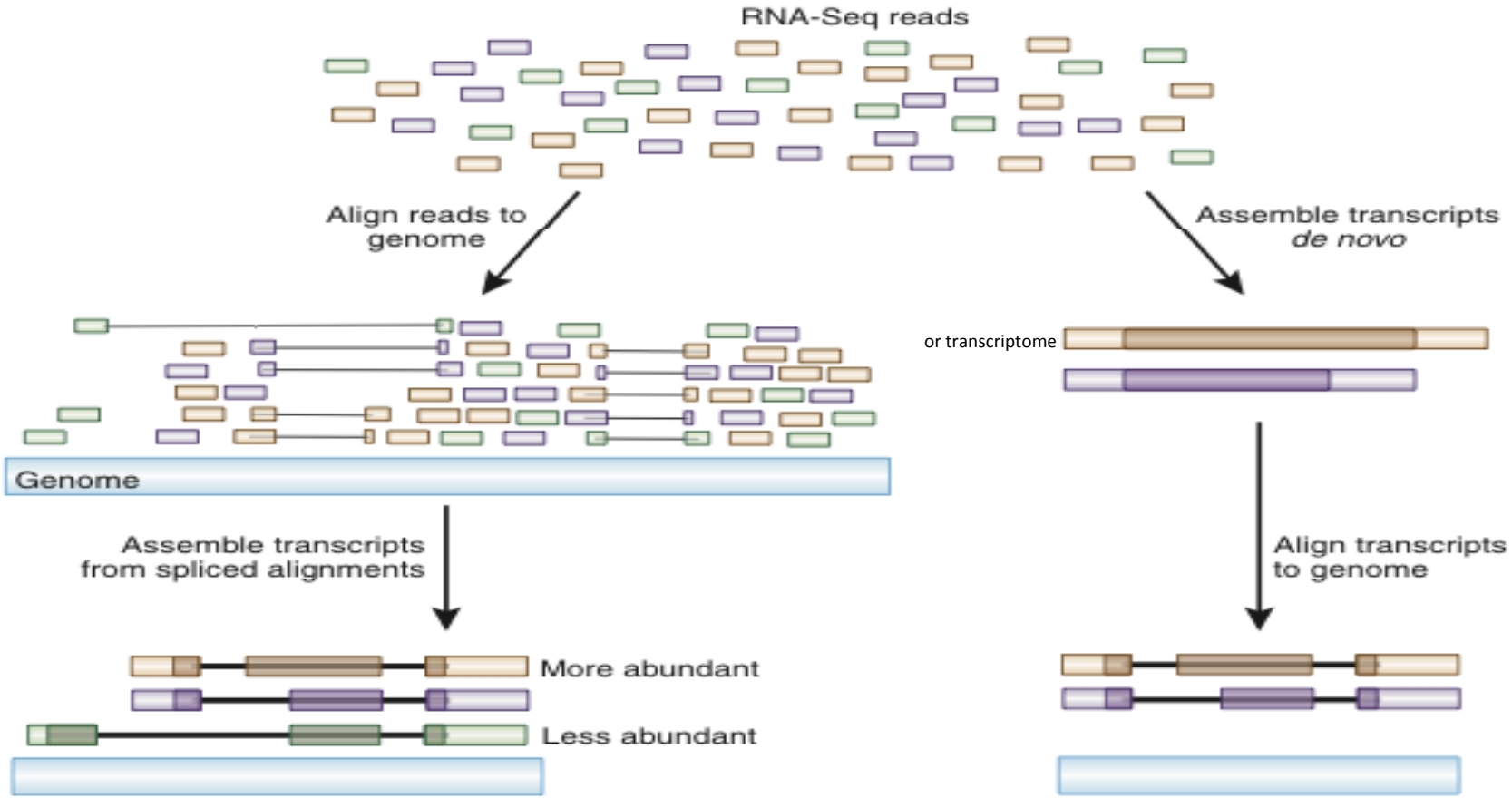
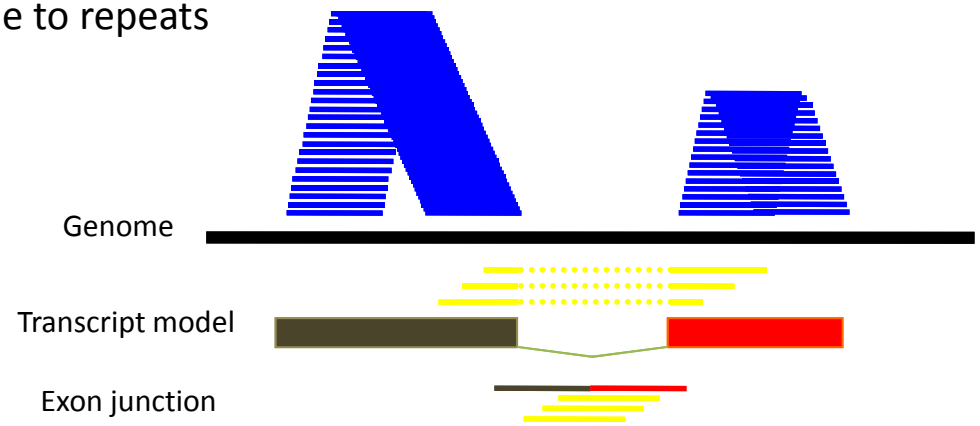


