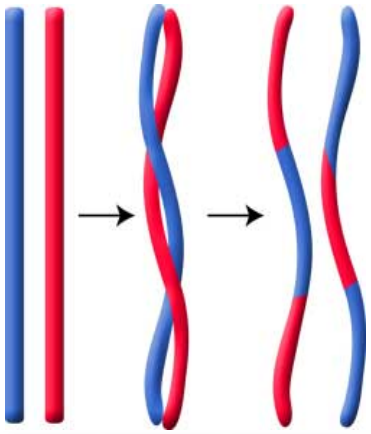


High Throughput Sequencing (NEXT GENERATION SEQUENCING)

Technologies and Applications

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

http://petang.cgu.edu.tw

Molecular Regulation and Bioinformatics laboratory

Home | MRBLab | Research | CV | Publications | Parasitology | Bioinformatics

The Molecular Regulation & Bioinformatics Laboratory (MRBLab) use bioinformatics approaches to integrate multi-omic (including genomics, proteomics, transcriptomics, metablomics, etc.) analysis techniques to study parasite-host interaction and to develop pipelines/visualization methods that makes computational biology accessible to research scientists with no programming experience

Molecular Regulation & Bioinformatics Laboratory

Dept. Parasitology, College of Medicine, Chang Gung University, TAIWAN



Ph.D. (CAM, UK)
Associate Prof.

Prof. Petrus Tang, Ph.D.

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

[Announcements]

- *Trichomonas videos*

1. *Trichomoniasis: Why is It the Neglected STD?* by M. Hobbs
2. *A Trich-y parasite: genomics, population genetics and evolution of Trichomonas vaginalis* by Jane Carlton

[Lectures]

- *Applications of Bioinformatics and Computational Techniques in Biomedical Research*

高通量定序分析

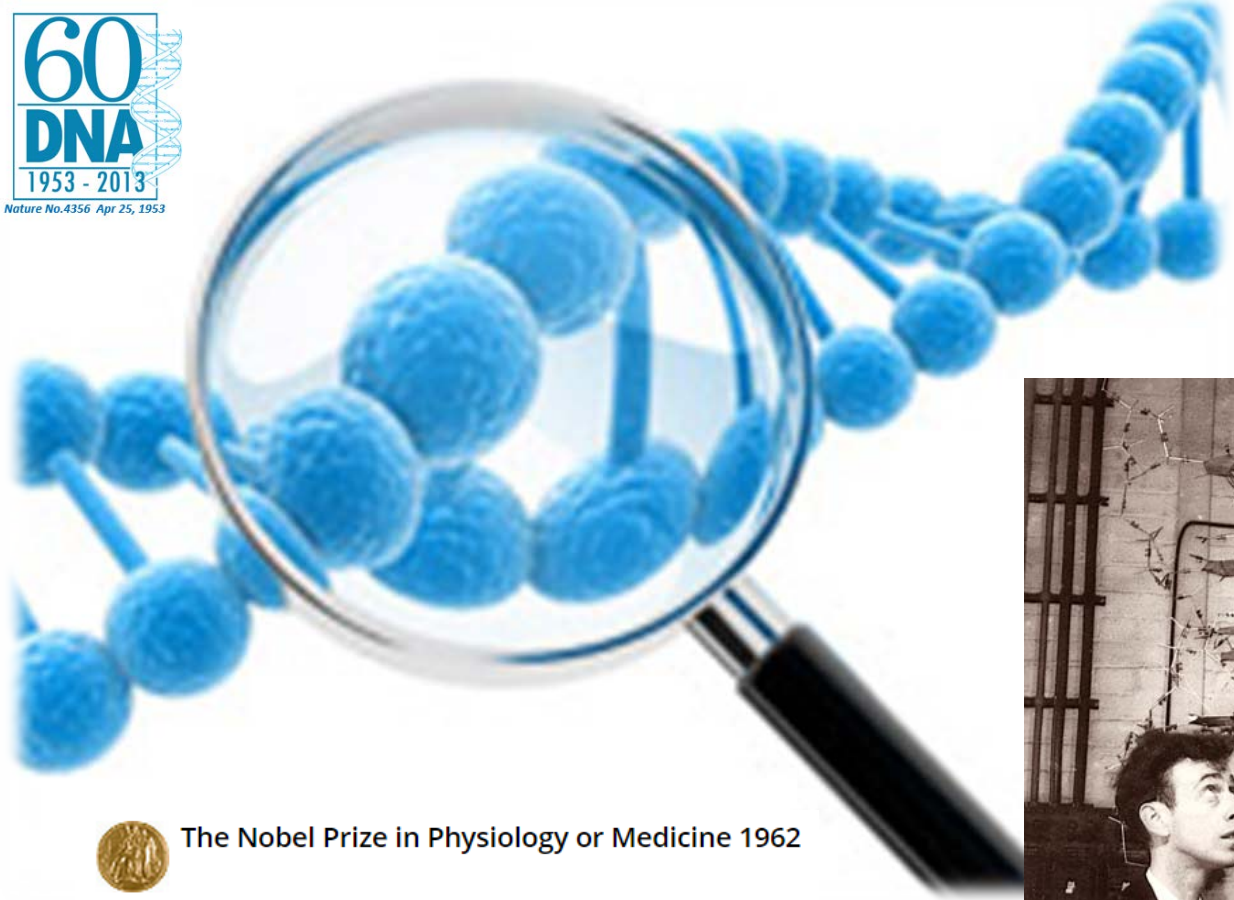
上課時間：每週五上午 08:10~12:00 （08:10~09:00－授課，09:10~12:00－實習）
上課地點：醫學院 9F 生物資訊中心 電腦教室

助教：蔡智宇(分機5690)

進度	課程進度	授課老師	講義下載 L:lecture T:tutorial E:exercise
1	New Generation Sequencing (NGS) Technologies & Biological Big Data	鄧致剛	
2	Databases and Tools for High Throughput Sequencing Analysis	葉元鳴	
3	Small RNA Analysis and Target Identification	黃柏榕	
4	Gene Expression Analysis for Non-model Organisms	鄧致剛	
5	Gene Expression Analysis	葉元鳴	
6	Transcriptome Functional Profiling - DAVID	朱俐潔	
7	Small Genome de novo Assembly & Annotation I	陳亭妏	
8	Small Genome de novo Assembly & Annotation II	陳亭妏	
9	MID-TERM EXAMINATION	鄧致剛	
10	Genome Annotation & Genome Browser	黃柏榕	
11	Chip Sequencing Analysis	陳亭妏	
12	Metagenome Analysis	葉元鳴	
13	Cancer Panel Sequencing Analysis	黃柏榕	
14	Exome Sequencing Analysis	黃柏榕	
15	Biological Big Data Visualization & Integration	李季青	
16	Galaxy	黃柏榕	
17	Student Presentation	鄧致剛	
18	Final-Examination	鄧致剛	

60 DNA 1953 - 2013

Nature No.4356 Apr 25, 1953



The Nobel Prize in Physiology or Medicine 1962



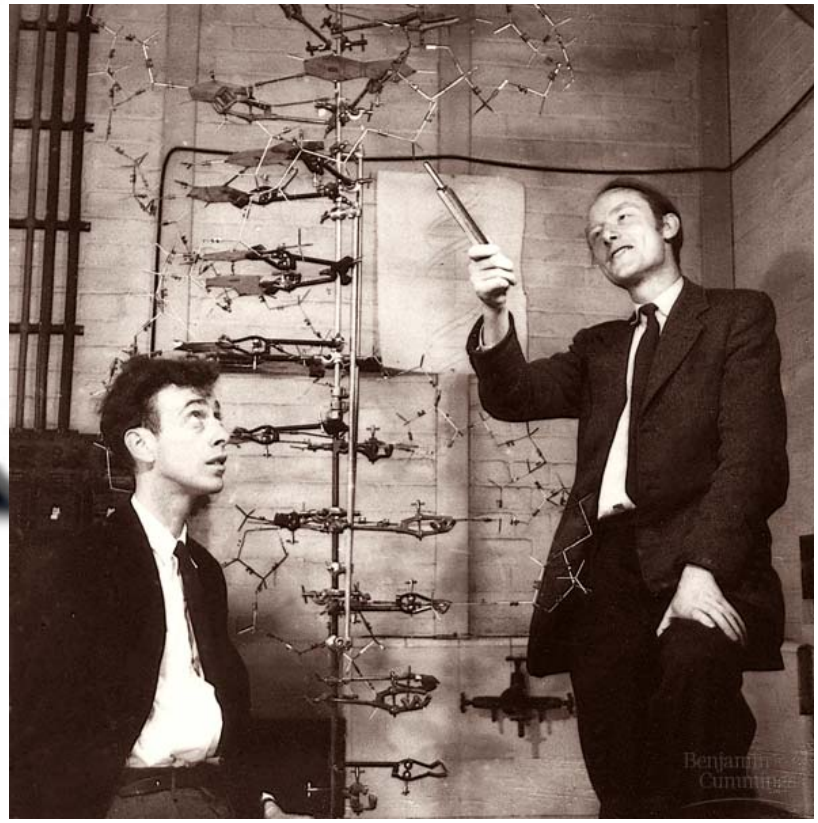
Francis Harry Compton Crick
Prize share: 1/3



James Dewey Watson
Prize share: 1/3

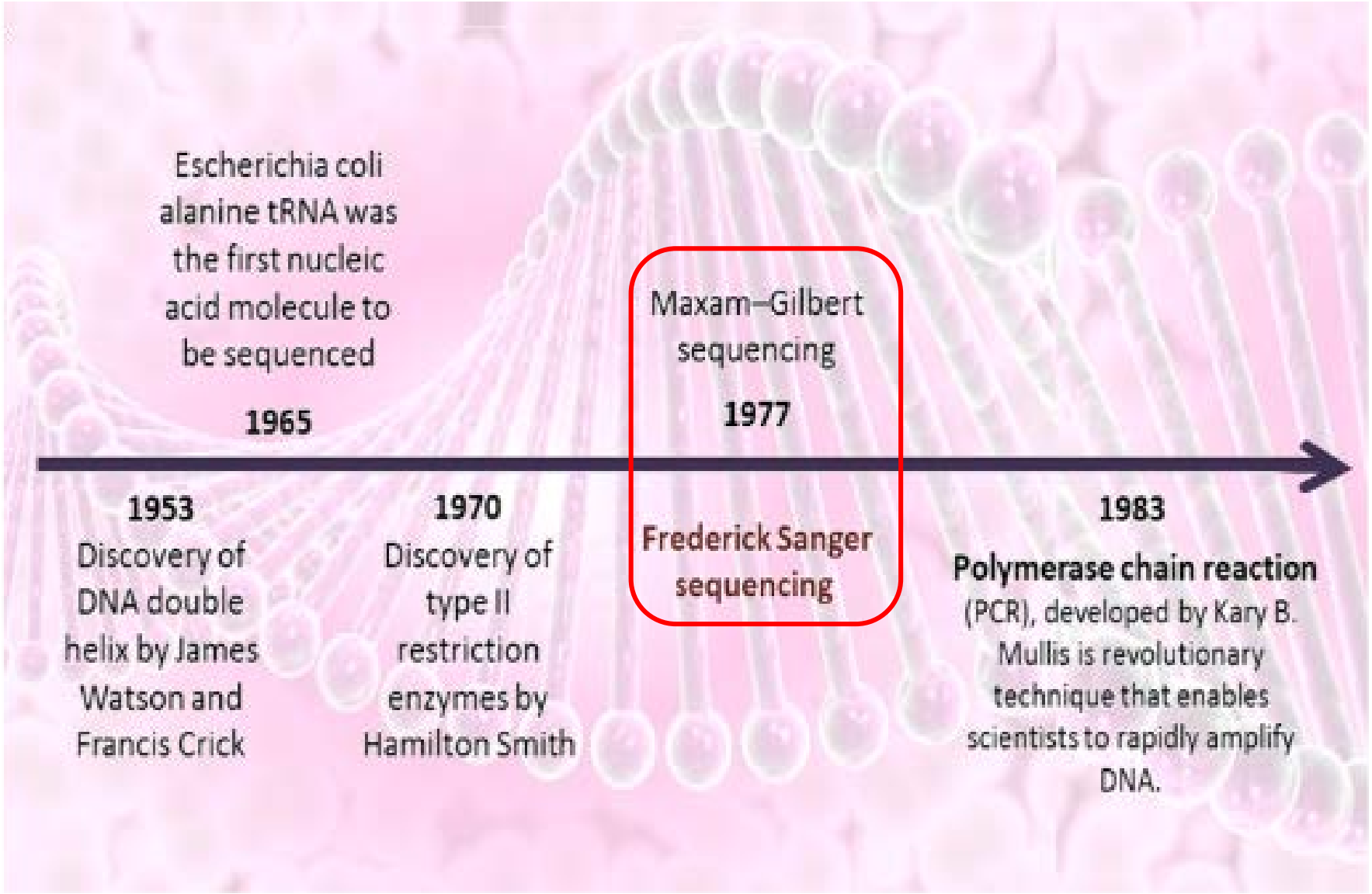


Maurice Hugh Frederick Wilkins
Prize share: 1/3

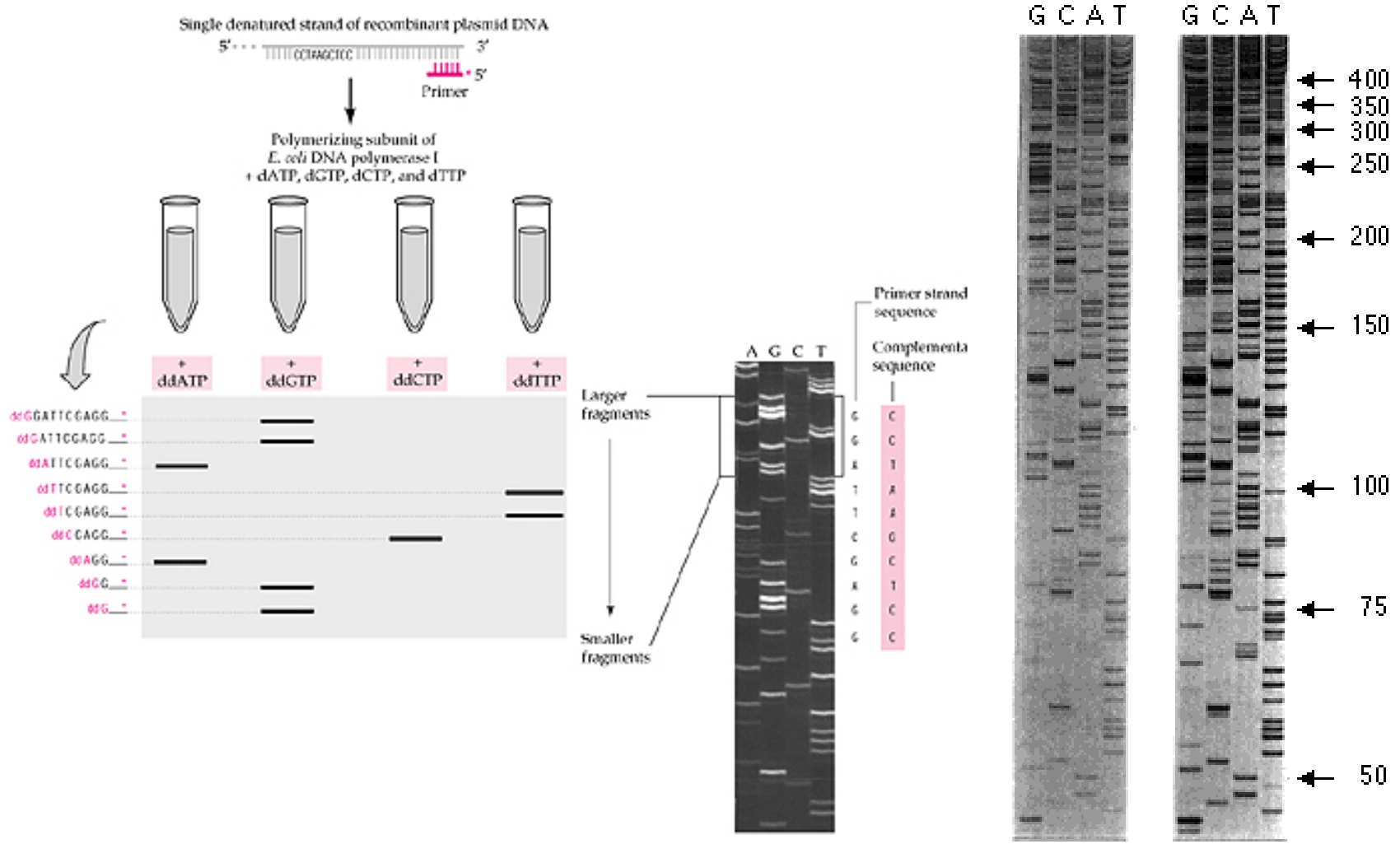


James Watson and Francis Crick with their DNA model at the Cavendish Laboratories in 1953

Do You Read Me?



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).

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
Recombinant DNA
1/2 of the prize
USA
Stanford University
Stanford, CA, USA



Walter Gilbert
Chemical sequencing
1/4 of the prize
USA
Harvard University,
Biological Laboratories
Cambridge, MA, USA



Frederick Sanger
Chain-termination sequencing
1/4 of the prize
United Kingdom
MRC Laboratory of
Molecular Biology
Cambridge, United

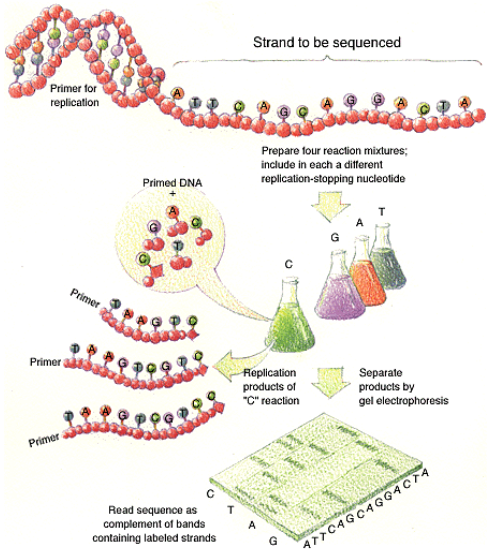
Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

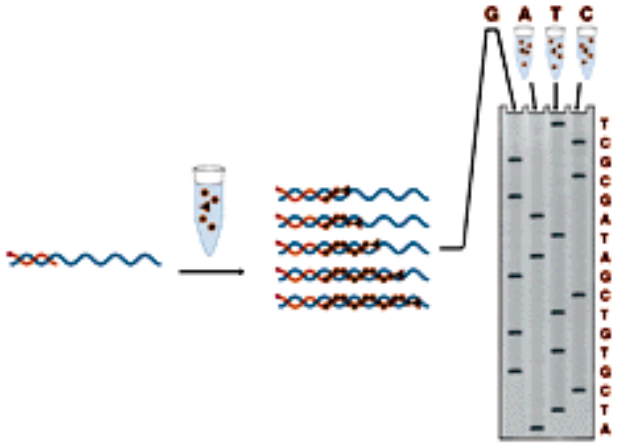
DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage φX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON
Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

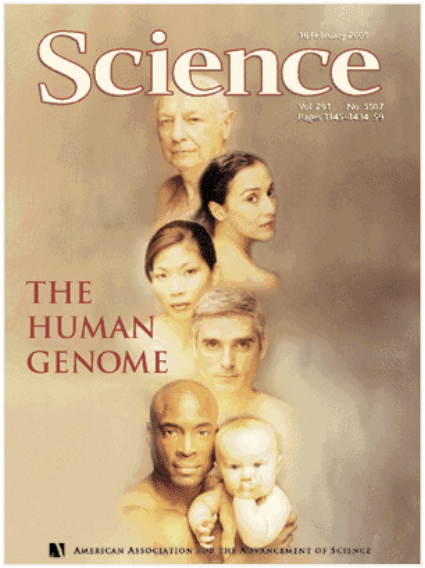
Contributed by F. Sanger, October 3, 1977