Image from Haas & Zody, 2010

*Assembly is the only option when working with a creature with no genome sequence, alignment of contigs may be to ESTs, cDNAs etc

Drawbacks for each strategy

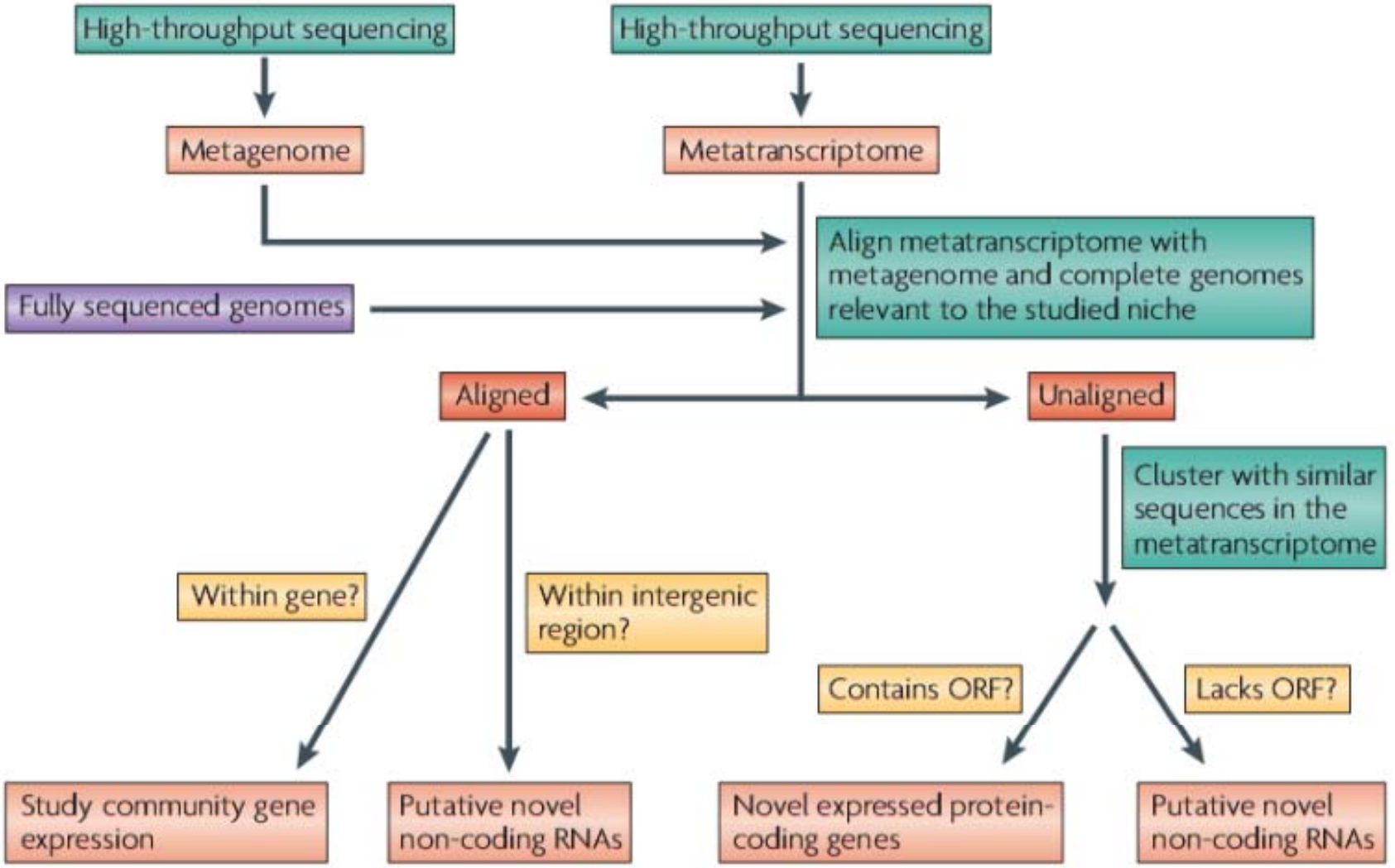
- **Alignment to genome**
  - Computationally expensive
  - It is **never** a good idea to simply align RNA-seq data to the genome Need a spliced aligner or a surrogate (such as including exon-exon junction sequences in ‘genome’)
- **Alignment to transcriptome**
  - Reads deriving from non-genic structures may be ‘forcibly’ (and erroneously) aligned to genes
    - Incorrect gene expression values
    - False positive SNVs
    - Many other potential problems
- **Assembly**
  - Low expression = difficult/impossible to assemble
  - Misassemblies/fragmented contigs due to repeats
  - Requires vast amounts of memory



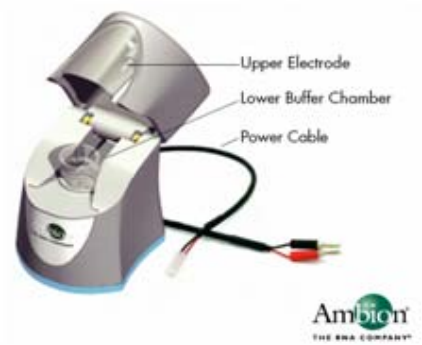
Benefits of each approach

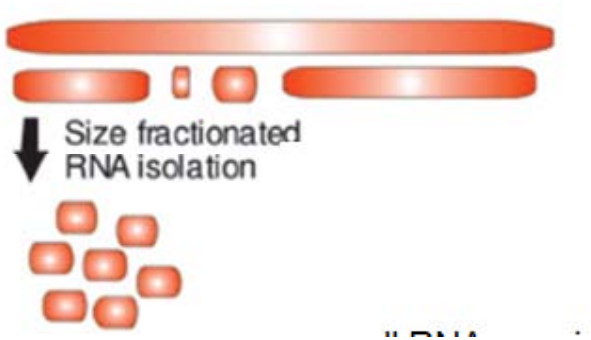
- **Alignment to genome**
  - Allow reads from unannotated loci, introns *et cetera* to align to their correct locations… potential for new biological insights
- **Alignment to transcriptome**
  - Computationally inexpensive
  - Spliced (exon junction) reads map correctly
  - Pairing distance and junction reads may help distinguish individual isoforms (informative/unique regions of transcripts)
- **Assembly**
  - Can provide a more long-range view of transcripts
  - Allows detection of chimeric transcripts and resolution of ‘breakpoints’
  - May not necessarily need a genome