February 2001: Completion of the Draft Human Genome

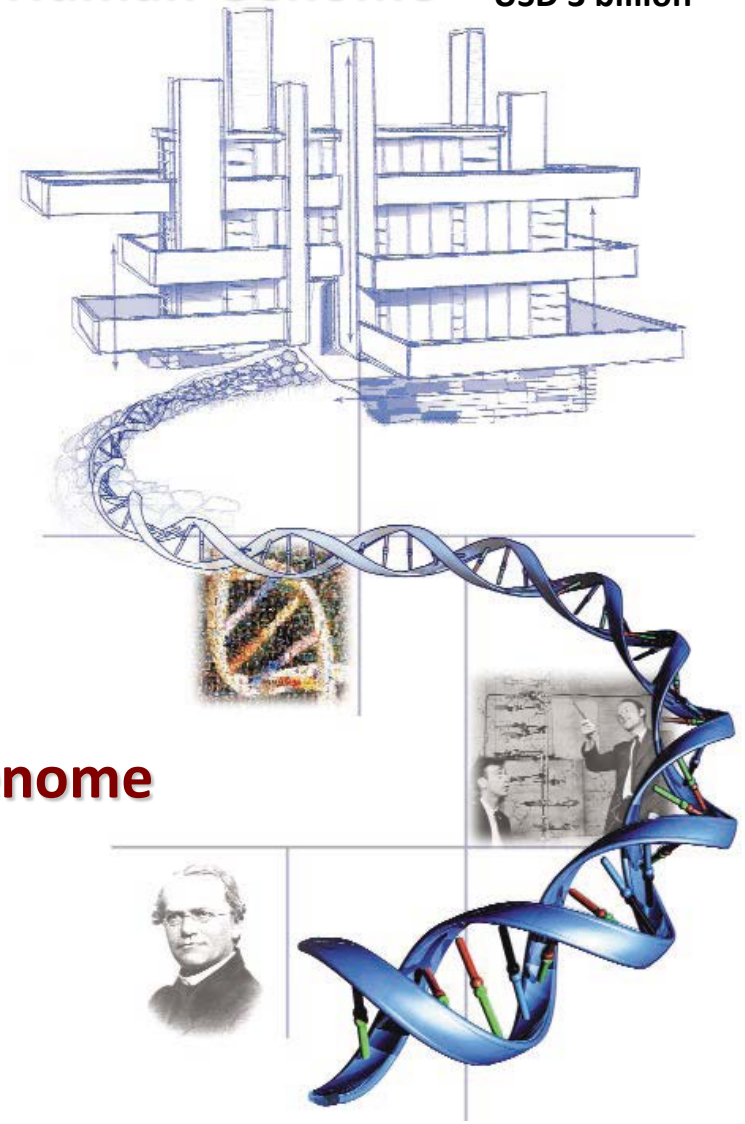
>10 years to finish
USD 3 billion



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

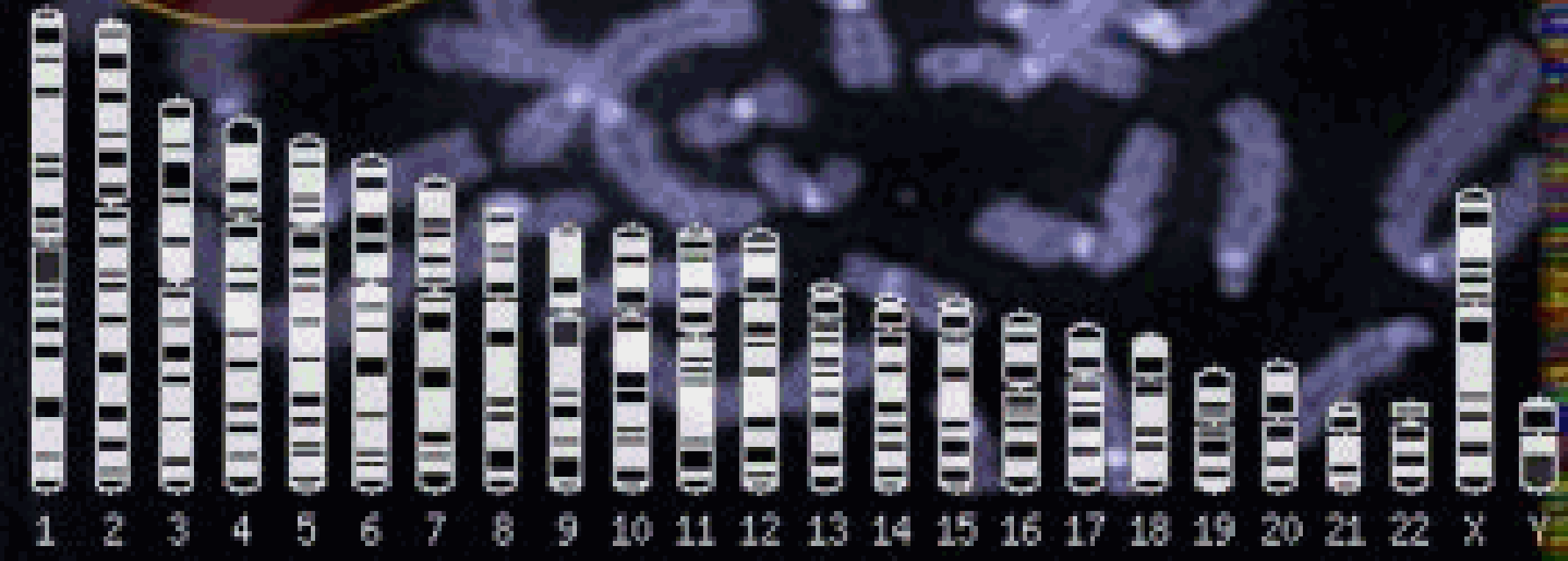
Nature, 23 April 2003
Vol. 422, Pages 1-13

Human Genome Project

3G

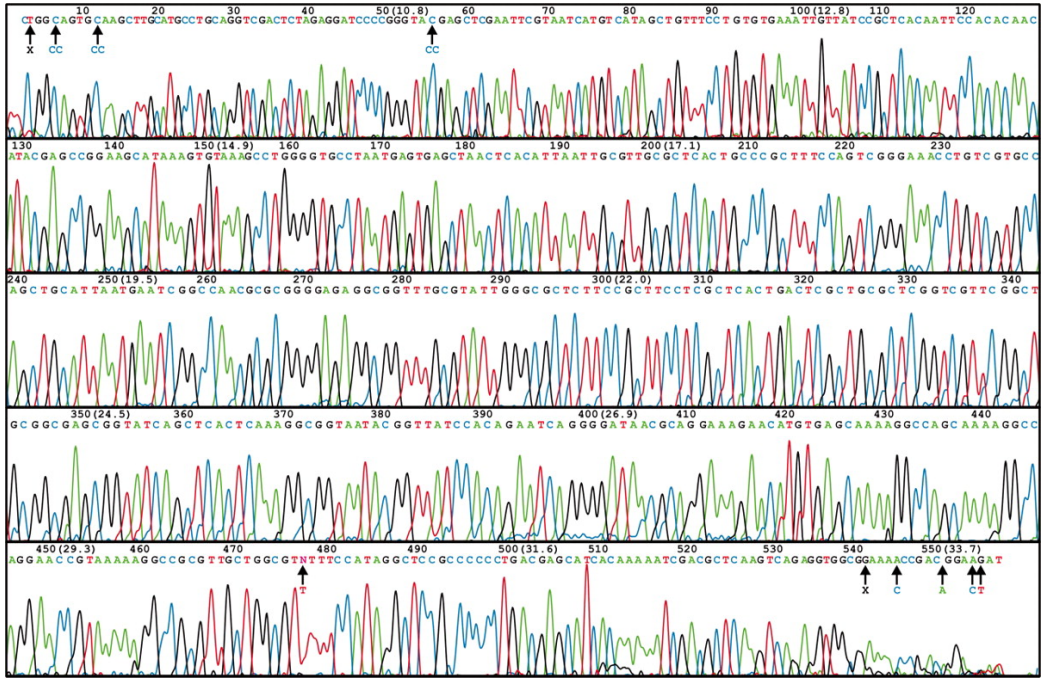
3,000,000,000 bp

3 billion base pairs



ABI 3730 XL DNA Sequencer

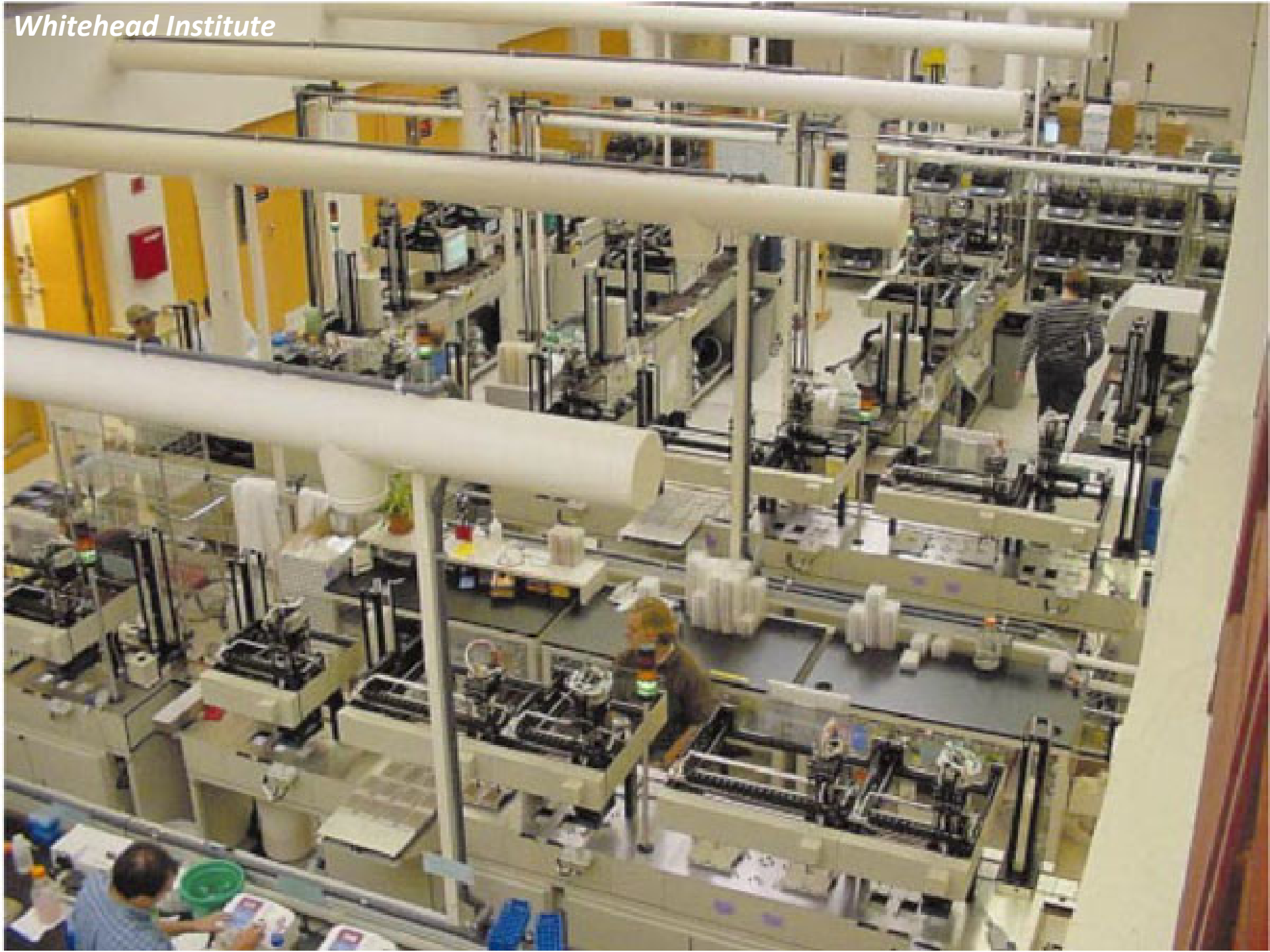




96/384 DNA sequencing in 2 hrs, approximately 600-1000 readable bps per run.

1-4 MB bps/day

A human genome of 3GB need 750 days to finish


Whitehead Institute



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[5], Michael L. Metzker[5], Susan L. Naylor[5], Raju S. Kucherlapati[5], David L. Nelson, & George M. Weinstock[5]

RIKEN Genomic Sciences Center: Yoshiyuki Sakaki[6], Asao Fujiyama[6], Masahira Hattori[6], Tetsushi Yada[6], Atsushi Toyoda[6], Takehiko Itoh[6], Chiharu Kawagoe[6], Hidemi Watanabe[6], Yasushi Totoki[6] & Todd Taylor[6]

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[3], David Kulp[34], Doron Lancet[41], Todd M. Lowe[42], Aoife McLysaght[37], Tarjei Mikkelsen[38], John V. Moran[43], Nicola Mulder[29], 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[30], 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

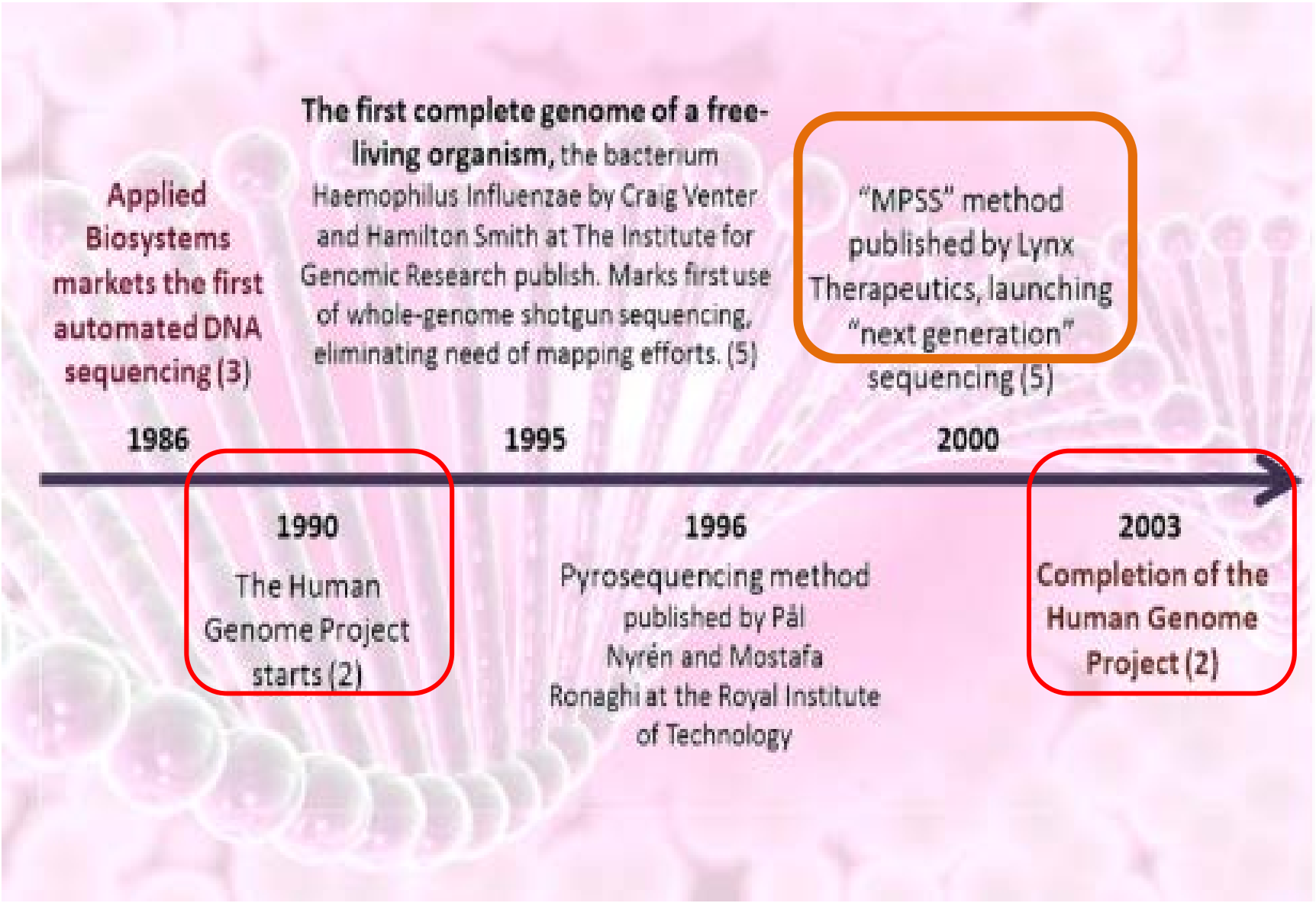
Abstract

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)

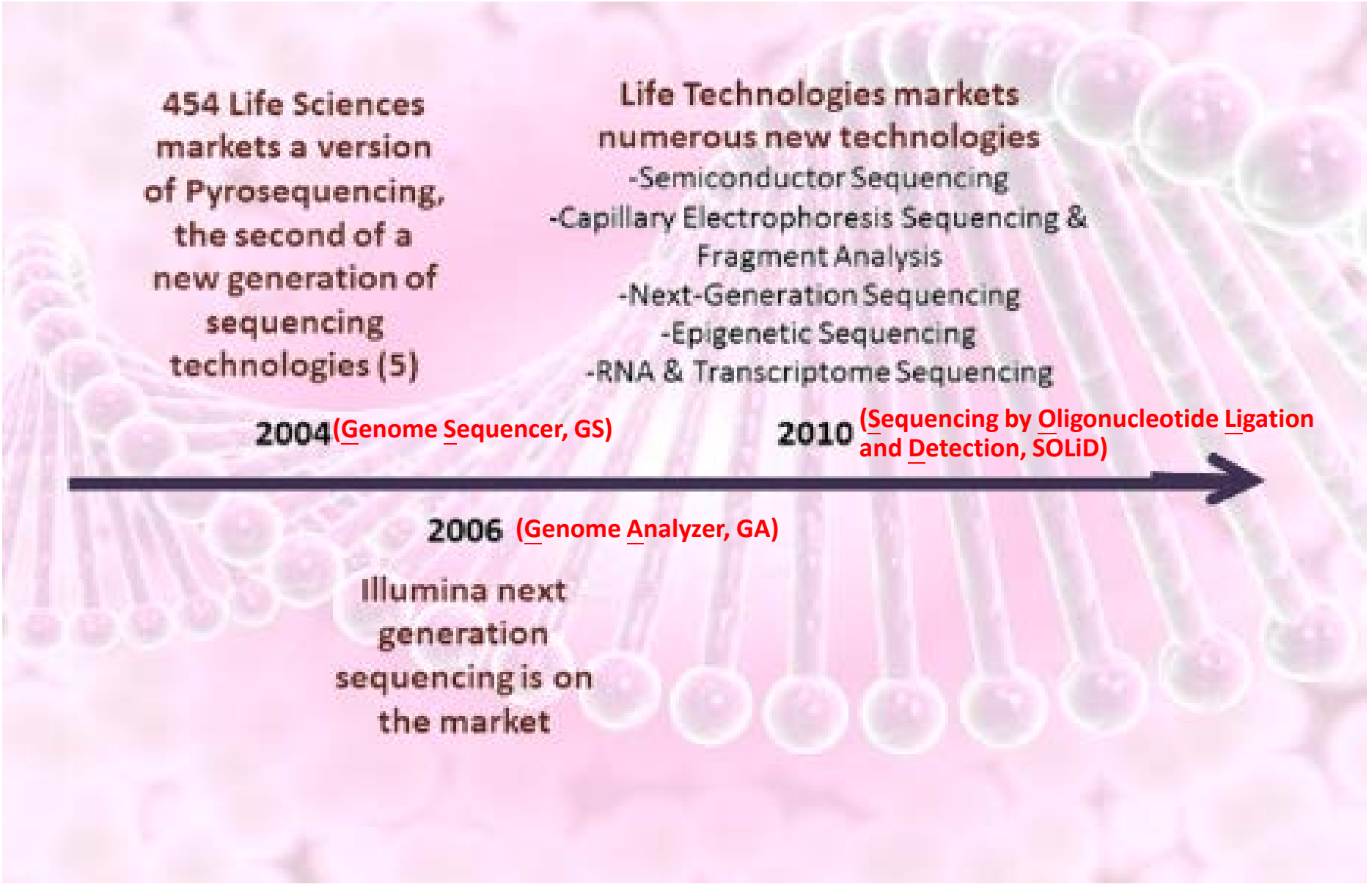


NEXT GENERATION SEQUENCING

NGS

Next Generation Sequencing Technology

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



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)

Throughput = total number of reads x read length

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

High throughput
Short read/run
Illumina
SOLiD

Low throughput-
Long read/run
Roche
(~4Gb/4 days)

Throughput of NGS machines: Roche (2012)

Roche 454 SEQUENCING



GS FLX+ System Performance	
Read Length	Up to 1,000 bp
Mode Read Length	700 bp
Throughput Profile	85% of total bases from reads >500 bp 45% of total bases from reads >700 bp
Typical Throughput	700 Mb
Reads per Run	1,000,000
Consensus Accuracy*	99.997%
Run Time	23 hours



	GS Junior System
Read Length	~400 bp
Throughput Profile	- 85% of total bases from reads >400 bp in length - 45% of total bases from reads >500 bp in length*
Throughput	~35 Mb

Throughput of NGS machines: Roche

Roche 454 SEQUENCING





Roche Shutting Down 454 Sequencing Business

Oct 15, 2013 | a GenomeWeb staff reporter

NEW YORK (GenomeWeb News) – Roche is shuttering its 454 Life Sciences sequencing operations and laying off about 100 employees, the company confirmed today.

The 454 sequencers will be phased out in mid-2016, and the 454 facility in Branford, Conn., will be closed "accordingly," Roche said in a statement e-mailed to *GenomeWeb Daily News*.