What can you get from Transcriptome Sequencing

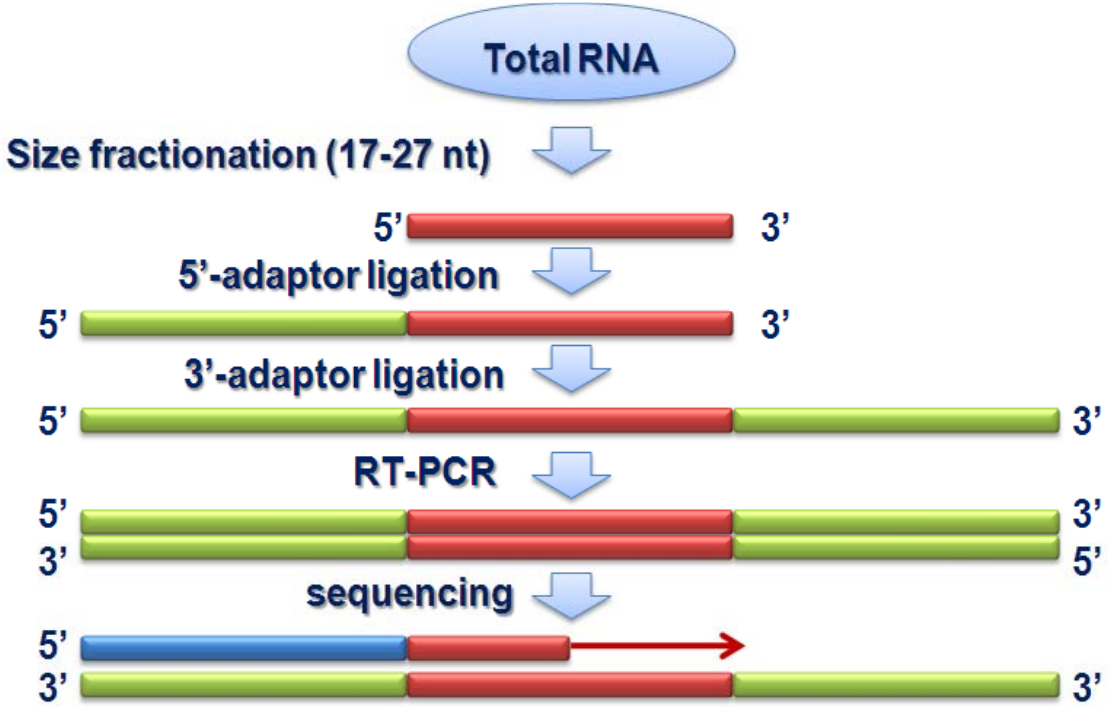


Small RNA Sequencing





Illumina SOLEXA sequencing by synthesis




```
@HWI-EAS82_3_FC204V1AAXX:6:1:886:345
AGAGTTCTACAGTCCGGACGATCTCGTATGCCGTC
+HWI-EAS82_3_FC204V1AAXX:6:1:886:345
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
```

40,000,000 reads of 35 bp long

Small RNA Sequencing

Workflow

Clean up

↓

Clustering

↓

ncRNA matching

(Rfam V10)

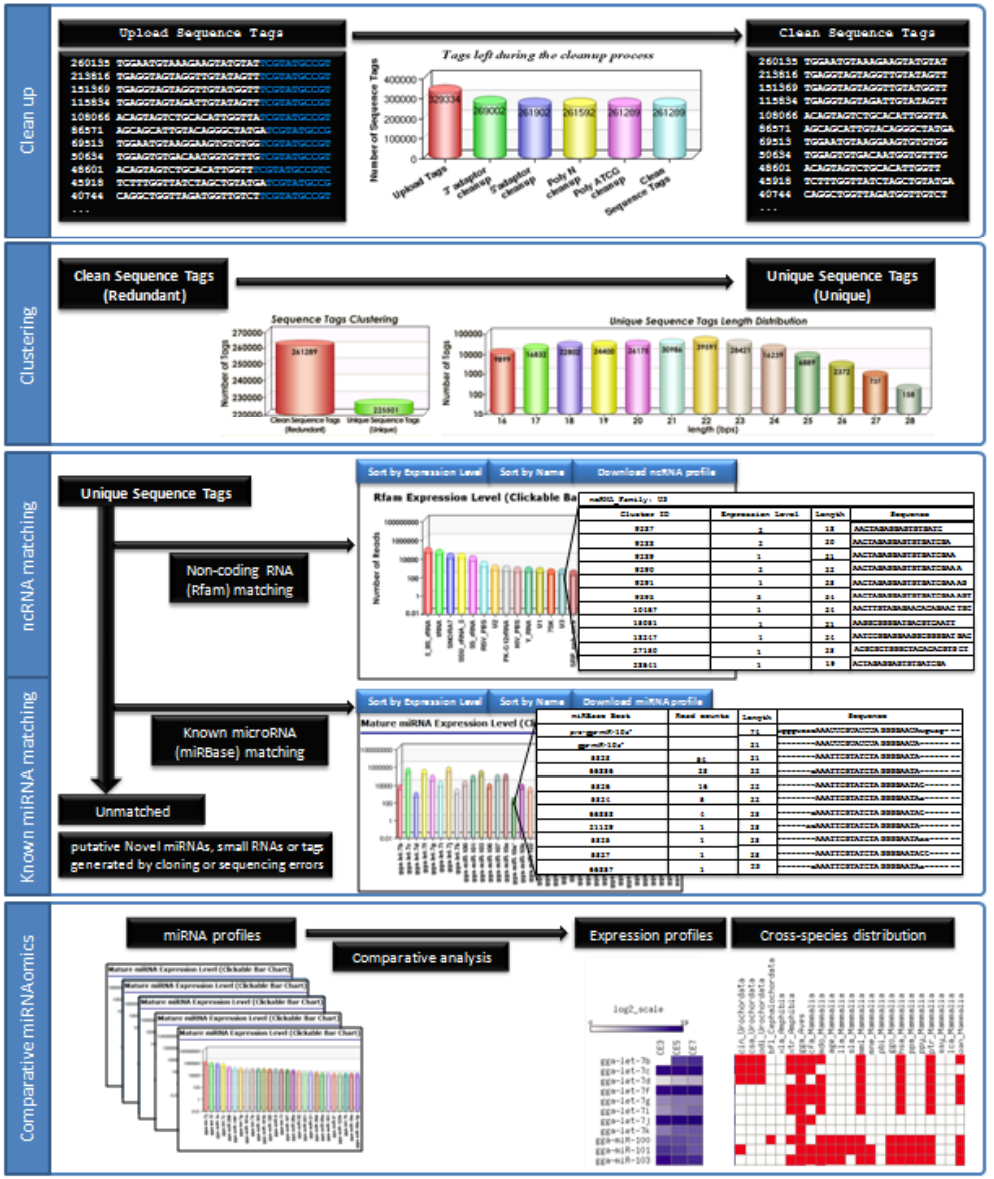
↓

Known miRNA Matching

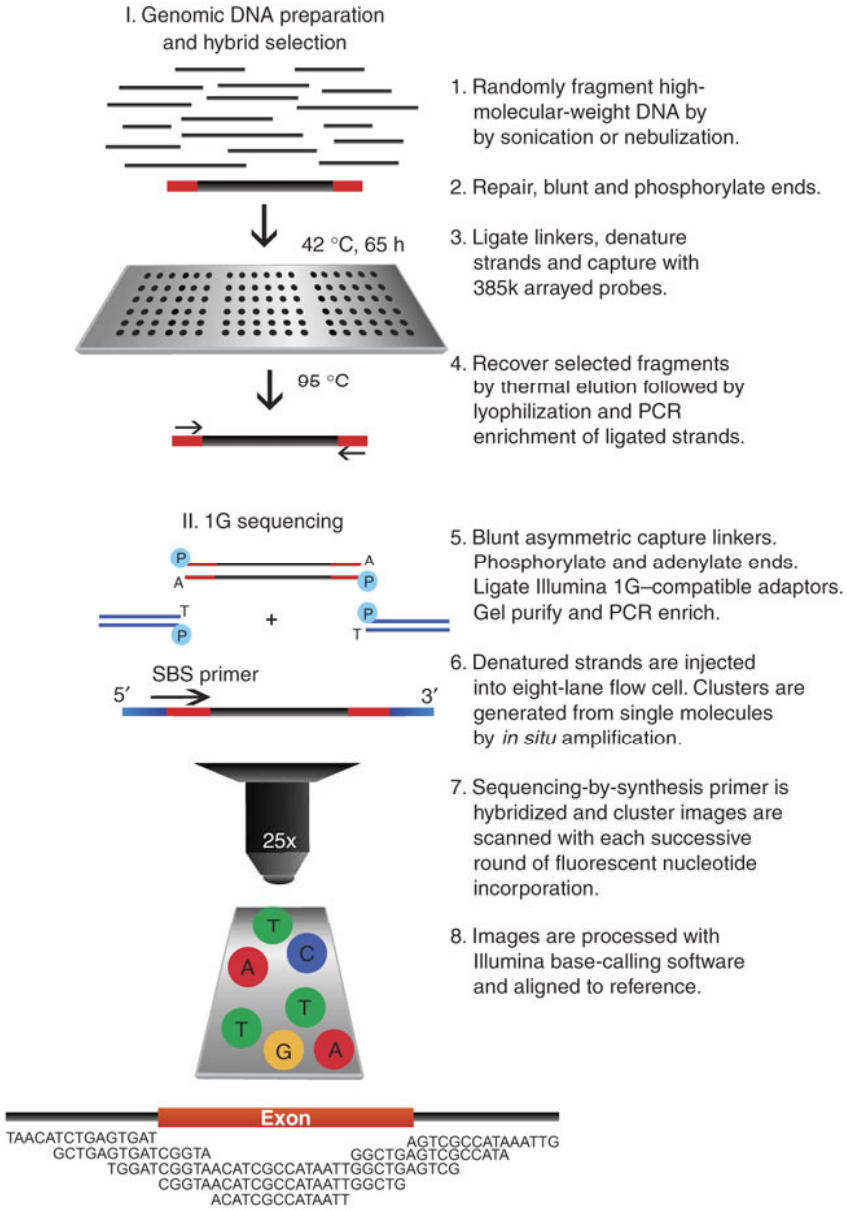
(miRBase V16)