Throughput of NGS machines: SOLiD 4 (2013)

life technol applied biosystems



	5500 SOLiD™ Sequencer with microbeads	5500xl SOLiD™ Sequencer with microbeads	5500xl SOLiD™ Sequencer with nanobeads
System accuracy[1]	up to 99.99%	up to 99.99%	up to 99.99%
Typical throughput/day[2]	10–15 Gb/day	20–30 Gb/day	30–45 Gb/day
Samples/run[3]	• 1 genome • 12 exomes • 6 transcriptomes	• 2 genomes • 24 exomes • 12 transcriptomes	• 3 genomes • 40 exomes • 20 transcriptomes
Independent lanes	1–6	1–12	1–12
Typical run time	• 1 day for 35 bp (1 lane) • 7 days for 60 bp x 60 bp (6 lanes) • 7 days for 75 bp x 35 bp (6 lanes)	• 1 day for 35 bp (1 lane) • 7 days for 60 bp x 60 bp (12 lanes) • 7 days for 75 bp x 35 bp (12 lanes)	• 1 day for 35 bp (1 lane) • 7 days for 60 bp x 60 bp (12 lanes) • 7 days for 75 bp x 35 bp (12 lanes)
Typical throughput/run[4]	90 Gb	180 Gb	300 Gb
Read length	• MP: 60 bp x 60 bp • PE: 75 bp x 35 bp • Fragment: 75 bp	• MP: 60 bp x 60 bp • PE: 75 bp x 35 bp • Fragment: 75 bp	• MP: 60 bp x 60 bp • PE: 75 bp x 35 bp • Fragment: 75 bp
Multiplexing	96 for RNA and DNA	96 for RNA and DNA	96 for RNA and DNA

In 2014 ABI SOLiD was replaced by the Ion Torrent sequencing platforms (Proton and Personal Genome Machine)

Next Generation Sequencing Technology

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

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

Method of the year 2007



***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

~~ABI~~ **SOLID**

http://marketing.appliedbiosystems.com/

illumina® 2015 NGS Machines: Hiseq 3000/4000



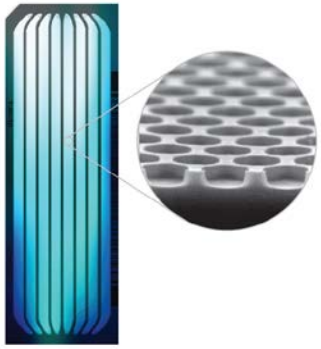


Table 1: Performance Parameters of the HiSeq 3000/4000 Systems.[a]

	HiSeq 3000 System	HiSeq 4000 System
Number of Flow Cells per Run	1	1 or 2
Output[b] 2 × 150 bp 2 × 75 bp 1 × 50 bp	 630–750 Gb 315–375 Gb 105–125 Gb	 1300–1500 Gb 650–750 Gb 215–250 Gb
Clusters Passing Filter (Single Reads)	2.1–2.5 billion	4.3–5 billion
Quality Scores	≥ 75% of bases above Q30 at 2 × 150 bp	≥ 75% of bases above Q30 at 2 × 150 bp
Daily Throughput	> 200 Gb	> 400 Gb
Run Time	< 1–3.5 days	< 1–3.5 days
Human Genomes per Run[c]	up to 6	up to 12
Exomes per Run[d]	up to 90	up to 180
Transcriptomes per Run[e]	up to 50	up to 100

2014 NGS Machines: Hiseq X



Table 1: HiSeq X System Sequencing Capacity.

	HiSeq X Ten System	HiSeq X Five System
Minimum Number of Instruments	10	5
Annual Genome Capacity	> 18,000	> 9000
Price per 30x genome	< $1000	< $1500

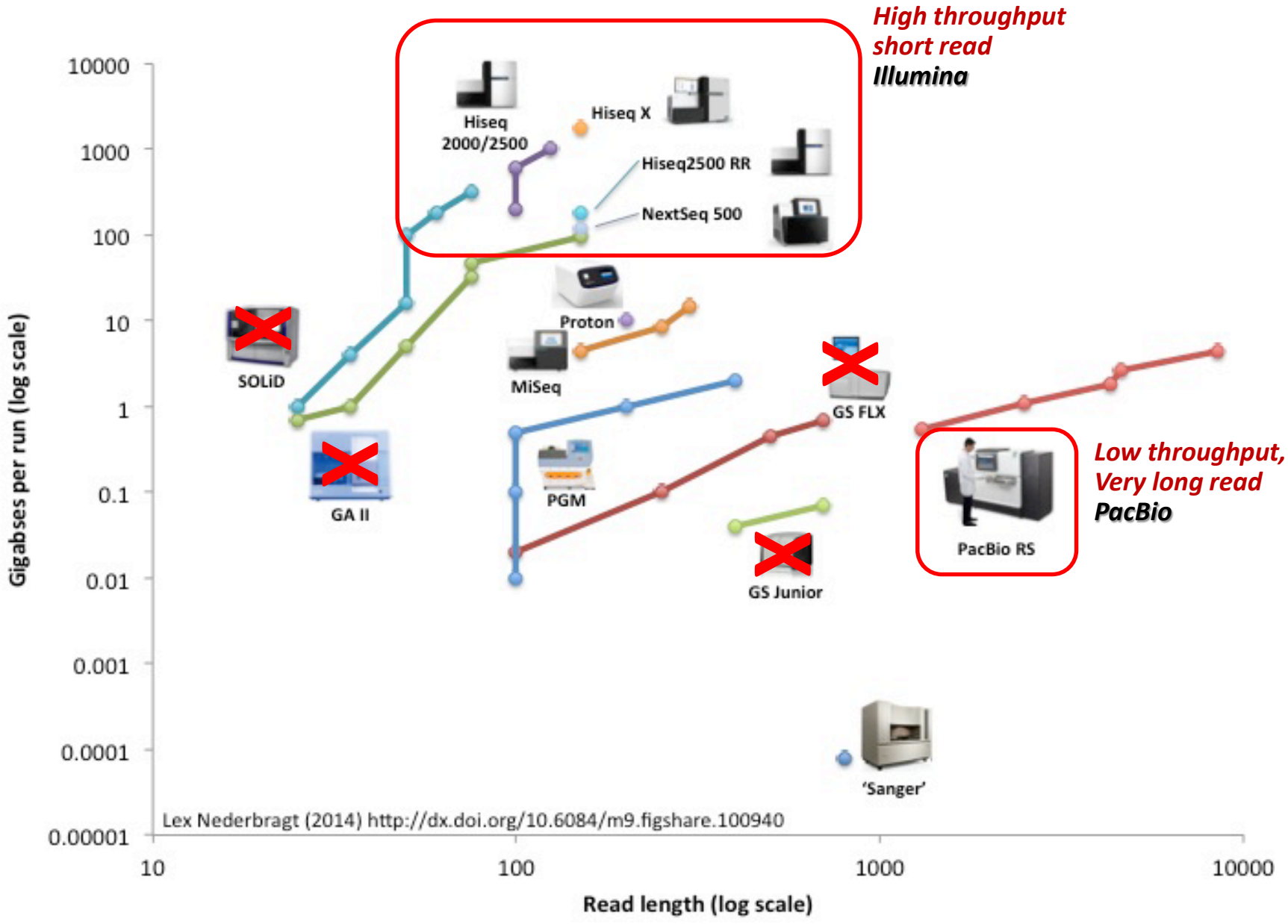
When operating at scale, HiSeq X Ten and HiSeq X Five Systems generate a staggering level of throughput, sequencing thousands of genomes per year.

Table 2: HiSeq X System Performance Parameters.[a]

Output per Run	Dual flow cell: 1.6-1.8 Tb
Single Reads Passing Filter	Dual flow cell: 5.3-6 billion
Supported Read Length	2 × 150 bp
Run Time	< 3 days
Quality	≥ 75% of bases above Q30 at 2 × 150 bp
Supported Library Preparation	TruSeq DNA PCR-Free Library Prep Kit TruSeq Nano DNA Library Prep Kit

a. Specifications based on Illumina PhiX control library at supported cluster densities (1255–1412 K clusters/mm²) on 1 HiSeq X System. Supported library preparation kit includes TruSeq Nano DNA Kit and TruSeq PCR-Free DNA Kit with 350 bp target insert size and HiSeq X Reagent Kit v2. The HiSeq X System was designed, optimized, and licensed for human WGS. Other applications and species are not permitted.

Throughput of NGS machines



NEXT GENERATION SEQUENCING

NGS

Basis Principles

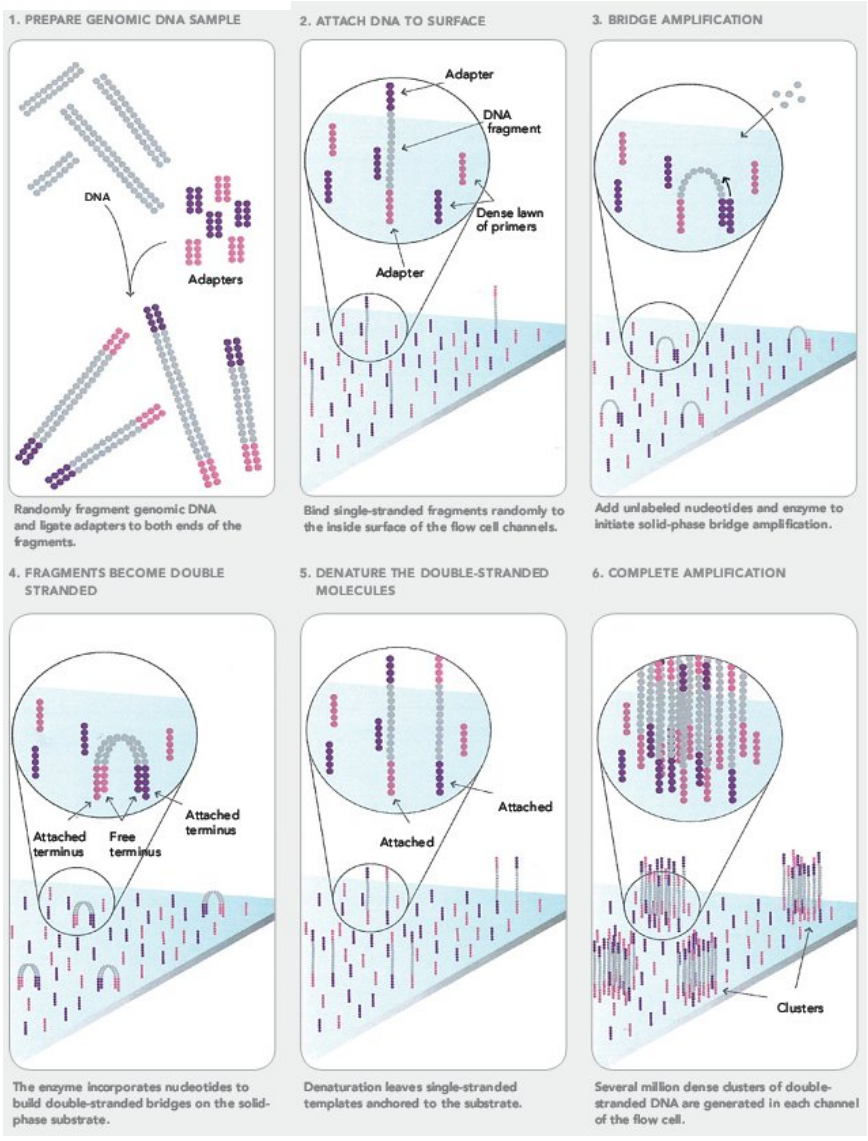
Temple fragmentation (library) ➡ Amplification ➡ Sequencing ➡ Detection

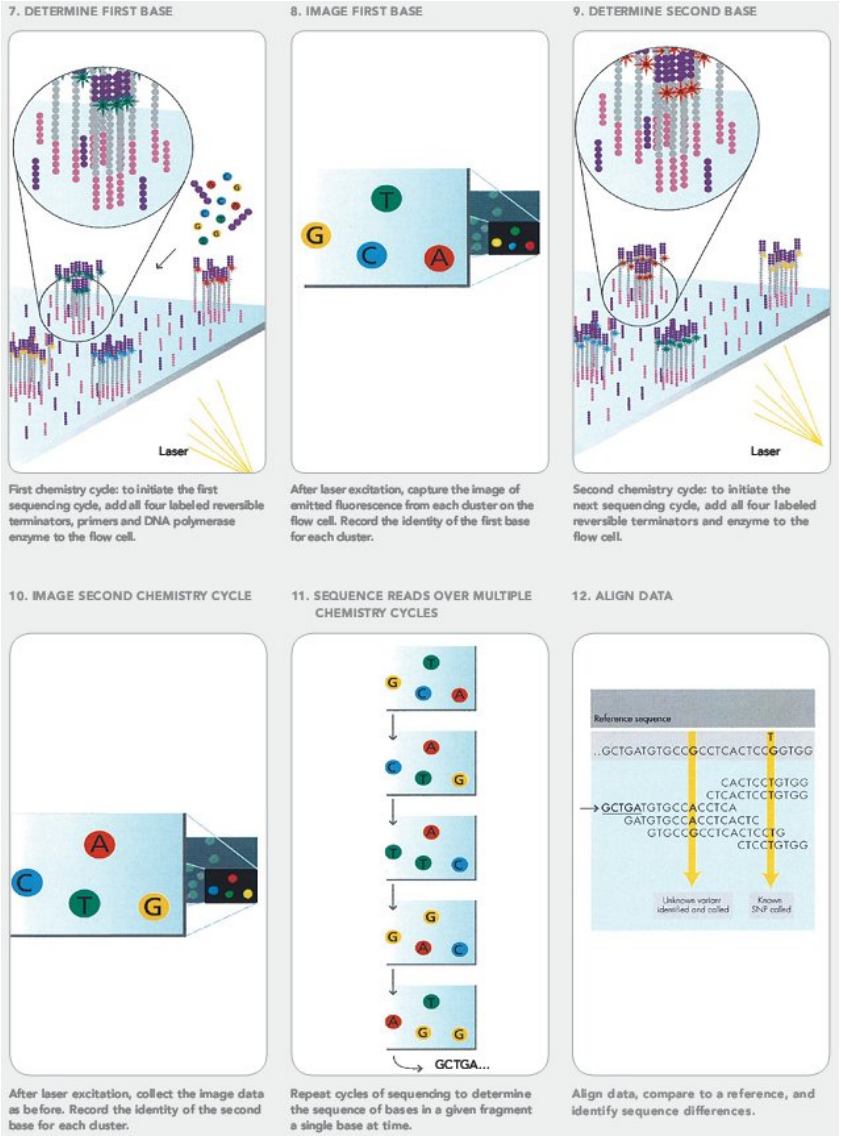
Illumina Sequencing by Synthesis (SBS)