↓

Comparative miRNAomics



Human Exome Sequencing

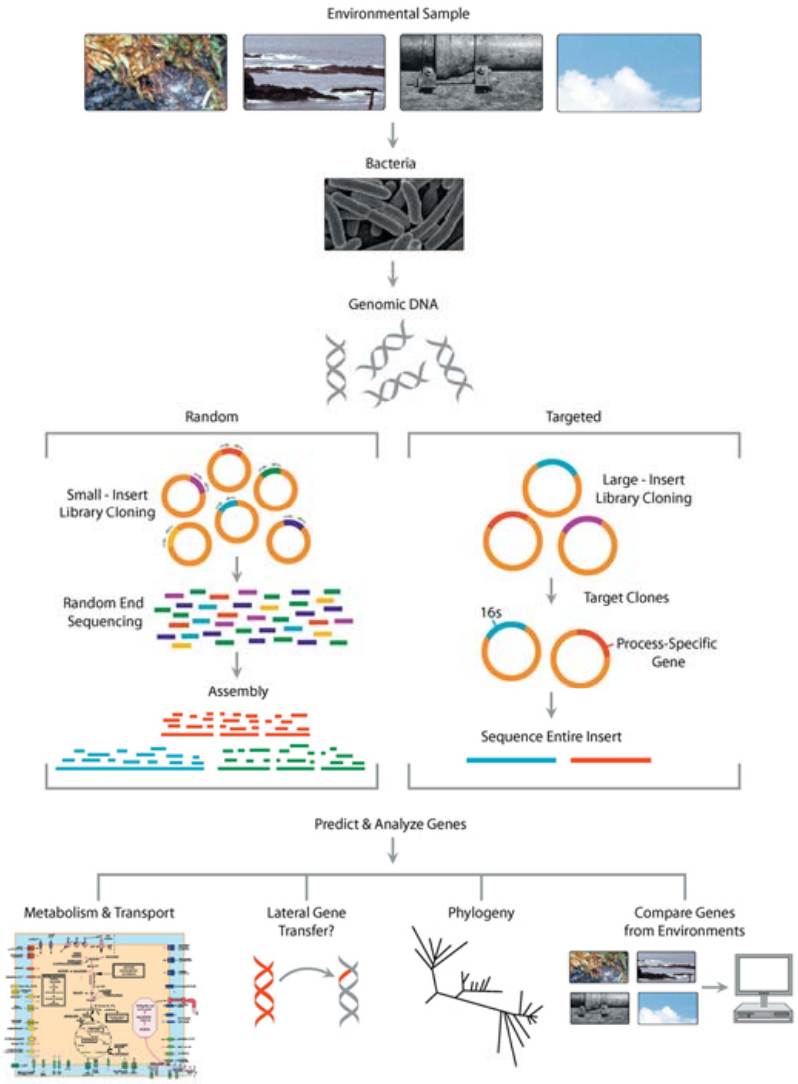


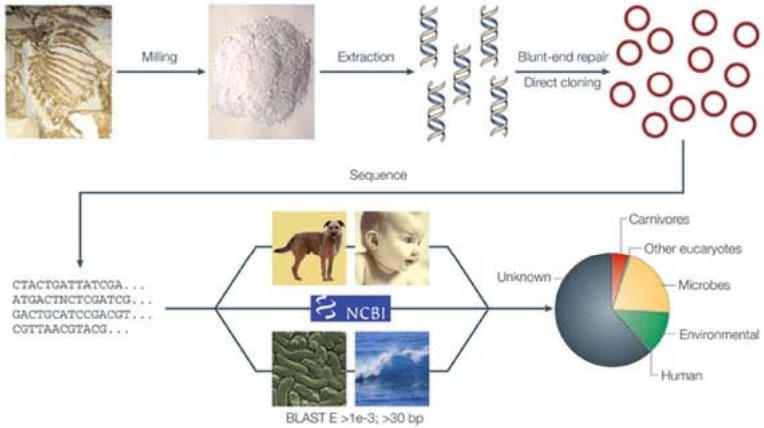
A single Nimblgen 2.1M array is capable of enrichment simultaneously more than 180,000 human coding-exon and 550 miRNA exons available in the CCDS database. The methods significantly improve the efficiency, the cost and throughput of target enrichment compared to conventional PCR-based methods.