illumina
Solexa



Solexa Genome Analyzer

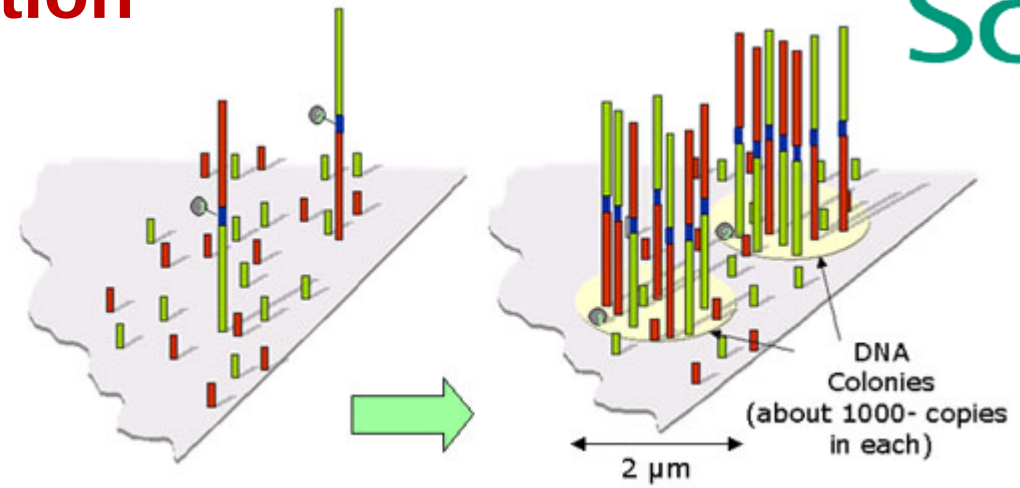




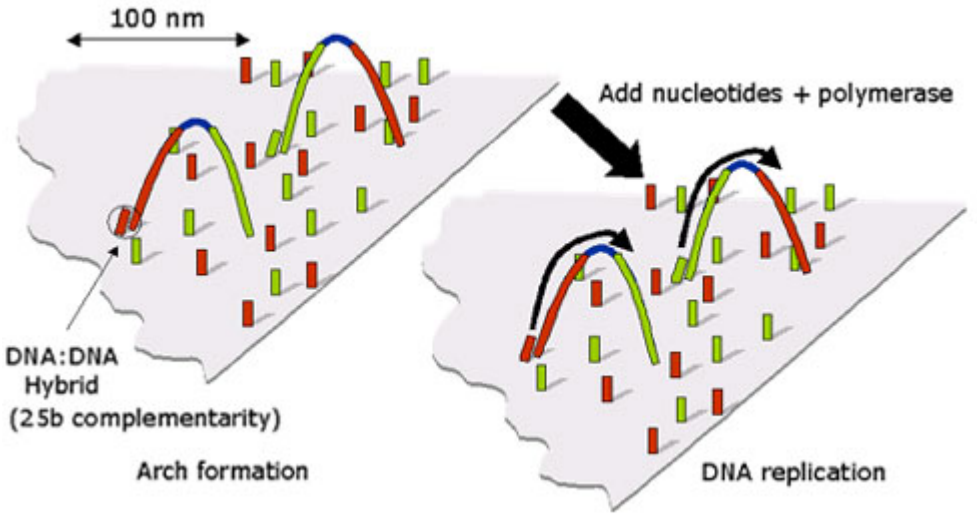
Cluster Generation

Solexa™

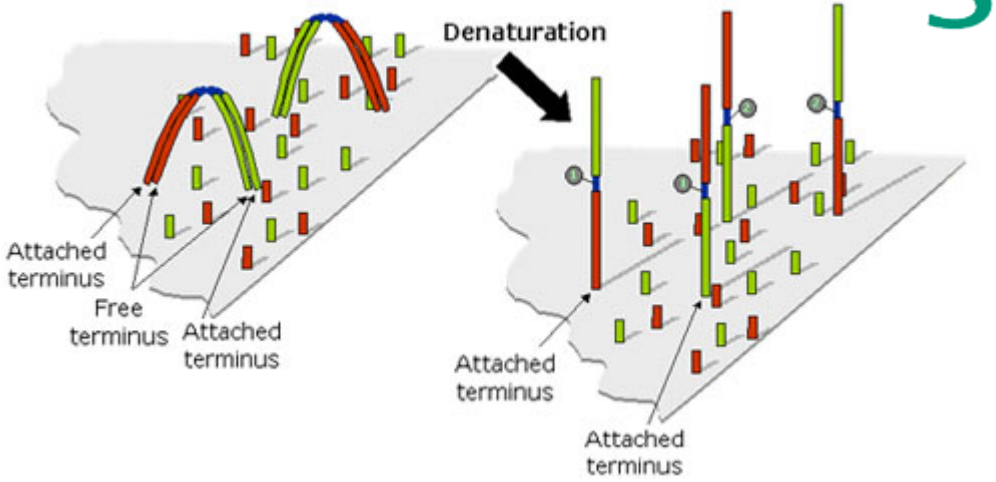
Clusters Array



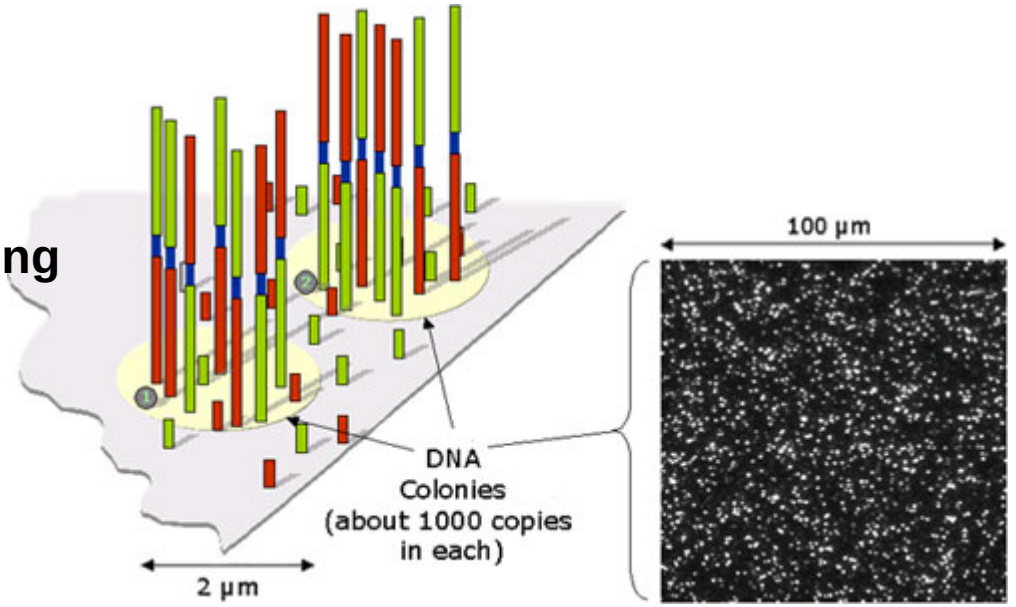
Template Bridging



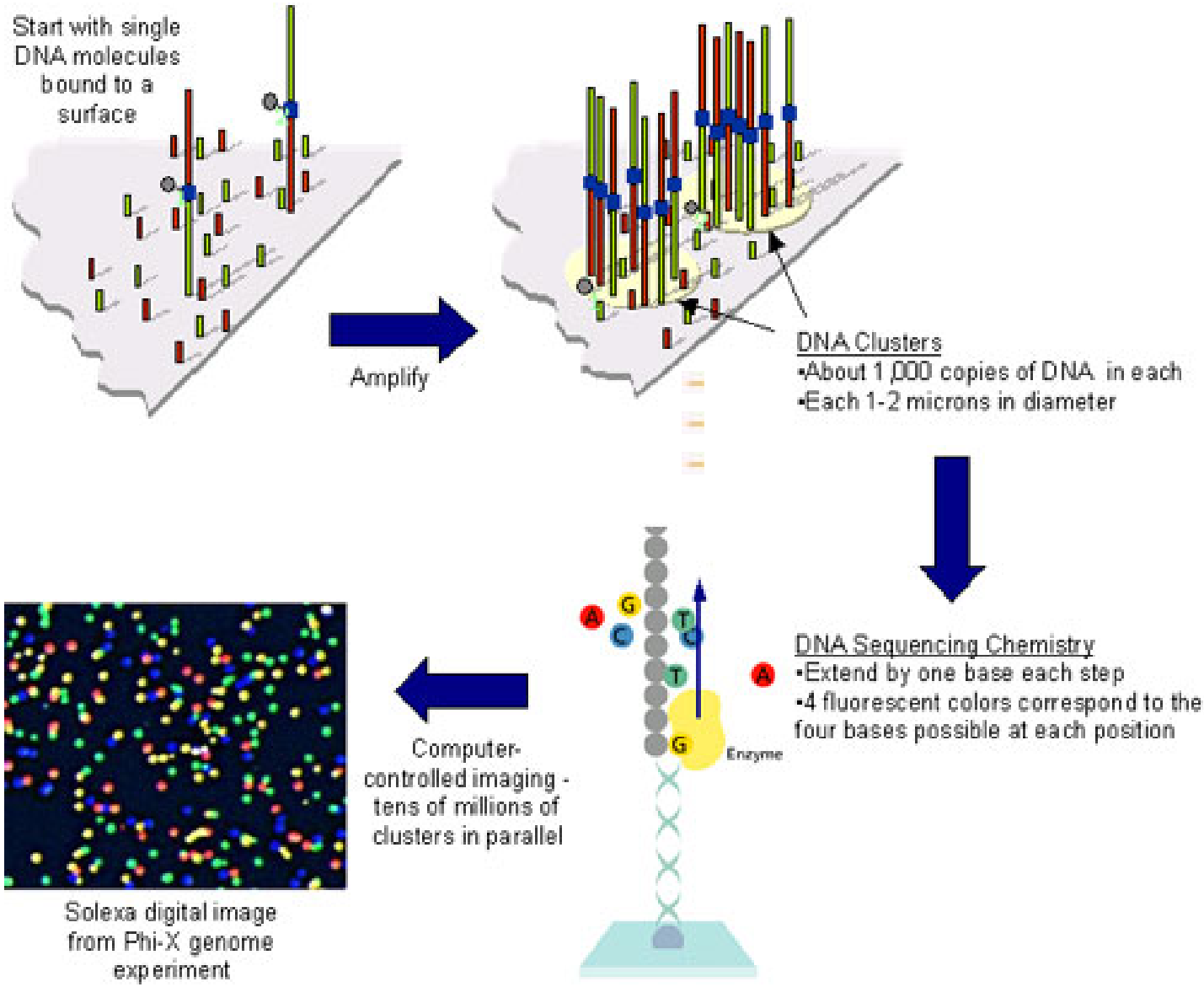
Denaturation



Ready For Sequencing



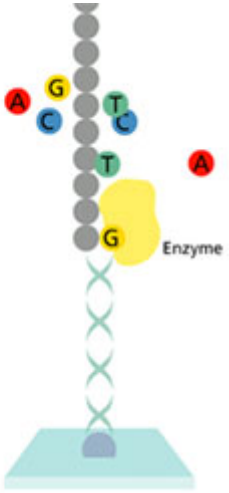
Sequencing-By-Synthesis



Solexa™

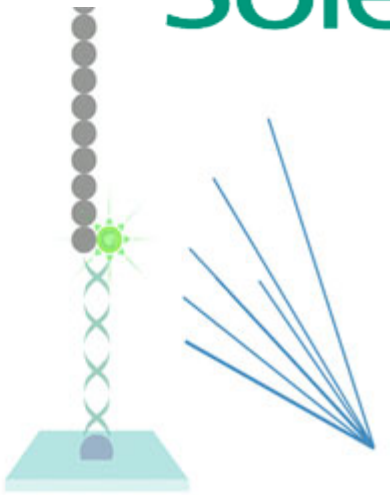
Cycle 1

Incorporation

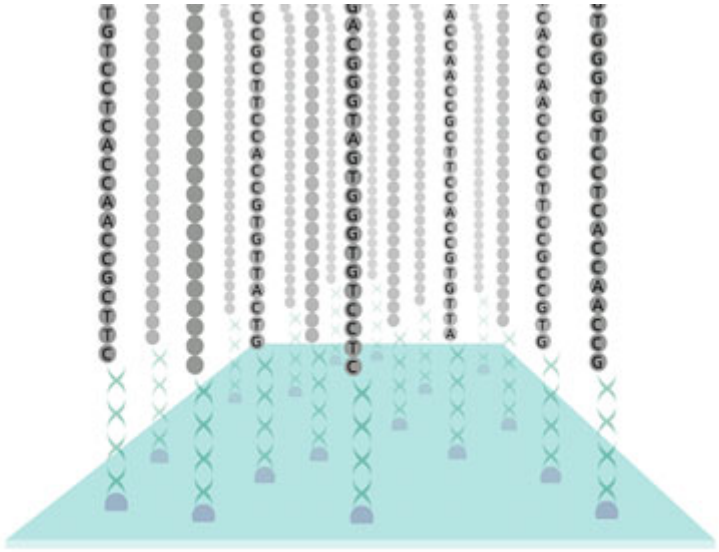


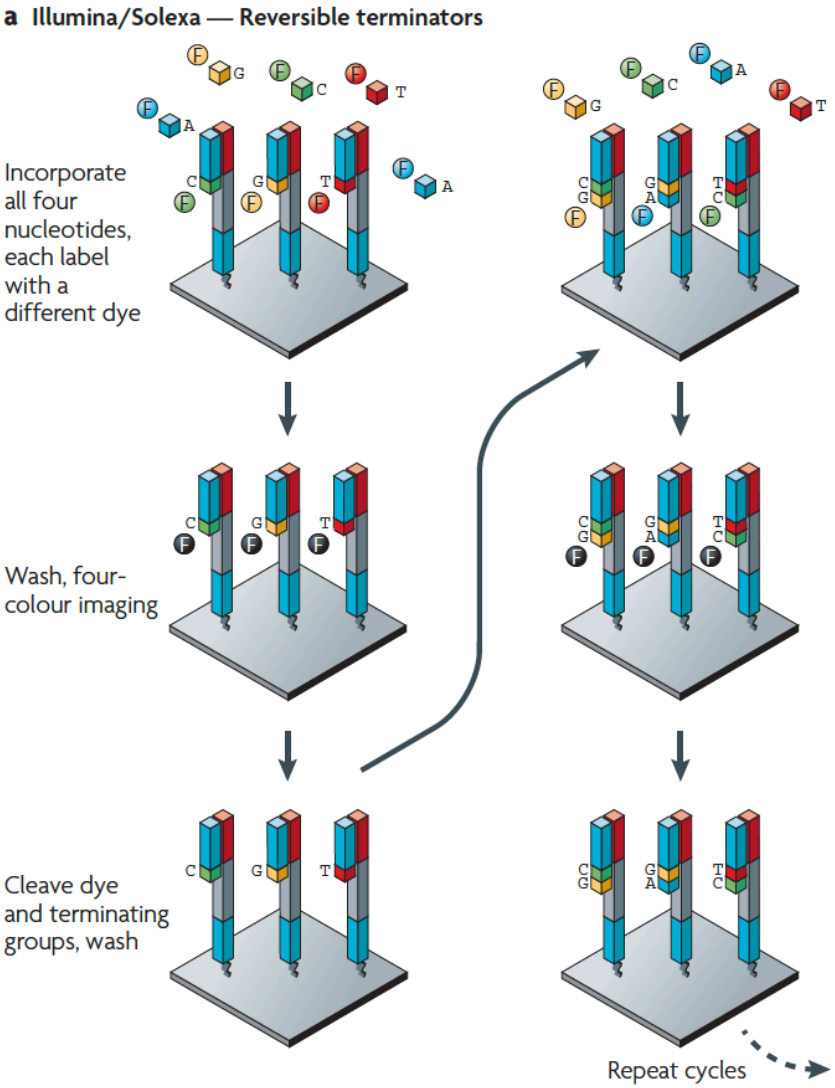
Cycle 1

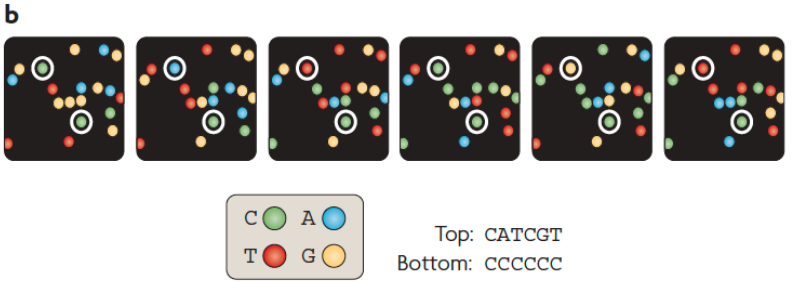
Fluorescent Detection



Sequence Generated At Every Site On The Array

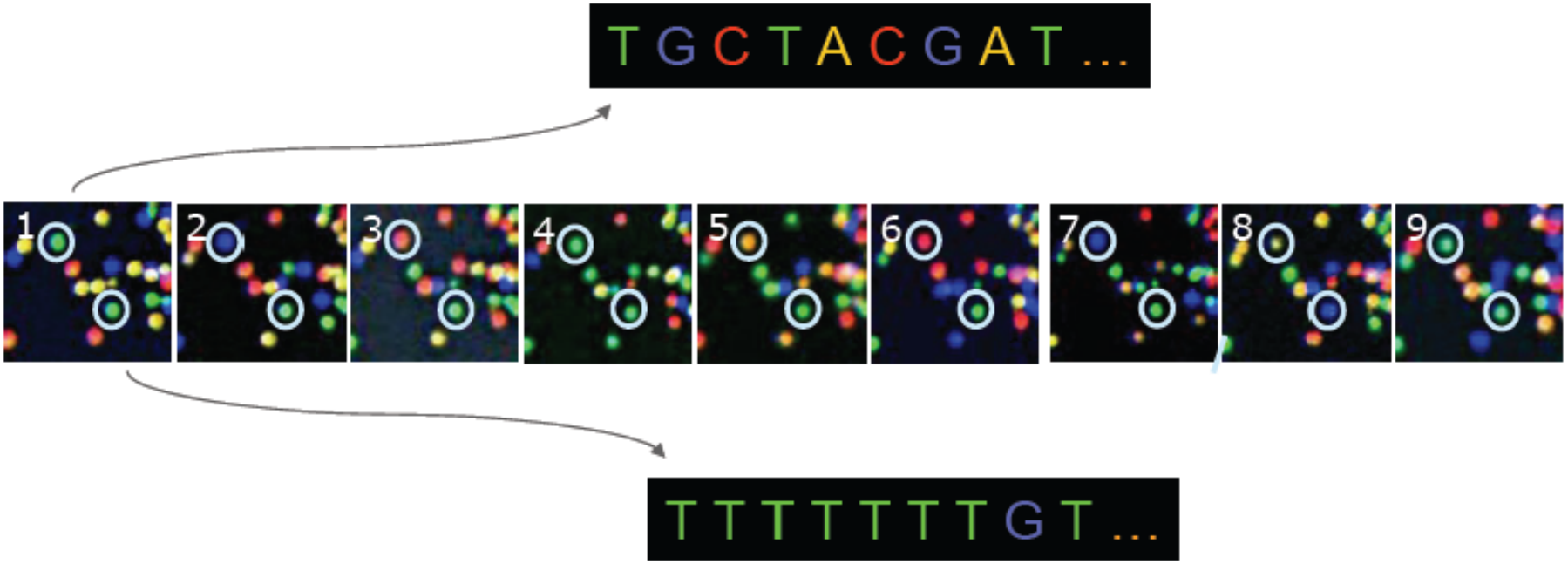






From Debbie Nickerson, Department of Genome Sciences, University of Washington, http://tinyurl.com/6zbzh4

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

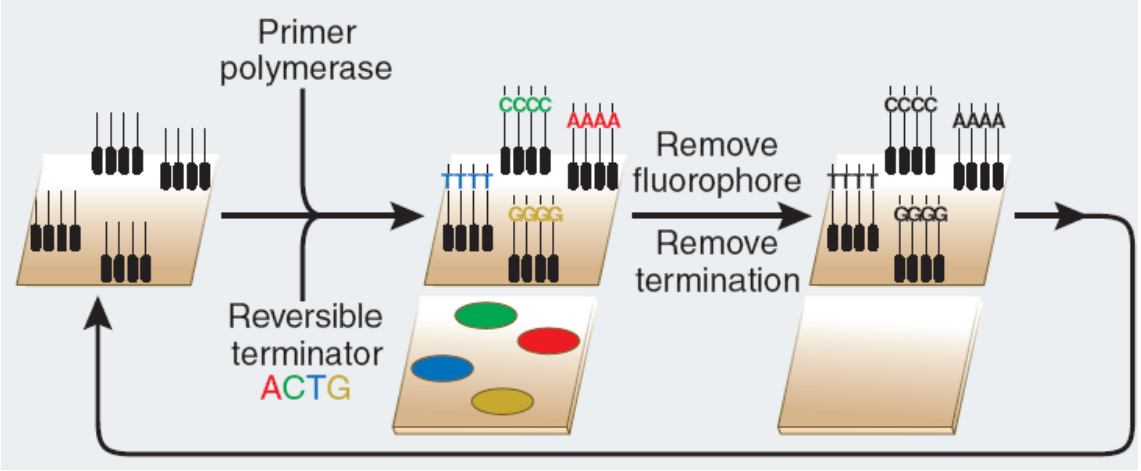
Solexa technology (sequencing-by-synthesis)

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.

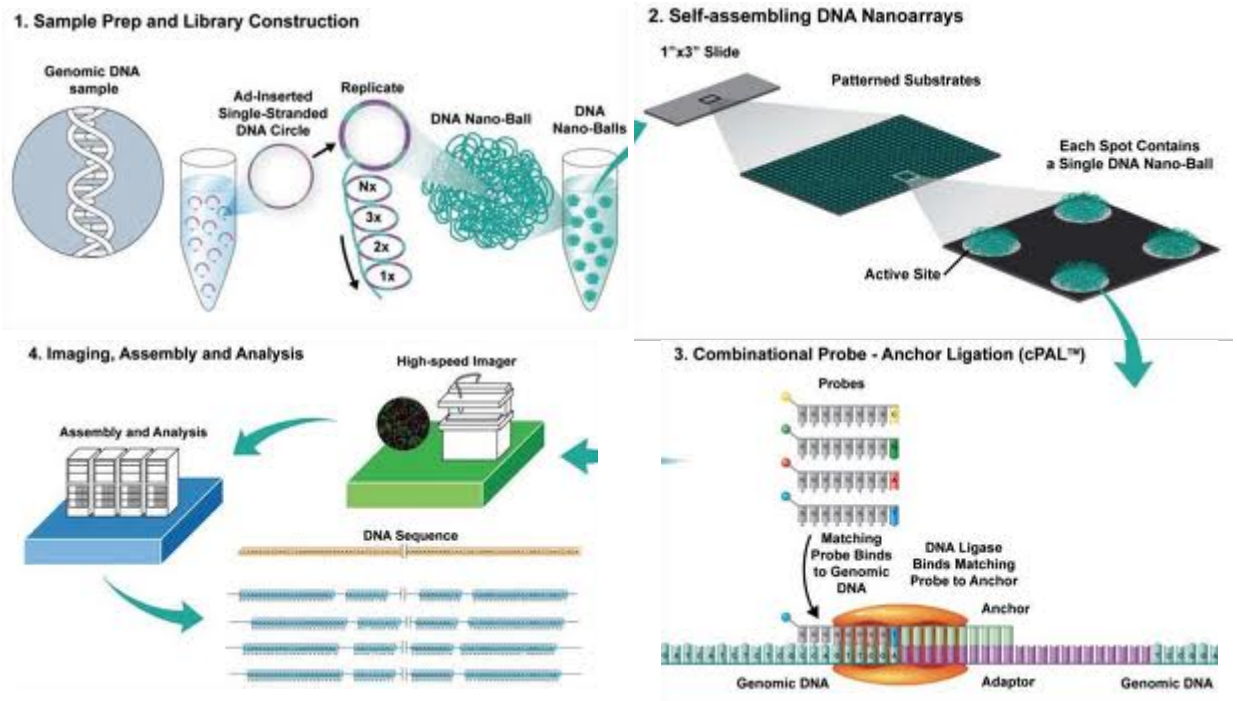
Sequencing with reversible terminators.

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.



Combinatorial Probe-Anchor Ligation (cPAL) sequencing

Complete genomics



Technology description: Drmanac, R et al. (2010) Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 327, 78-81

Evolution of Sequencing Technology



Sanger dideoxy-sequencing

ABI 3730XL

Massive parallel sequencing

~~Roche 454 FLX~~, ~~Life Technologies SOLiD,~~
Illumina Hiseq, Life Tech Ion Torrent

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,, Electron Optica, Electronic BioSciences, GenapSys, Genia, GnuBIO, IBM, LaserGen, Lightspeed Genomics, NABsys, NobleGen Biosciences, QuantuMDx, Reveo, Stratos Genomics, Two Pore Guys, ZS Genetics……

NEXT-NEXT-NEXT-

NEW GENERATION SEQUENCING

NGS

Basis Principles

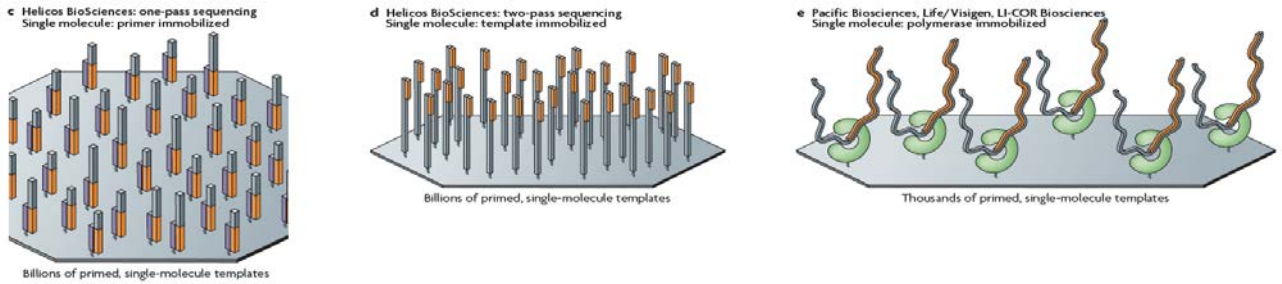
Temple fragmentation (library) ➡ Amplification ➡ Sequencing ➡ Detection

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



~~VisGen Biotechnologies~~

Pacific Biosciences

~~Helicos Biosciences~~

Helicos Sells Certain MDx IP to Sequenom for $1.3M, Warns of Potential Bankruptcy

April 11, 2012

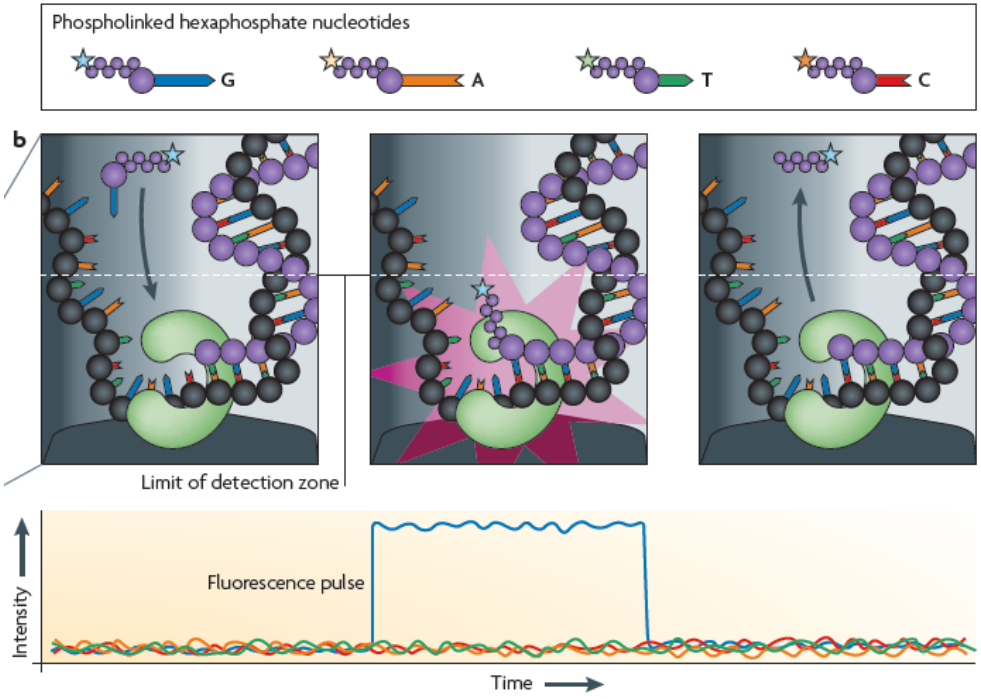
Helicos BioSciences said in a regulatory filing that it has sold certain patent applications covering molecular diagnostic applications of its technology to Sequenom for $1.3 million. The patent applications cover methods for detecting fetal nucleic acids and diagnosing fetal abnormalities. Helicos also warned that if it is "unable to successfully raise additional capital, we may have to cease operations and/or seek bankruptcy protection."

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

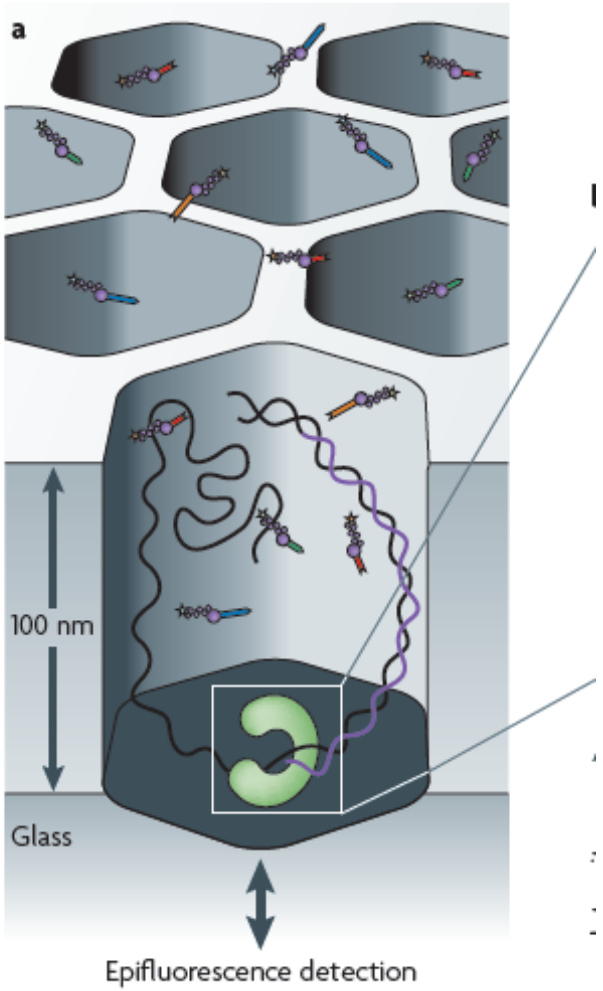
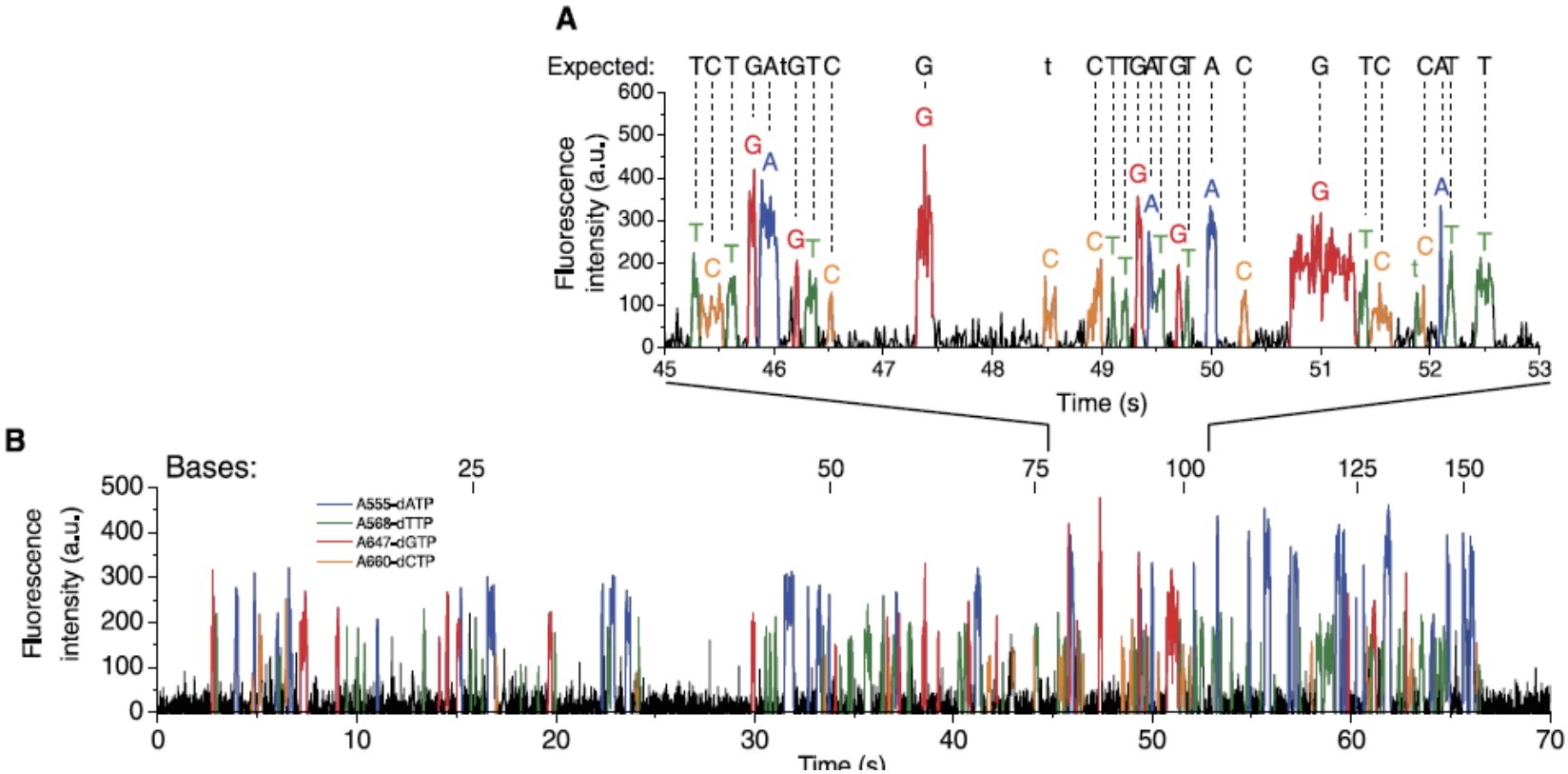




Figure 10. Processive Synthesis with Phospholinked Nucleotides.

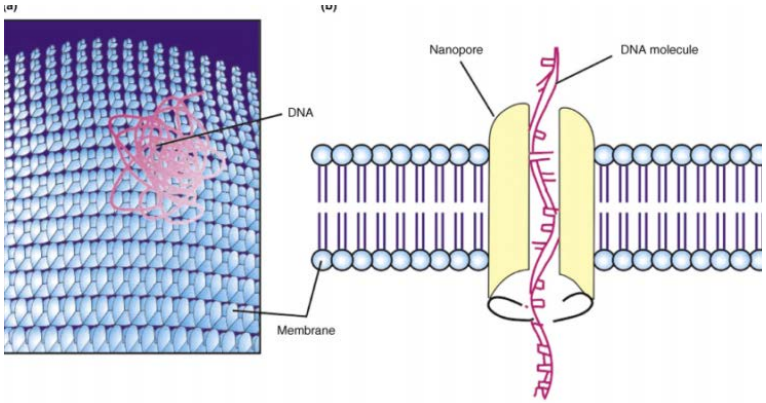
Step 1: Fluorescent phospholinked labeled nucleotides are introduced into the ZMW.
Step 2: The base being incorporated is held in the detection volume for tens of milliseconds, producing a bright flash of light.
Step 3: The phosphate chain is cleaved, releasing the attached dye molecule.
Step 4-5: The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



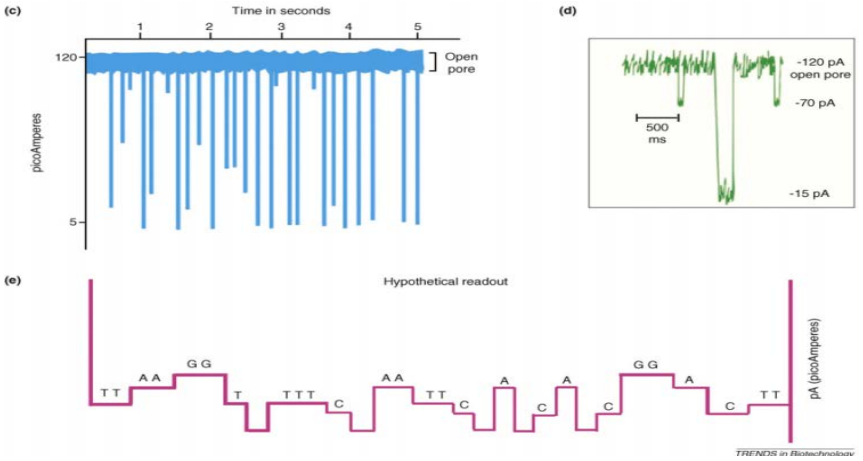
Nanopore Sequencing

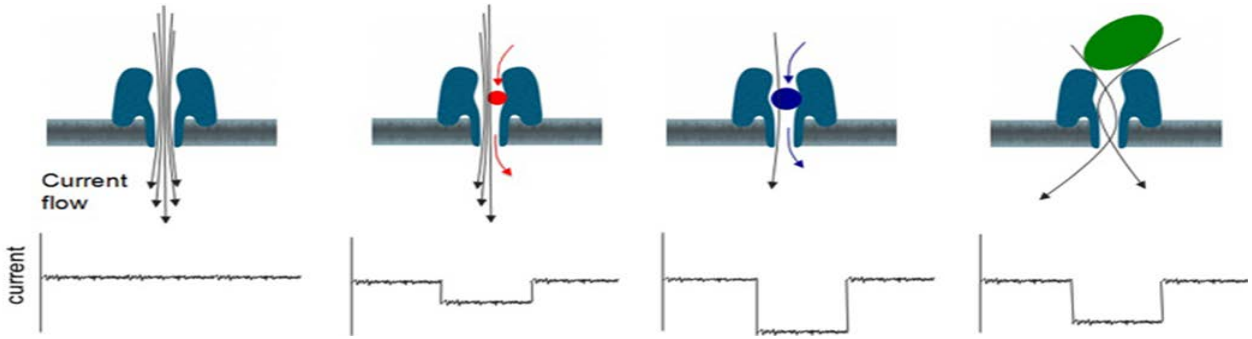
Oxford Nanopore

Nucleic acids driven through a nanopore.



Differences in conductance of pore provide readout.





This diagram shows a protein nanopore set in an electrically resistant membrane bilayer. An ionic current is passed through the nanopore by setting a voltage across this membrane.

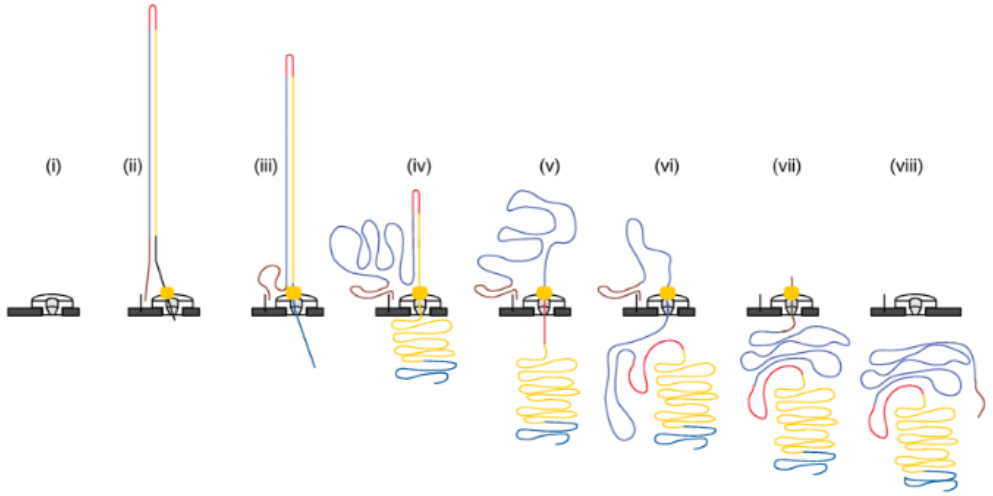
If an analyte passes through the pore or near its aperture, this event creates a characteristic disruption in current. By measuring that current it is possible to identify the molecule in question. For example, this system can be used to distinguish the four standard DNA bases and G, A, T and C, and also modified bases. It can be used to identify target proteins, small molecules, or to gain rich molecular information for example to distinguish the enantiomers of ibuprofen or molecular binding dynamics.

Bio·IT World
Indispensable Technologies Driving Discovery, Development, and Clinical Trials

Oxford Strikes First in DNA Sequencing Nanopore Wars

By Kevin Davies

February 17, 2012 | Breaking a near total vow of silence after three years in stealth mode, Oxford Nanopore has offered stunning details of its nanopore next-generation sequencing (NGS) technology to a packed house at the premier genome sequencing conference in Marco Island, Florida.



In a talk at the Advances in Genome Biology and Technology conference this morning, chief technology officer Clive Brown presented details of what the British company calls its "new generation" of single-strand sequencing technology, featuring accurate long reads of single-stranded DNA molecules, including the completion of two viral genomes.

In the second half of this year, the company will commercially release its previously announced GridION instrument, or node, as well as the world's smallest DNA sequencing instrument – a portable cartridge called the MinION that doubles (and works as) a USB drive.





Oxford Nanopore's disruptive MinION USB device



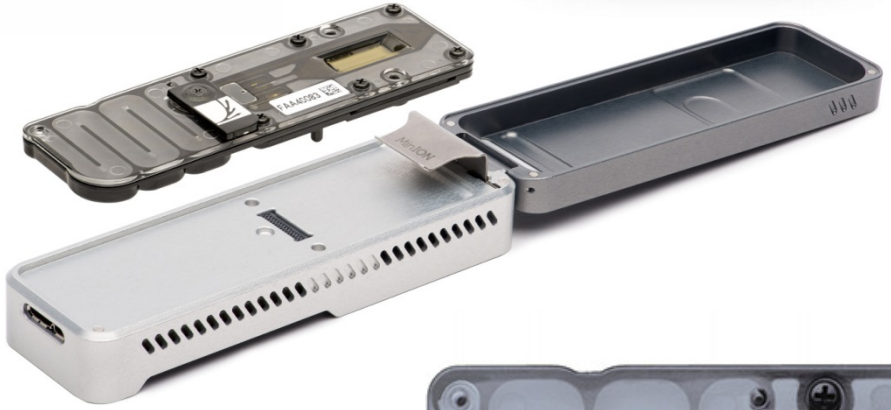


Weight

87g (103g with a flow cell)

Size

W 105, H 23, D 33mm





NATURE | NEWS

Pint-sized DNA sequencer impresses first users

Portable device offers on-the-spot data to fight disease, catalogue species and more.

Erika Check Hayden

05 May 2015





R9.4 flow cell
Throughput: 5-10 Gb

2014 *E. coli* genome
Avg. read length: 5.4 kb
Longest read: 10 kb

2015
Mean read length: 2 kb
Median read length : 1 kb
Longest read: 98 kb

Wellcome Trust Centre for Human Genetics and Genomics plc First to Sequence Multiple Human Genomes Using Hand-Held Nanopore Technology

Demonstrates potential for wide-scale whole-genome sequencing in humans using nanopore approaches

December 01, 2016 07:01 PM Eastern Standard Time

OXFORD, England--(BUSINESS WIRE)--Oxford University's Wellcome Trust Centre for Human Genetics (WTCHG) and the leading genome analytics company Genomics plc today announced the first sequencing and analysis of multiple human genomes using nanopore technology. The announcement, made on Thursday at the Oxford Nanopore Community Meeting in New York City, marks a major breakthrough in sequencing technology in opening up the potential of wide-scale whole-genome sequencing in humans using nanopore approaches.

"It is a significant breakthrough to sequence multiple human genomes with nanopore technology. Moving forward, the idea that there will be alternative technologies with new properties in this space will be very appealing to researchers and to clinical geneticists."

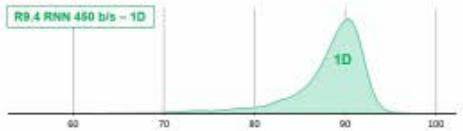
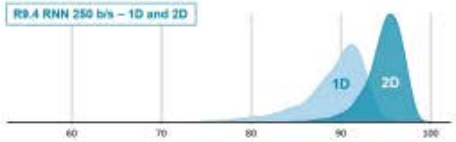
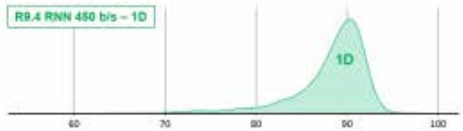
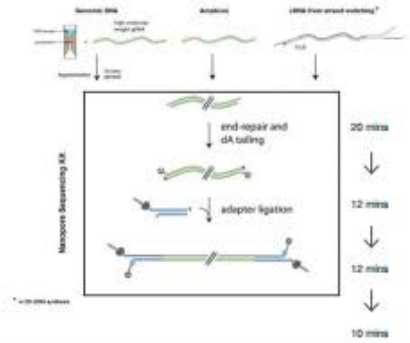
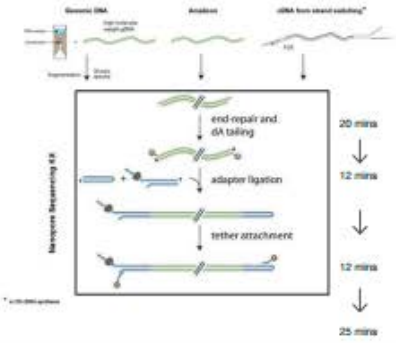
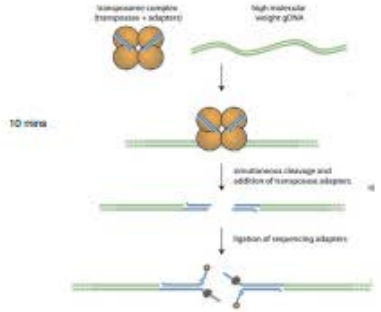
Tweet this

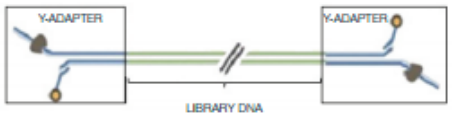
Today's announcement raises the possibility of a major change to both the economics and the science of DNA sequencing. Researchers at the WTCHG and Genomics plc used the MinION sequencer developed by Oxford Nanopore Technologies. The portable MinION is a hand-held device about the size of a Mars bar, and is powered by a laptop computer to which it connects via a USB port. It is available for $1,000 plus the cost of consumables. While nanopore sequencing has been available for two years, it has previously only been practicable to apply it to relatively small genomes or regions of DNA, such as the much smaller genomes of viruses and bacteria. Recent upgrades to the technology have now made it possible for researchers to conduct larger-scale studies such as sequencing complete human genomes by using multiple MinIONs.

Ligation Sequencing Kit 1D, Ligation Sequencing Kit 2D or Rapid Sequencing Kit?

Sequencing is used for a wide range of applications and each has its own requirements for the type of data needed to answer the biological question. Oxford Nanopore offers two ranges of kits: Rapid 1D and Ligation Sequencing Kits which provide 1D or 2D read data, allowing selection of the right tool for the right experiment. Nanopores sequence any strand of DNA presented to them, independent of the length or source. As a result a user can influence the outcome of their sequencing run by selecting the right kit and protocol for their experimental goals.

The table below provides information to guide the selection :

Property	Ligation Sequencing Kit 1D	Ligation Sequencing Kit 2D	Rapid Sequencing Kit
	SQK-LSK108 the simplest method for genomic DNA, amplicon and/or cDNA with control over read length and amounts of input material	**SQK-LSK208** the highest accuracy for genomic DNA, amplicon and cDNA, with control over read length and amounts of starting material	**SQK-RAD002** rapid and easy method for genomic DNA
Input sample	gDNA, amplicon, cDNA	gDNA, amplicon, cDNA	gDNA
Yield on R9.4 flow cells	10 GB / run	5 GB / run	10 GB / run
Accuracy	 >92%	 >96%	 >92%
Library prep	50 mins	70 mins	<10 mins
Protocol overview	A multi-step, 50 minute protocol 	A multi-step, 70 minute protocol 	A two-step, 10 minute protocol 

Property	Ligation Sequencing Kit 1D	Ligation Sequencing Kit 2D	Rapid Sequencing Kit
Amount of sample	1000 ng	1000 ng (20-1000 ng for general protocols) (5000 ng for ultra-long reads)	200 ng
Efficiency of library preparation	*** (All the adapted fragments are available for sequencing)	** (Some fragments lost due to double hairpin molecules which will not be sequenced)	*** (All the adapted fragments are available for sequencing)
Protocol / Kit adaptions	Native barcoding PCR barcoding Low input DNA Long reads (>50kb) Read length selection Wash Kit	Native barcoding PCR barcoding Low input DNA Long reads (>50 kb) Read length selection Wash Kit	Wash Kit
Additional consumables required for library prep	Agencourt AMPure XP beads (A63880) NEBNext End repair / dA-tailing Module (E7546), NEB Blunt/TA Ligase Master Mix (M0367). Optional Covaris g-TUBE (#520079), NEB FFPE Repair Mix (M6630)	Agencourt AMPure XP beads (A63880) NEBNext End repair / dA-tailing Module (E7546), NEB Blunt/TA Ligase Master Mix (M0367), MyOne C1 Streptavidin beads (65001/65001). Optional Covaris g-TUBE (#520079), NEB FFPE Repair Mix (M6630)	NEB Blunt/TA Ligase Master Mix (M0367)
Read lengths v fragment length **These are for illustrative purposes only**	Read length is the same as fragment length Read Length: 8 kb Fragment Length: 8 kb DNA sheared using Covaris g-TUBE 	Read length is double the fragment length Read Length: 16 kb Fragment Length: 8 kb DNA sheared using Covaris g-TUBE 	Read length is the same as fragment length Read Length: 4.5 kb Fragment Length: 4.5 kb DNA tagmentation by transposase 

The GridION system

Oxford Nanopore's proprietary nanopore-based sensing chemistries are operated on an electronics-based platform, the GridION system. This enables the scaled-up measurements of multiple nanopores and the sensing, processing and analysis of data in real time.

A single instrument, a GridION node, operates with a single-use cartridge that contains the necessary reagents to perform an experiment.



Scalability
A node can be employed as a single desktop instrument, or scaled up in a similar way to computing installations.