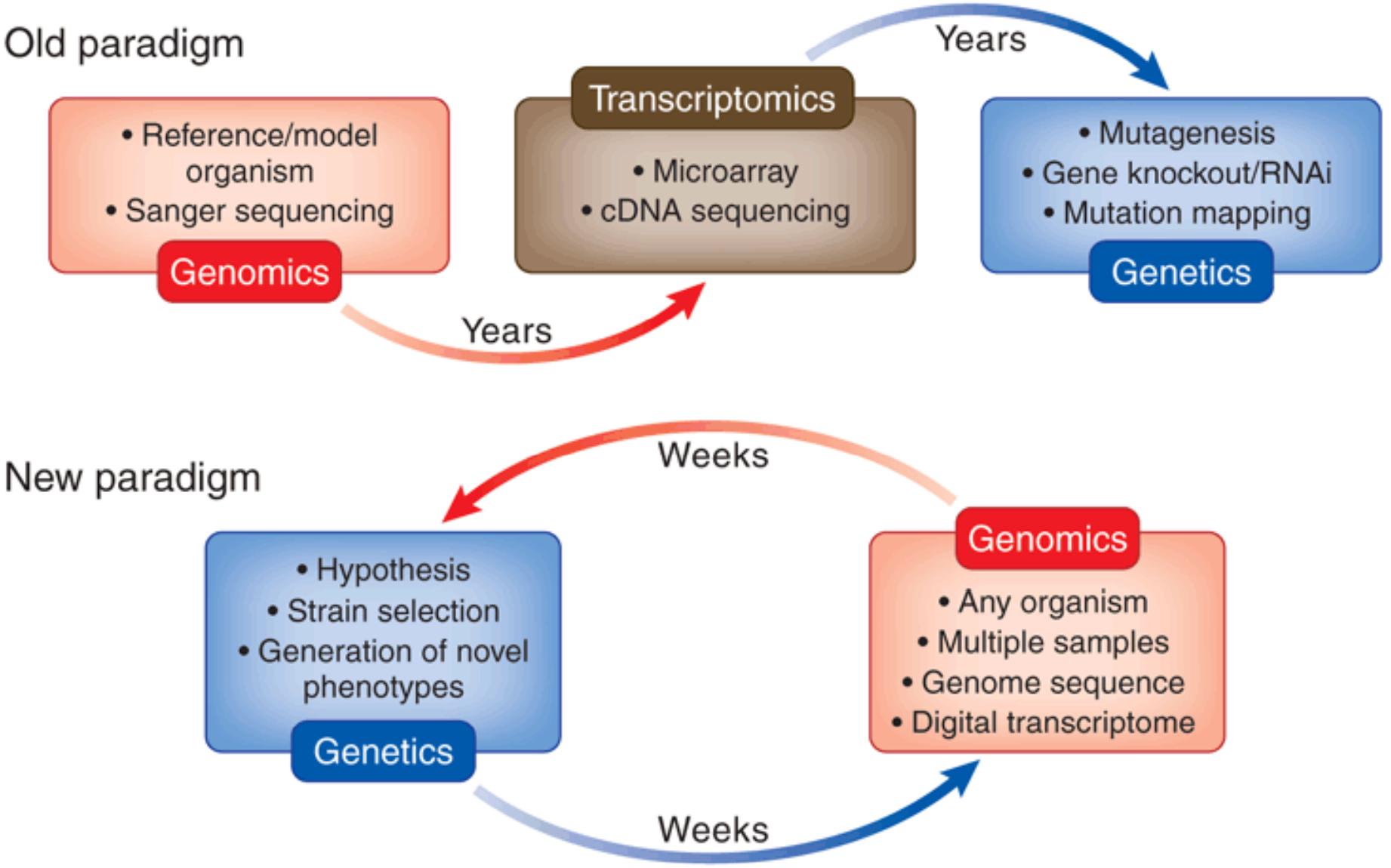
The human exome captured library preparation consists of three major processes and lasts seven days, including library preparation for sequence capture, microarray hybridization, elution and library construction for sequencing, using Illumina Genome Analyzer can directly get high quality sequence data for the entire human exome.

This approach can find the exact genes and mutations causing several complex human diseases, such as cancer, diabetes, obesity and so on.

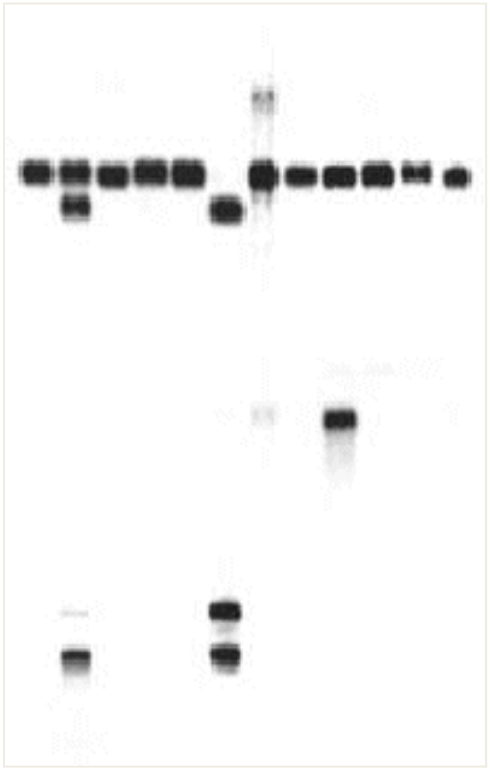
Metagenomics






Old paradigm
• Reference/model organism
• Sanger sequencing
Genomics
Years
Transcriptomics
• Microarray
• cDNA sequencing
Years
• Mutagenesis
• Gene knockout/RNAi
• Mutation mapping
Genetics
New paradigm
Weeks
• Hypothesis
• Strain selection
• Generation of novel phenotypes
Genetics
Genomics
• Any organism
• Multiple samples
• Genome sequence
• Digital transcriptome
Weeks