As a desktop instrument for the individual researcher, it can write data to the network or a locally attached disc (directly through a USB or through a network) and work in a small lab.

Each node is a network device and multiple nodes can be aggregated together into larger co-operating units or clusters, communicating with each other in a peer-to-peer fashion over the user's network. The system is designed to interface, or even co-locate, with standard or high-performance IT infrastructure. Workflow overheads, and total costs, including IT, scale linearly with the GridION system.

graphic: a single node (left) may be used as a desktop device, or installed in conjunction with other nodes that communicate with each other through a network (centre, right).

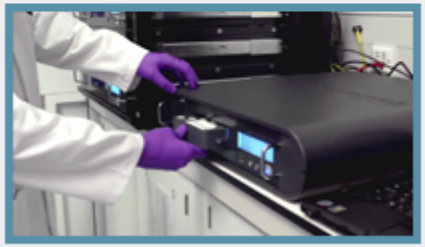
A single node as a desktop installation



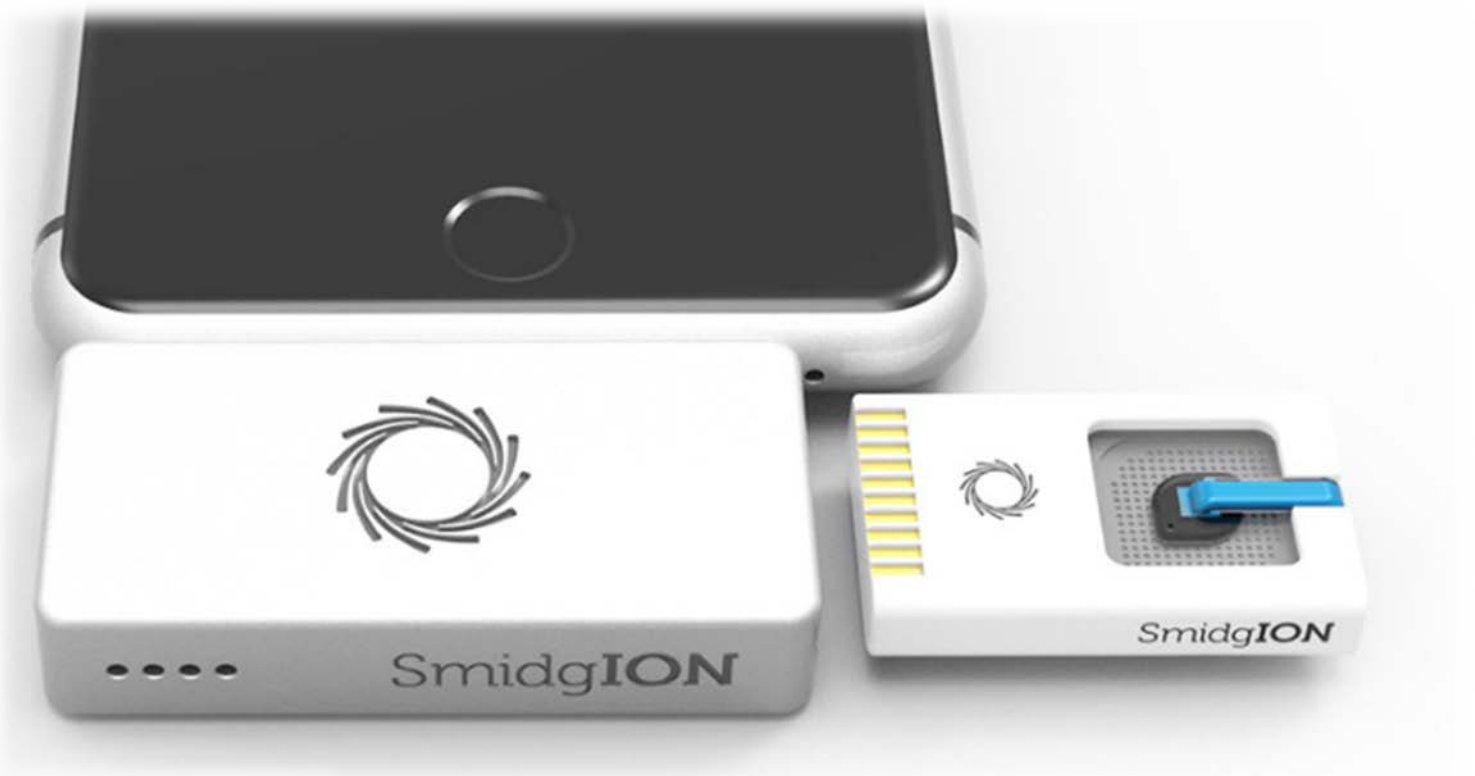
Nodes may be clustered through a network for larger throughput installations



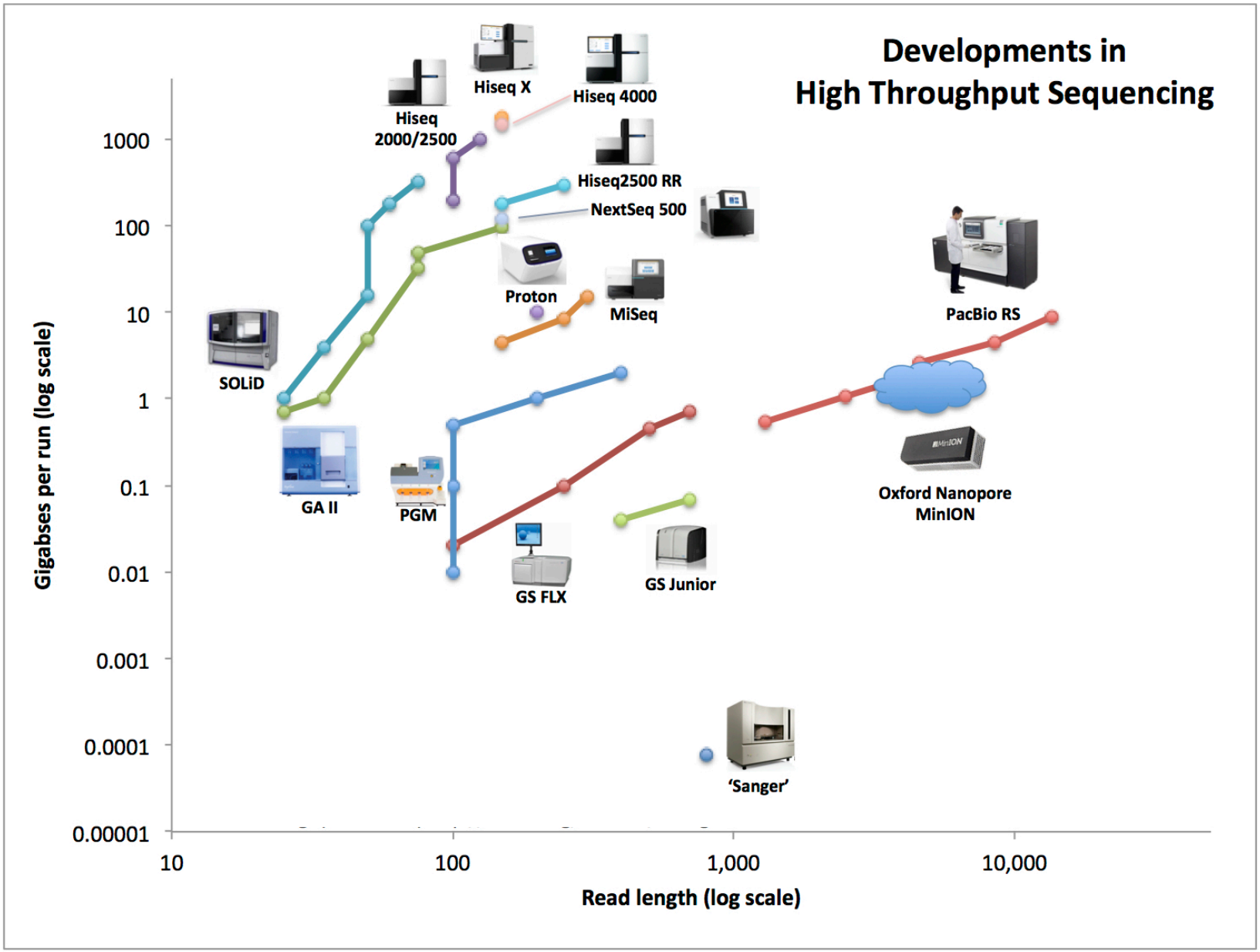
Each GridION node operates with a single dispoable cartridge.



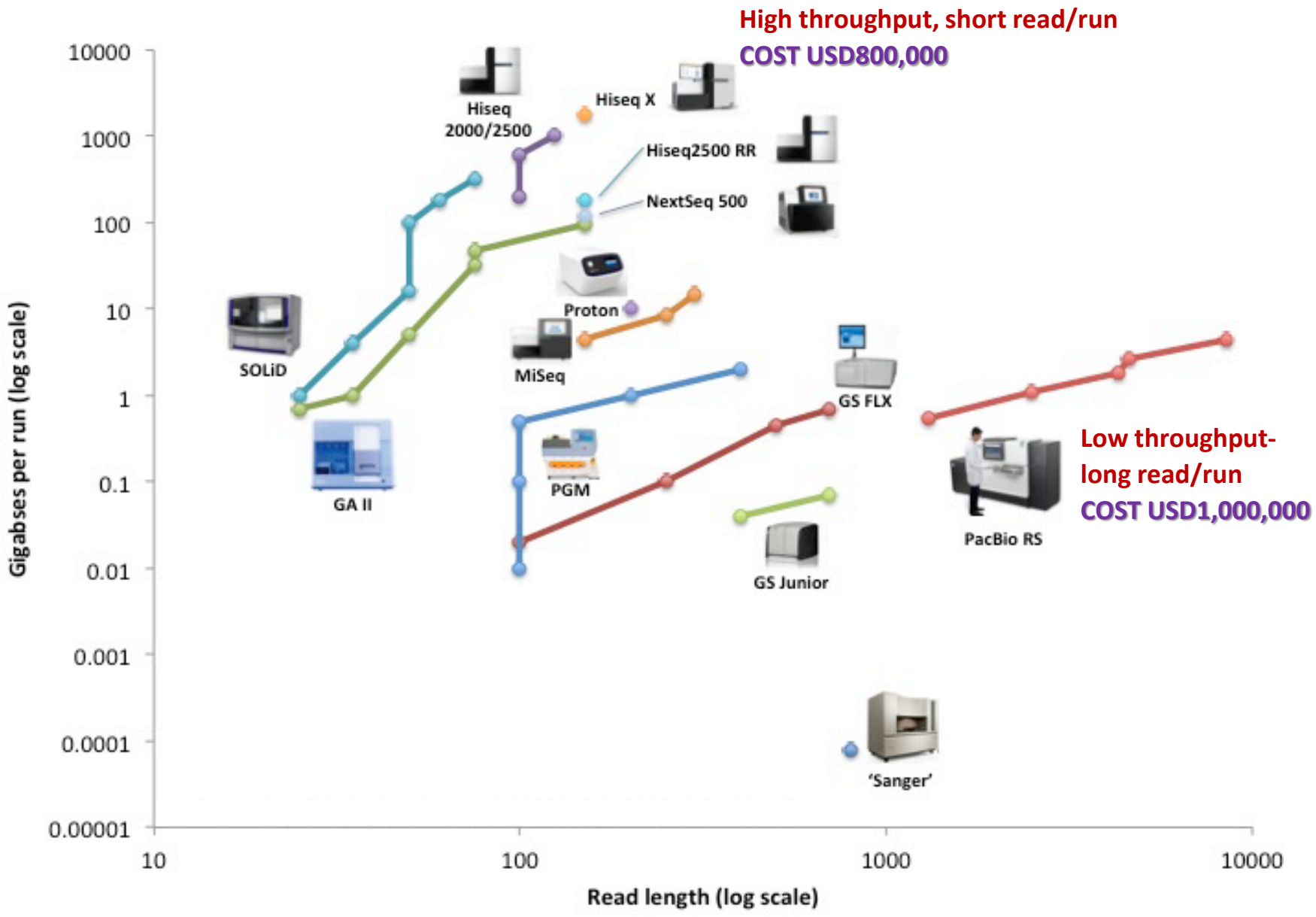
SmidgION: nanopore sensing for use with mobile devices



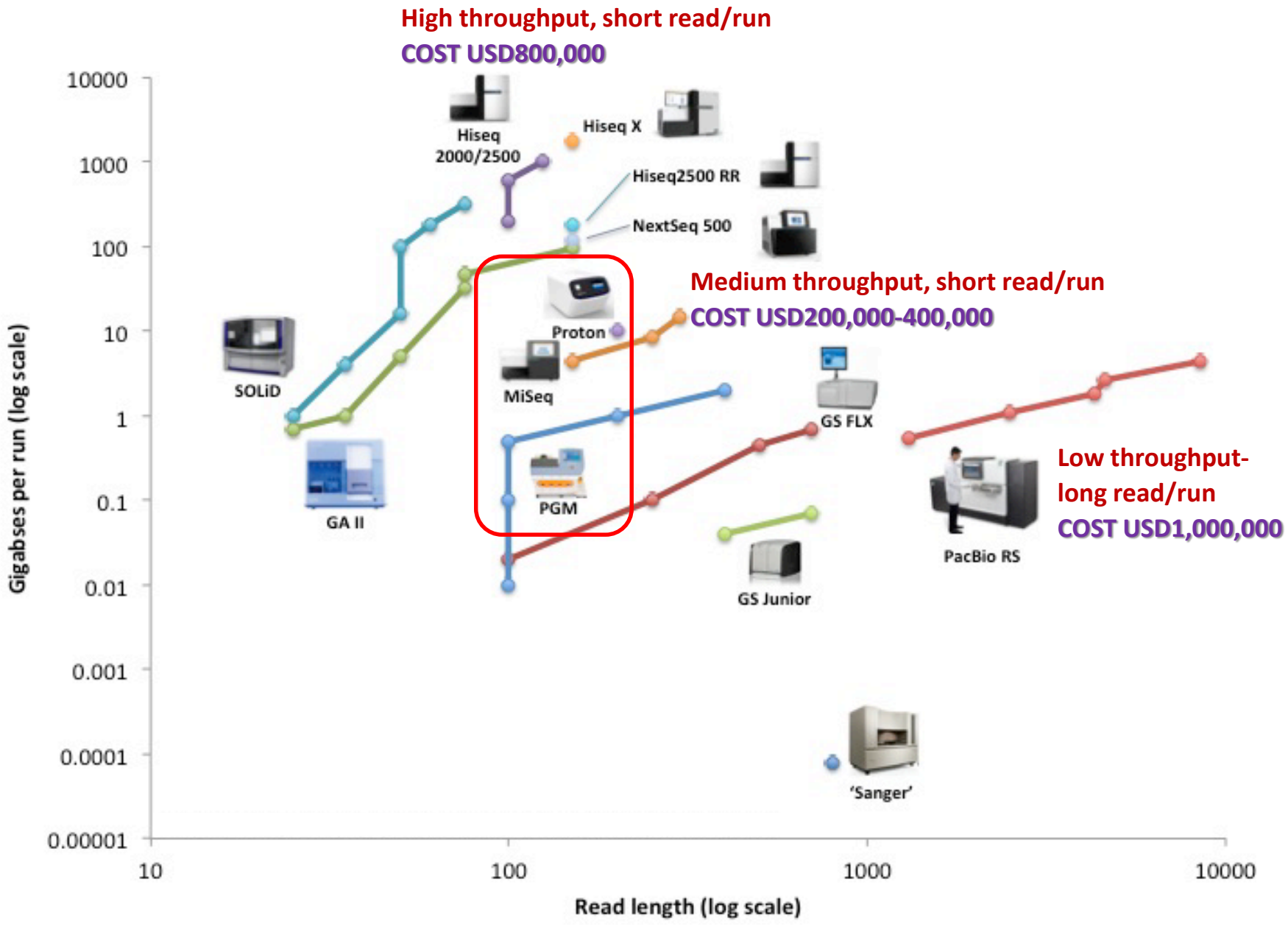



Developments in
High Throughput Sequencing
Hiseq X
Hiseq 4000
Hiseq
2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
PacBio RS
SOLiD
GA II
PGM
GS FLX
GS Junior
Oxford Nanopore
MinION
'Sanger'
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1,000
10,000
Read length (log scale)

Cost of NGS machines



Cost of NGS machines



Personal Sequencer

illumina®



Miseq/MiniSeq
Personal Sequencing System

ion torrent

by life technologies™



Ion Torrent/Ion Proton
Personal Genome Machine

illumina®

MiSeq Benchtop Sequencer





MISEQ REAGENT KIT V2

READ LENGTH	TOTAL TIME*	OUTPUT
1 × 36 bp	~4 hrs	540-610 Mb
2 × 25 bp	~5.5 hrs	750-850 Mb
2 × 150 bp	~24 hrs	4.5-5.1 Gb
2 × 250 bp	~39 hrs	7.5-8.5 Gb

MISEQ REAGENT KIT V3

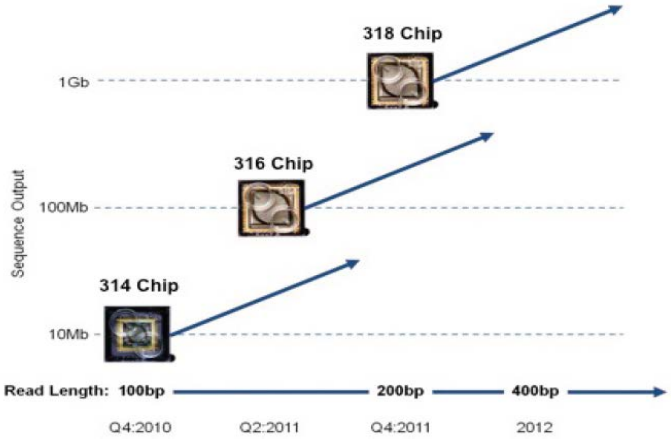
READ LENGTH	TOTAL TIME*	OUTPUT
2 × 75 bp	~24 hrs	3.3-3.8 Gb
2 × 300 bp	~65 hrs	13.2-15 Gb

ion torrent

by life technologies™



Chip Type	314	316	318
Wells (in millions)	1.2	6.2	11.1
Bases (in Mb)	>10	>100	>1000



10 Mb to 1 Gb

100-fold scaling means you can choose the amount of sequencing throughput required for your specific application.

ION PGM™ SEQUENCER

SMALL GENOMES
SETS OF GENES
GENE EXPRESSION
ChIP-SEQ



90-minute run times

Rapid 100-base sequencing on the Ion 314™ Chip.

35–400* bp reads

If your application requires short-read or long-read sequencing, or single-end, paired-end, or mate-paired sequencing, the Ion PGM™ Sequencer delivers the greatest flexibility.

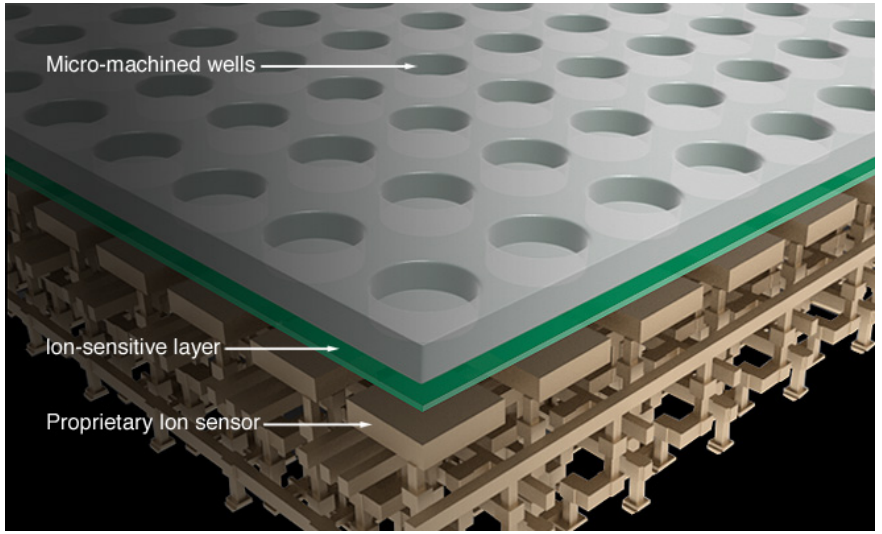

ion torrent
by life technologies™
ion torrent
by life technologies™
ion torrent
by life technologies™

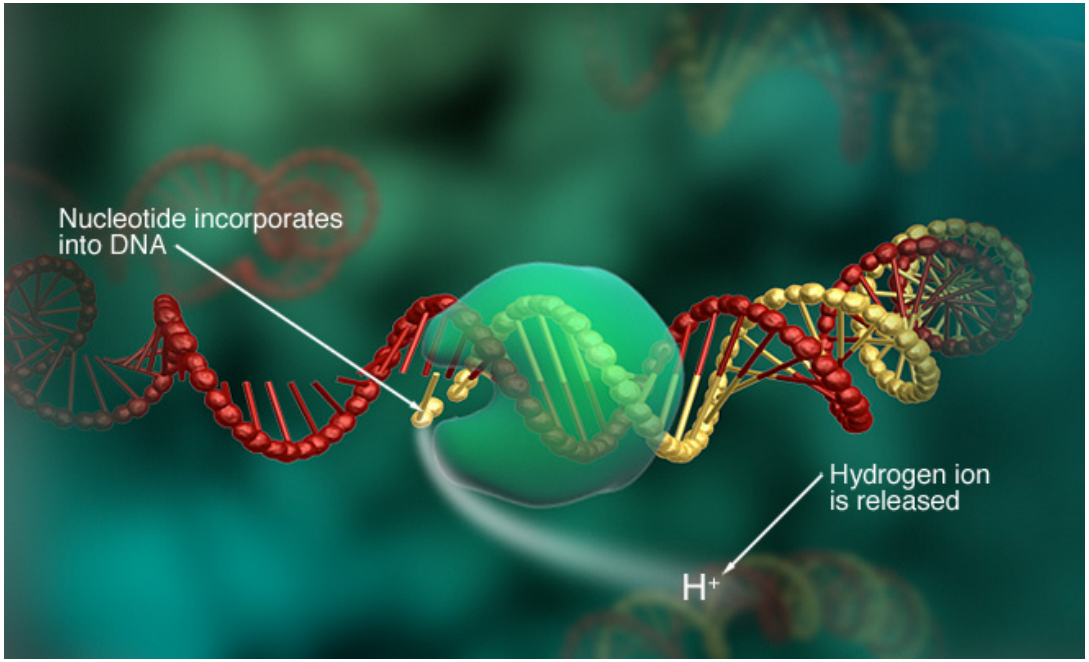
Ion Torrent





314 chip
316 chip



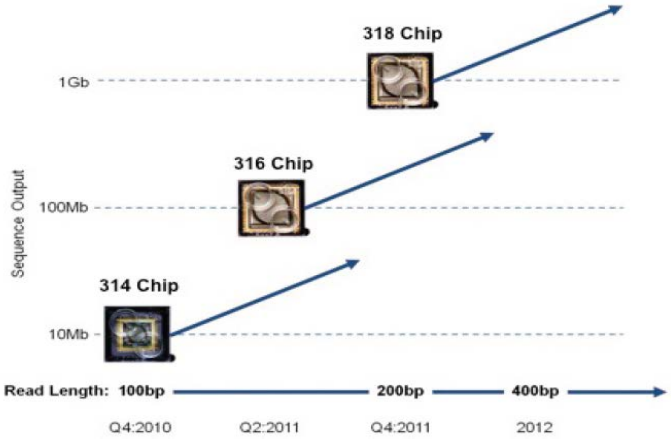


ion torrent

by life technologies™



Chip Type	314	316	318
Wells (in millions)	1.2	6.2	11.1
Bases (in Mb)	>10	>100	>1000



10 Mb to 1 Gb

100-fold scaling means you can choose the amount of sequencing throughput required for your specific application.

ION PGM™ SEQUENCER

SMALL GENOMES
SETS OF GENES
GENE EXPRESSION
ChIP-SEQ



90-minute run times

Rapid 100-base sequencing on the Ion 314™ Chip.

35–400* bp reads

If your application requires short-read or long-read sequencing, or single-end, paired-end, or mate-paired sequencing, the Ion PGM™ Sequencer delivers the greatest flexibility.

Ion Proton™ Sequencer (2012)

The Only Benchtop Genome Center

"Cost, speed, and accuracy are key elements in the use of DNA sequencing. The technological advances in the new Ion Proton™ instrument promise to be game-changing for both research and clinical applications."

DR. RICHARD LIFTON
YALE SCHOOL OF MEDICINE, USA

Human exome sequencing

Using the next generation of semiconductor technology, the Ion Proton™ I Chip will deliver whole-exome sequencing in just a few hours.

ION PROTON™ SEQUENCER

HUMAN GENOMES
HUMAN EXOMES
WHOLE TRANSCRIPTOMES

Human genome sequencing

The Ion Proton™ II Chip will enable fast, affordable, whole-genome sequencing on your benchtop.



2-hour run times

Rapid 100-base sequencing runs on the Ion Proton™ I Chip.

Ion Proton™ System performance specifications* with Ion Proton™ I Chip at commercial launch	
Throughput	Up to 10 Gb (Note: The Ion Proton™ II Chip will be available about six months after the Ion Proton™ I Chip. The Ion Proton™ II Chip will enable sample-to-variant analysis of a human genome in a single day, at up to 20x coverage.)
Read length	Up to 200-base fragment reads
Sequencing run time	As little as 2 hours for 100-base reads

illumina® Miniseq (2016)



Cluster Generation and Sequencing

MINISEQ SYSTEM HIGH-OUTPUT KIT*

READ LENGTH	TOTAL TIME†	OUTPUT
2 × 150 bp	~24 hrs	6.6-7.5 Gb
2 × 75 bp	13 hrs	3.3-3.75 Gb
1 × 75 bp	7 hrs	1.65-1.875 Gb

MINISEQ SYSTEM MID-OUTPUT KIT*

READ LENGTH	TOTAL TIME†	OUTPUT
2 × 150 bp	17 hrs	2.1-2.4 Gb

Reads Passing Filter

MINISEQ SYSTEM HIGH-OUTPUT KIT

Single Reads	22-25 Million
Paired-End Reads	44-50 Million

MINISEQ SYSTEM MID-OUTPUT KIT

Single Reads	7-8 Million
Paired-End Reads	14-16 Million

Quality Scores††

MINISEQ SYSTEM HIGH-OUTPUT KIT

> 80% bases higher than Q30 at 2 x 150 bp
> 85% bases higher than Q30 at 2 x 75 bp
> 85% bases higher than Q30 at 1 x 75 bp

MINISEQ SYSTEM MID-OUTPUT KIT

> 80% bases higher than Q30 at 2 x 150 bp

illumina®



MiniSeq System

MAX OUTPUT
8 Gb

MAX READ NUMBER
25 million

MAX READ LENGTH
2×150 bp



MiSeq Series

MAX OUTPUT
15 Gb

MAX READ NUMBER
25 million

MAX READ LENGTH
2×300 bp



NextSeq System

MAX OUTPUT
120 Gb

MAX READ NUMBER
400 million

MAX READ LENGTH
2×150 bp



HiSeq Series

MAX OUTPUT
1500 Gb

MAX READ NUMBER
5 billion

MAX READ LENGTH
2×150 bp



HiSeq X Series

MAX OUTPUT
1800 Gb

MAX READ NUMBER
6 billion

MAX READ LENGTH
2×150 bp

illumina®







Product	MiniSeq	MiSeq	NextSeq 500*		HiSeq 2500		HiSeq 3000	HiSeq 4000	HiSeq X Five†	HiSeq X Ten†
Description	Simplicity and accessibility for examining single genes or entire pathways	Speed and simplicity for targeted and small-genome sequencing	Speed and simplicity for everyday genomics		Power and efficiency for large-scale genomics		Maximum throughput and lowest cost for production-scale genomics		Maximum throughput and lowest cost for population- and production-scale human WGS	
Key methods	Targeted DNA and targeted RNA sequencing	Small genome, amplicon, targeted gene panel sequencing	Everyday genome, exome, transcriptome sequencing, and more		Production-scale genome, exome, transcriptome sequencing, and more				Population-scale human WGS	
Run mode	—	—	Mid-output	High-output	Rapid run	High-output	—	—	—	—
Flow cells processed per run	1	1	1	1	1 or 2	1 or 2	1	1 or 2	1 or 2	1 or 2
Output range	0.6–7.5 Gb	0.3–15 Gb	20–39 Gb	30–120 Gb	10–300 Gb	50–1000 Gb	125–750 Gb	125–1500 Gb	900–1800 Gb	900–1800 Gb
Run time	4–24 hours	5–55 hours	15–26 hours	12–30 hours	7–60 hours	< 1–6 days	< 1–3.5 days	< 1–3.5 days	< 3 days	< 3 days
Reads per flow cell‡	25 million§	25 million§	130 million	400 million	300 million	2 billion	2.5 billion	2.5 billion	3 billion	3 billion
Maximum read length	2 x 150 bp	2 x 300 bp	2 x 150 bp	2 x 150 bp	2 x 250 bp	2 x 125 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp

GeneReader



Early-access customers in 2013,
Commercial launch in early 2014.

20 flowcells configuration
20-400M reads per flow cell
12 minutes per base sequencing-by-synthesis
(50-bases ~10 hours)

Droplet-based Sequencing

Microfluidics, the science that deals with the behavior, precise control and manipulation of fluids that are geometrically constrained to a small, typically sub-millimeter scale, emerged in the beginning of the 1980s and is used widely in technologies such as inkjet print heads, DNA chips, lab-on-a-chip technology, micro-propulsion and micro-thermal technologies. GnuBIO has adapted technology developed in Dr. David Weitz laboratory at Harvard. This technology is combined with proprietary molecular biology assays to create a unique next-generation DNA sequencing technology. The science behind this GnuBIO sequencing technology utilizes droplet microfluidics to perform the biochemical reactions for sequencing inside of tiny picoliter -sized aqueous drops. Each droplet acts as a discrete reaction vessel, like a miniature test tube, where the sequencing assay is performed.



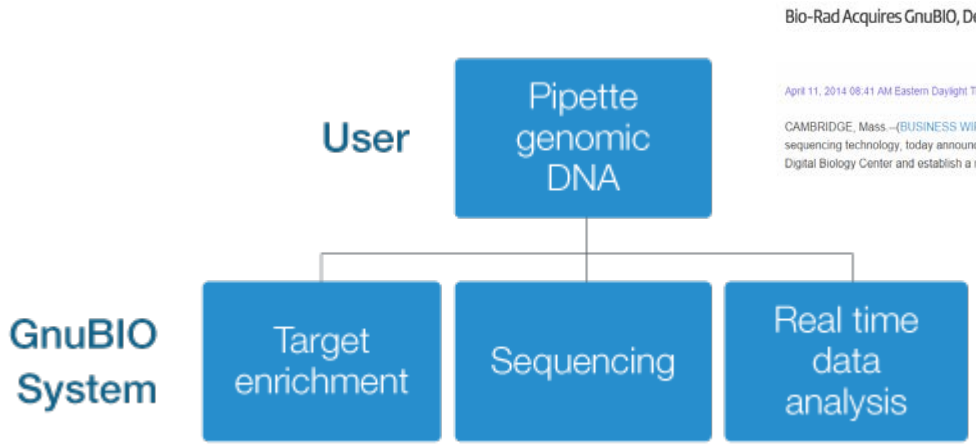
The following results are the raw and unfiltered data from the GnuBIO prototype system:

99.993% per base accuracy: This is the true per base accuracy, and is not based on consensus accuracy
100% sequence coverage
100% of the reads were used
Read lengths of 126 base pairs (read lengths of 1Kb can routinely be achieved today)

Droplet-based Sequencing

Microfluidics, the science that deals with the behavior, precise control and manipulation of fluids that are geometrically constrained to a small, typically sub-millimeter scale, emerged in the beginning of the 1980s and is used widely in technologies such as inkjet print heads, DNA chips, lab-on-a-chip technology, micro-propulsion and micro-thermal technologies. GnuBIO has adapted technology developed in Dr. David Weitz laboratory at Harvard. This technology is combined with proprietary molecular biology assays to create a unique next-generation DNA sequencing technology.

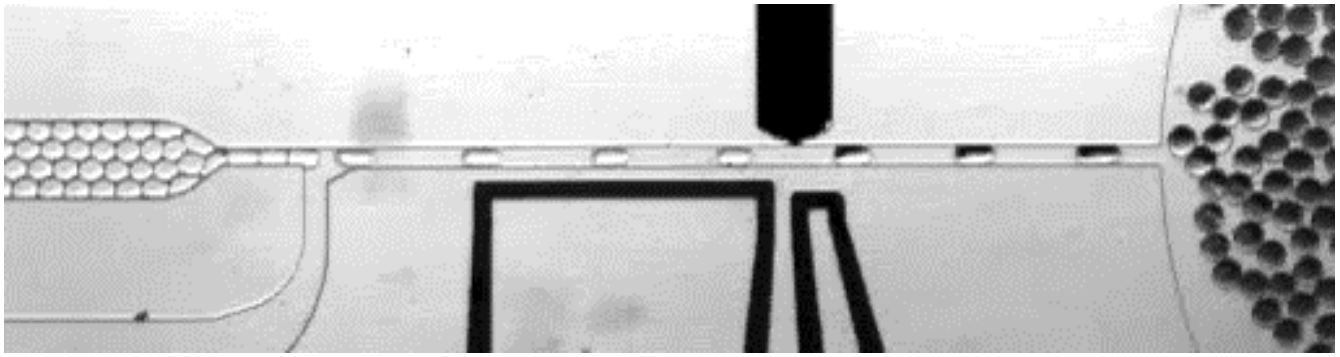
The science behind this GnuBIO sequencing technology utilizes droplet microfluidics to perform the biochemical reactions for sequencing inside of tiny picoliter -sized aqueous drops. Each droplet acts as a discrete reaction vessel, like a miniature test tube, where the sequencing assay is performed.



Bio-Rad Acquires GnuBIO, Developer of Droplet-Based DNA Sequencing Technology

April 11, 2014 08:41 AM Eastern Daylight Time

CAMBRIDGE, Mass.--(BUSINESS WIRE)--GnuBIO, a privately-held life sciences company that has developed a droplet-based DNA sequencing technology, today announced that it was acquired by Bio-Rad Laboratories, Inc. With this acquisition, Bio-Rad will expand its Digital Biology Center and establish a research and development site in Cambridge, Massachusetts.



Storage & Computing Power

Next gen sequencers generated Giga bp to Tera bp of data

Illumina HiSeq 2000	SOLiD 5500xl	454 FLX Titanium	Ion Torrent Personal Genome Machine	PacBio RS
100 bp reads	75bp x 35bp or 60bp x 60 bp	250-400 bp reads	200 bp reads	1000 bp reads
8 days	7 days	10 hours	2 hours	< 1 hour



TABLE 1. Average file sizes for various analyses

	Image Data Size†	Primary Analysis Data*	Secondary Analysis Data
1 slide - 1 tag (fragm			20 GB
1 slide - 2 tags (mate			40 GB
2 slides - 1 tags (frag			40 GB
2 slides - 2 tags (mat			80 GB

http://allseq.com/knowledgebank

AllSeq NEED SEQUENCING NGS CONFERENCES KNOWLEDGE BANK ABOUT US BLOG SIGN IN search...

Home ▸ Knowledge Bank

The AllSeq Knowledge Bank

The Knowledge Bank is managed by employees at AllSeq. AllSeq matches people's sequencing service needs with labs that have extra capacity on their NGS instruments. See how it works. This totally independent position in the sequencing market has lead to great knowledge about the topic.

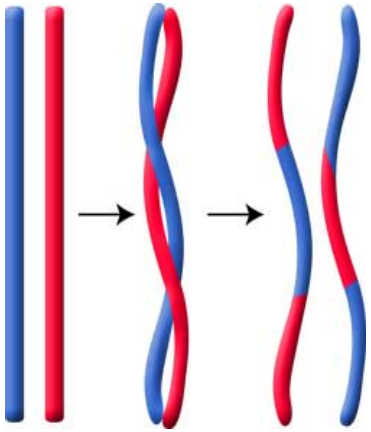
-The AllSeq-team

Knowledge Bank

Sequencing Platforms

Emerging Technologies

Electron Optica
Electronic BioSciences
GenapSys
Genia
GnuBIO
IBM
LaserGen
Lightspeed Genomics
NABsys
NobleGen Biosciences
Oxford Nanopore Technologies
Qiagen
xQuantuMDx
Reveo
Stratos Genomics
Two Pore Guys
ZS Genetics



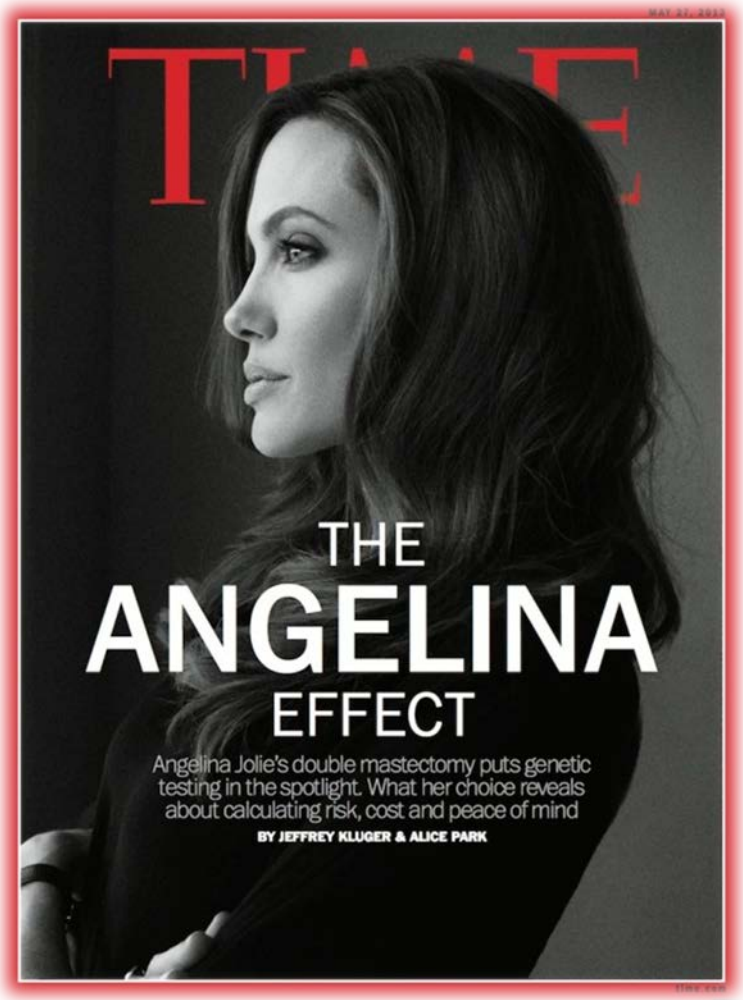
NEXT GENERATION SEQUENCING

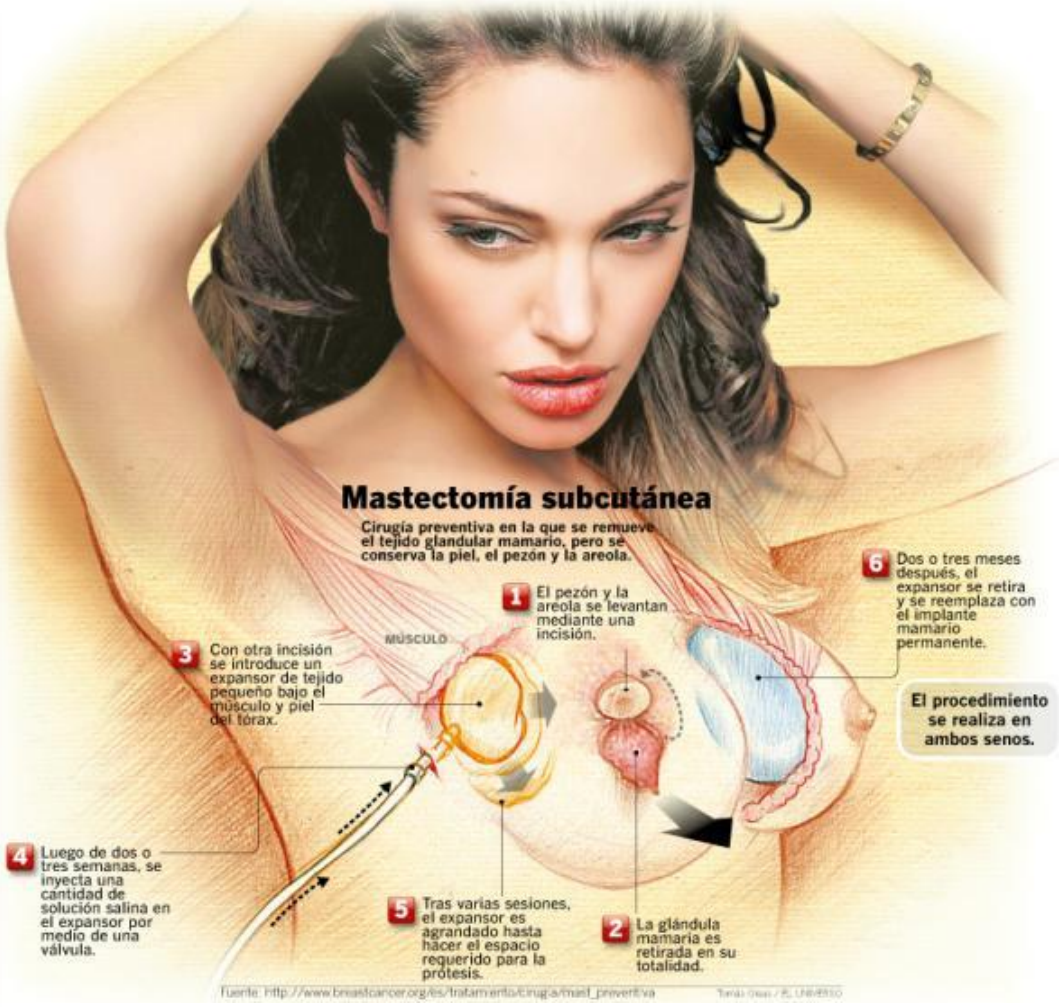
Applications & Analysis

Biomarker Discovery
Translational Bioinformatics
NGS
Biological Pathways
Evolution
Bioinformatics
Cancer Genome
Visualization
Disease
Personal Genome
CNVs
SNPs
Next-Gen Sequencing
Variants
ChipSeq
Metagenome
Epigenome
Network Medicine
Genomics
Pharmacogenomics
Network Biology
Semantic Biology
Genome-Phenome-Envirome
Medical Informatics
Drug Repositioning
RNA-Seq
Toxicogenomics
Rare Disease
Bio-data Mining