The “old” biology



The most challenging task for a scientist is to get good data

The “new” biology

The most challenging task for a scientist is to make sense of lots of data

Computing Power

Next gen sequencers generated Giga bp to Tera bp of data

HiSeq2000 (launched in Q1 2010)

Read Length	Run Time	Output
1 x 35 bp	~1.5 days	26-35 Gb
2 x 50 bp	~4 days	75-100 Gb
2 x 100 bp	~8 days	150-200 Gp

Up to 25GB per day for a 2 x100 bp run

CPU

Disk space

RAM

Bandwidth

CPU

- The speed of your CPU determines how quickly it can process instructions
- Many bioinformatics operations fall into the “embarrassingly parallel” category
- Getting a results faster is as simple as adding more CPUs
- => clusters
- Though clusters still dominate, increasing attention is being paid to servers with large number of cores

For current throughputs, you will need ~8 <dual quad core> nodes per sequencer to handle raw data.

For data analysis, 1-2 nodes is enough for general purpose.

How about use Graphic Processing Unit (GPU) instead?

RAM

- Typical sizing is 2GB of RAM per core
- This works fine for most aligners
  - Certain aligners require a minimum number of cores for optimal efficiency
  - 16-48GB RAM per node
- Assemblers typically need much more RAM
- If you don't have enough RAM, the CPU will need to make use of the disk storage –
- When a computer has run out of RAM it is said to be "swapping"
- Some aligners have a minimum memory limit for optimal performance
- Human de novo requires lots of RAM!

Disk Space

- Sequencers generate a lot of data
- Including quality values and additional files (alignments etc) create multiplier
- Binary formats such as BAM are helping, but there are limited binary formats for basic data
  - Life and Illumina are both moving to binary formats for basic data
- **Debates on what needs to be stored**
- RAID
- Scaling

Bandwidth

Bandwidth of a connection represents the
maximum rate of transfer between two points

- E.g. An aligner can process X reads per second on a single CPU at a data rate of Y bytes/sec
  - ~200 million reads in 10 hours
  - Each read 50 bases at 10 bytes per base
  - 2.7 MB/sec
- Design 100TB storage and connect it to a CPU resource
- Design bandwidth to be 10 Mb/sec – plenty of spare bandwidth
- Now we want to complete the job in an hour and get permission to buy 10 more CPUs – great!

BUT
For the 10 CPUs to run at maximum speed, they need to be supplied data at 27MB/sec
Our bandwidth is 10MB/sec
Therefore, no matter how many CPUs we buy, the job will never run faster than ~ 2.5 hours .

How about optic fibers?

http://petang.cgu.edu.tw

Molecular Regulation and Bioinformatics Laboratory

The Molecular Regulation & Bioinformatics Laboratory (MRBLab) use bioinformatics approaches to integrate data generated by high-throughput technologies to compare the gene, protein and miRNA expression levels of protozoan as a basis for the development of new chemotherapeutic agents, to elucidate the interactome of pathogen-host and to study the biology of longevity in protozoan.

Molecular Regulation & Bioinformatics Laboratory
Dept. Parasitology, College of Medicine, Chang Gung University, TAIWAN

PETRUS TANG
Ph.D. (CAM, UK)

Contact Information

Address:
Dept. of Parasitology, College of Medicine,
Chang Gung University.
259 Wenhwa 1st. Road. Kweishan,
Taoyuan 333. Taiwan.
Phone number:
+886-3-2118800#5136
Fax:
+886-3-2118122
E-mail:
petang@mail.cgu.edu.tw

LATEST NEWS

[Announcement]

[Lecture]
- *Bioinformatics Databases*

[Publication]
- May2010 - *"DSAP: Deep-Sequencing Small RNA Analysis Pipeline" published in Nucleic Acids Research.*
- Mar2010 - *"The Genome of Trichomonas vaginalis" in Anaerobic Parasitic Protozoa: Genomics and Molecular Biology*
- Mar2010 - *"Proteomic analysis of the effect of cyanide on Klebsiella oxytoca" published in Current Microbiology.*
- Feb2010 - *"Trichomonas vaginalis vast BspA-like gene family: evidence for functional diversity from structural organisation and transcriptomics" published in BMC Genomics*
- Jan2010 - *"Loss of Bikunin in urine as useful marker of bladder carcinoma" published in J. Urology.*


THANK YOU