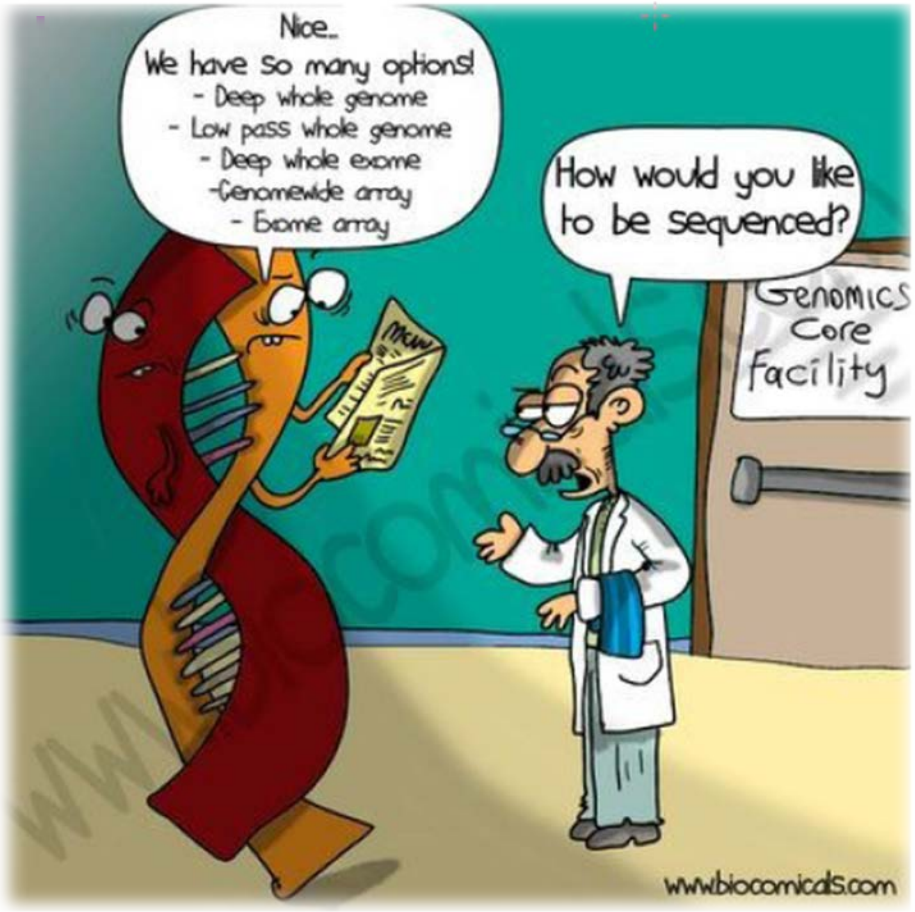
Applications on Biomedical Sciences





High-throughput Sequencing (HTseq) Platforms

- 454 Sequencing / Roche
  - GS Junior System
  - GS FLX+ System
- Illumina (Solexa)
  - HiSeq System
  - Genome analyzer IIx
  - MySeq
- Applied Biosystems - Life Technologies
  - SOLiD 5500 System
  - SOLiD 5500xl System
- Ion Torrent - Life Technologies
  - Personal Genome Machine (PGM)
  - Proton
- Helicos
  - Helicos Genetic Analysis System
- Pacific Biosciences
  - PacBio RS
- Oxford Nanopore Technologies
  - GridION System
  - MinION



Applications on Biomedical Sciences


ABI SOLiD 5500XL


Illumina Hiseq 2000


Roche 454 FLX


ABI Ion-Torrent


Illumina Miseq

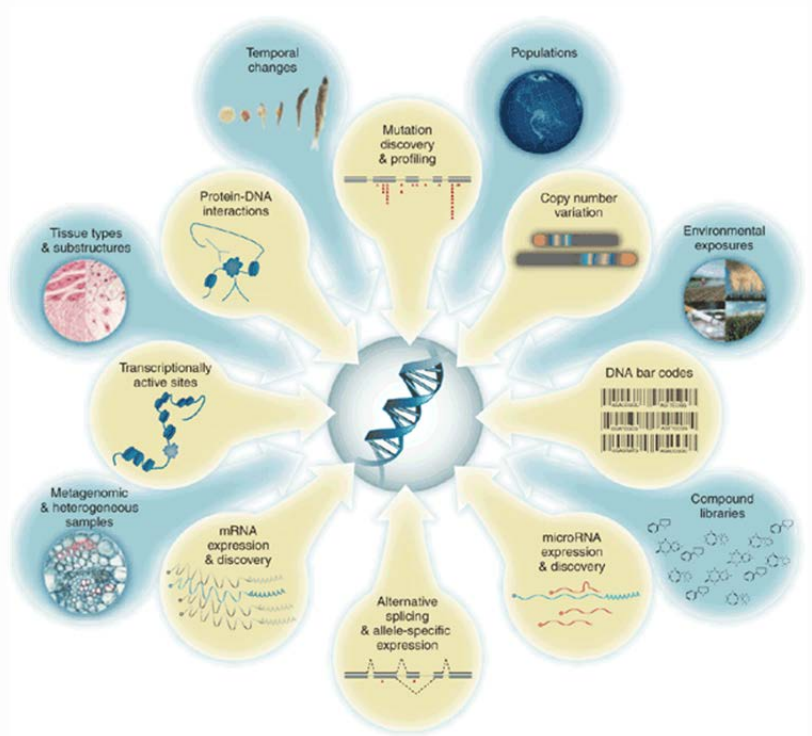

ABI 3730XL

DNA
Exome Sequencing
Genome Resequencing
De novo Genome Sequencing
Meta-Genomic Sequencing
Chip-Sequencing

RNA
Small RNA Sequencing
Transcriptome Sequencing
De novo Transcriptome Sequencing
Meta-Transcriptome Sequencing

BIOINFORMATICS


IBM X3850



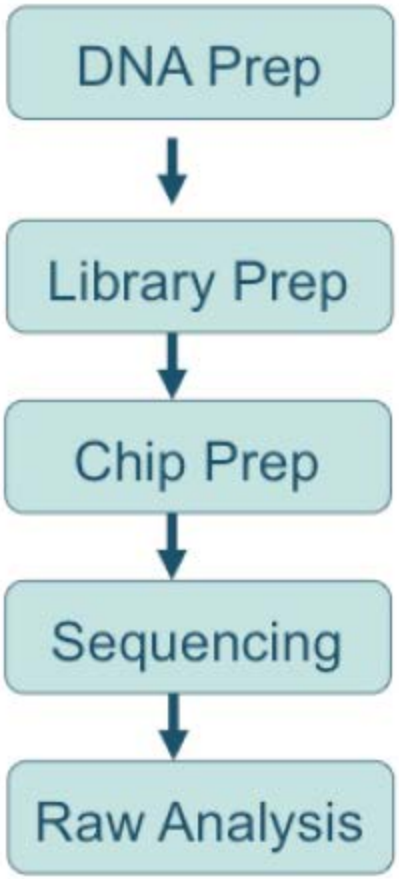
高通量定序分析

上課時間：每週五上午 08:10~12:00 （08:10~09:00－授課，09:10~12:00－實習）
上課地點：醫學院 9F 生物資訊中心 電腦教室

助教：蔡智宇(分機5690)

進度	課程進度	授課老師	講義下載 L:lecture T:tutorial E:exercise
1	New Generation Sequencing (NGS) Technologies & Biological Big Data	鄧致剛	
2	Databases and Tools for High Throughput Sequencing Analysis	葉元鳴	
3	Small RNA Analysis and Target Identification	黃柏榕	
4	Gene Expression Analysis for Non-model Organisms	鄧致剛	
5	Gene Expression Analysis	葉元鳴	
6	Transcriptome Functional Profiling - DAVID	朱俐潔	
7	Small Genome de novo Assembly & Annotation I	陳亭妏	
8	Small Genome de novo Assembly & Annotation II	陳亭妏	
9	MID-TERM EXAMINATION	鄧致剛	
10	Genome Annotation & Genome Browser	黃柏榕	
11	Chip Sequencing Analysis	陳亭妏	
12	Metagenome Analysis	葉元鳴	
13	Cancer Panel Sequencing Analysis	黃柏榕	
14	Exome Sequencing Analysis	黃柏榕	
15	Biological Big Data Visualization & Integration	李季青	
16	Galaxy	黃柏榕	
17	Student Presentation	鄧致剛	
18	Final-Examination	鄧致剛	

HTseq Experiment



Randomly shear DNA + end repair + size select

Append sequencing adapters

Layout of library on sequencing slide or wells

For each library fragment – determine the order and identity of bases at either end of the fragment



Image processing + base calling

- Base calls + associated quality: Fastq/BAM

Great… I got my data now what…

- Data and information management is slowly moving out of infancy in genomics science…. at the toddler stage…

- **The Good news**
  - Some data formats are being accepted widely

- **The Bad news**
  - Still many competing standards in some areas
  - Interoperability of data standards is almost non-existent
  - Governance is questionable

Data Format Types

- Raw Sequence Data e.g. fasta

```
>xyz some other comment
ttcctctttctcgactccatcttcgcggtagctgggaccgccgttcagtcgccaatatgc
agctctttgtccgcgcccaggagctacacaccttcgaggtgaccggccaggaaacggtcg
cccagatcaaggctcatgtagcctcactggagggcatt
```

- Aligned data e.g. BAM
  - SAM (Sequence Alignment/Map) format has become the *de facto* standard for storing alignment data.
  - BAM is a binary version of SAM allowing more efficient storage.

SAM format

```
ERR005646.11088674      147     1       161099954       60
54M     =       161099742       -266
TTTTCTGAACAGGGATGATATTTGTAATTTCATAGAATTAAGAGATATCTGACT
89=<;@>EECFCBBFFCAEFBGB=FFFC?@AB@G=FFB@CABABA?A@<>>=;=
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0
XO:i:0  XG:i:0  MD:Z:54 RG:Z:ERR005646  OQ:Z:D?
FFEEEFFFFFFFFFFEFFDFECFFFE;EEEEFCFFEEEEFEFECEEC=E;EF
ERR005646.5518024       147     1       161099956       60
54M     =       161099847       -163
TTCTGAACAGGGATGATATTTGTAATTTCATAGAATTAAGAGATATCTGACTCT  :
68=<A@@A???AB?A>ABBB>@CABCAAA>B@BAB@BA@A@A@A@=A=A=>;<
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0
XO:i:0  XG:i:0  MD:Z:54 RG:Z:ERR005646
OQ:Z:ECEEEEEEEEDEEEEE>EEEEEEEEEEEEE@EEEEEBEEEEEEEEECCCBEEEE
```

- Processed data e.g. BED

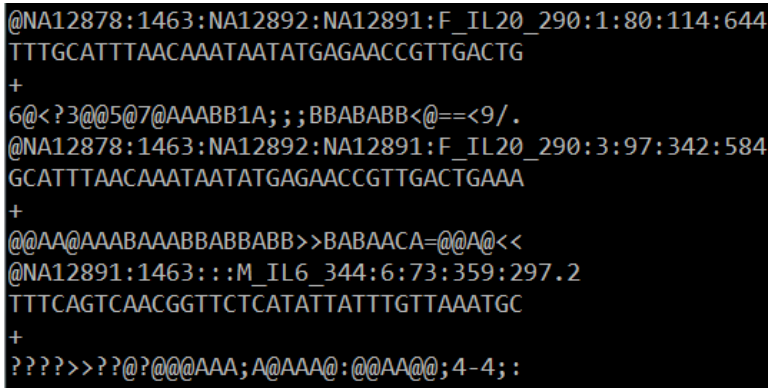
HTseq Raw Data Format

- fasta (Sanger)
- ~~csfasta (SOLiD)~~
- fastq (Solexa)
- sff (454)
- …. And about 30 other file formats

http://emboss.sourceforge.net/docs/themes/SequenceFormats.html

Fasta/Fastq

- FASTA
  - Header line “>”
  - Sequence
- FASTQ
  - Add QVs encoded as single byte ASCII codes
- Most aligners accept FASTA/Q as input
- Issue: data is volumous (2 bytes per base for FASTQ)
- Do PHRED scaled values provide the most information?



Fastq: Illumina

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

```
@HWI-E4_9_30WAF:1:1:8:308
TCCACATCAGAGGCCATGGCCACCAGGCCCAGGAT
+HWI-E4_9_30WAF:1:1:8:308
aaaaXaaabaa^aaLaaLLa^a^^VV\aaaaaaaa
@HWI-E4_9_30WAF:1:1:9:947
CCAATGTGGTCATAGGTGACAACCTTCTCCTCGCT
+HWI-E4_9_30WAF:1:1:9:947
aZaaaaaaZaab^aaaWaaaaaaaaaaaa\aaa
@HWI-E4_9_30WAF:1:1:9:1505
GGAAGCCAGGACCCACCATGAGTAGCATACATCTG
F:1:1:9:1505
```

Fastq: Illumina

```
@HWI-E4_9_30WAF:1:1:8:308
TCCACATCAGAGGCCATGGCCACCAGGCCCAGGAT
+HWI-E4_9_30WAF:1:1:8:308
aaaaXaaabaa^aaLaaLLa^a^^VV\aaaaaaaa
```

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

sff (text format): 454

```
>F7K88GK01BMPI0
Run Prefix: R_2009_12_18_15_27_42_
Region #: 1
XY Location: 0551_2346

Run Name: R_2009_12_18_15_27_42_FLX########_Administrator_yourrunname
Analysis Name: D_2009_12_19_01_11_43_XX_fullProcessing
Full Path: /data/R_2009_12_18_15_27_42_FLX########_Administrator_yourrunname/
D_2009_12_19_01_11_43_XX_fullProcessing/

Read Header Len: 32
Name Length: 14
# of Bases: 500
Clip Qual Left: 15
Clip Qual Right: 490
Clip Adap Left: 0
Clip Adap Right: 0

Flowgram: 1.03 0.00 1.01 0.02 0.00 0.96 0.00 1.00 0.00 1.04 0.00 0.00 0.97 0.00 0.96 0.02
0.00 1.04 0.01 1.04 0.00 0.97 0.96 0.02 0.00 1.00 0.95 1.04 0.00 0.00 2.04 0.02 0.03 1.05
Flow Indexes: 1 3 6 8 10 13 15 18 20 22 23 26 27 28 31 31 34 35 37 37 37 40 43 45 47 47 47 50
53 53 53 55 58 60 63 66 67 67 67 67 70 71 71 74 74 76 79 82 83 86 86 88 88 91 93 96 97...
Bases:
tcagatcagacacgCCACTTTGCTCCCATTTCAGCACCCCACCAAGCACAAGGCTGTCATCCCAATTGGACGGACAGATATGAGGTTAGCATT
GGAAACCAATTCAGTCCCTAATTATTCACGACTGAACCCAGCGACAATTGGACATGGATTCATTTTTCAACTTGATTTGTTGTTGTAAAAGCA
...
Quality Scores: 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 38 38 38 40 40 40 39 39
```

454 fasta with quality file

Fasta:

```
>FTJD6BE02HHD3W length=409 xy=2951_1562 region=2 run=R_2009_04_01_11_28_49_
AGAAAGTCAGCGGCAAATTTGGTTTTAGACGAATTGTCCCTTTGACATAACGACTAAAGG
AGTCAACAGATTTTCGTATAACTTCGTATAATGTATGCTATACGAAGTTATTACGCTATT
...
```

Qual:

```
>FTJD6BE02HHD3W length=409 xy=2951_1562 region=2 run=R_2009_04_01_11_28_49_
40 40 39 39 39 40 40 40 40 40 40 40 40 38 31 26 26 16 16 16 20 20
14 14 14 14 27 33 32 35 36 33 36 35 36 38 35 20 20 21 24 24 22 36
39 40 38 38 38 40 40 40 40 40 40 37 37 37 33 33
29 36 38 38 38 38 38 38 38 35 20 21 21 21 31 36 37 40 40 35 37 37
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
...
```

Sanger-style Phred scores $Q_{\text{sanger}} = -10 \log_{10} p$

454 base quality?

Fasta:

```
>FTJD6BE02HHD3W length=409 xy=2951_1562 region=2 run=R_2009_04_01_11_28_49_
AGAAAGTCAGCGGCAAATTTGGTTTTAGACGAATTGTCCCTTTGACATAACGACTAAAGG
AGTCAACAGATTTTCGTATAACTTCGTATAATGTATGCTATACGAAGTTATTACGCTATT
...
```

Qual:

chance of being wrong: $1:10^{4.0} = 1:10000$

```
>FTJD6BE02HHD3W length=409 xy=2951_1562 region=2 run=R_2009_04_01_11_28_49_
40 40 39 39 39 40 40 40 40 40 40 40 40 38 31 26 26 16 16 16 20 20
14 14 14 14 27 33 32 35 36 33 36 35 36 38 35 20 20 21 24 24 22 36
39 40 38 38 38 40 40 40 40 40 40 37 37 37 33 33
29 36 38 38 38 38 38 38 38 35 20 21 21 21 31 36 37 40 40 35 37 37
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
...
```

chance of being wrong: $1:10^{3.5} = 1:3162$

Sanger-style Phred scores

All Platforms have Errors



Illumina



SoLID/ABI-Life



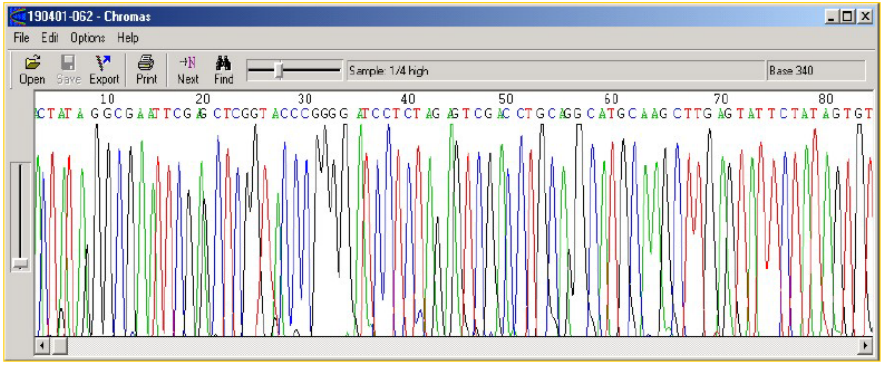
Roche 454



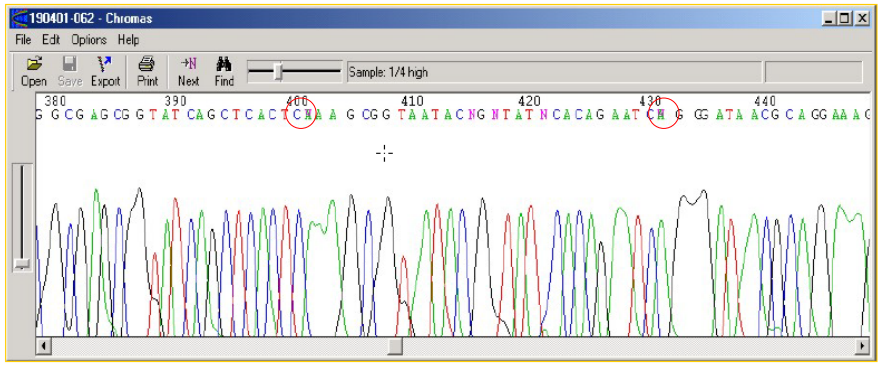
Ion Torrent

1. Removal of low quality bases/ Low complexity regions
2. Removal of adaptor sequences
3. Homopolymer-associated base call errors (3 or more identical DNA bases) causes higher number of (artificial) frameshifts

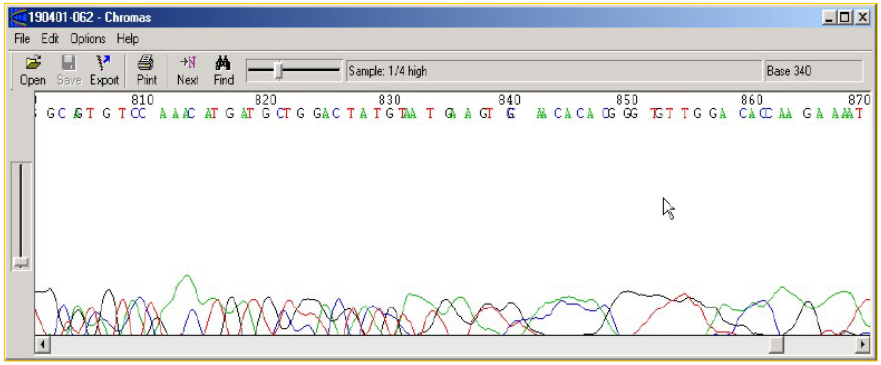
Trace File



High quality region - NO ambiguities (Ns)



Medium quality region - SOME ambiguities (Ns)



Poor quality region - LOW confidence

Accessing Quality: phred scores

Phred quality scores were originally developed by the program Phred to help in the automation of DNA sequencing in the Human Genome Project. Phred quality scores are assigned to each base call in automated sequencer traces.[1][2] Phred quality scores have become widely accepted to characterize the quality of DNA sequences, and can be used to compare the efficacy of different sequencing methods. Perhaps the most important use of Phred quality scores is the automatic determination of accurate, quality-based consensus sequences.

http://en.wikipedia.org/wiki/Phred_quality_score

$$Q = -10\log_{10} P$$

P=error probability of a given base call

Genome Research **8**: 175-185, 1998

Base-Calling of Automated Sequencer Traces Using *Phred*. I. Accuracy Assessment

Brent Ewing,[1] LaDeana Hillier,[2] Michael C. Wendl,[2] and Phil Green[1,3]

[1]Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA;
[2]Genome Sequencing Center, Washington University School of Medicine, Saint Louis, Missouri 63108 USA

Genome Research **8**: 186-194, 1998

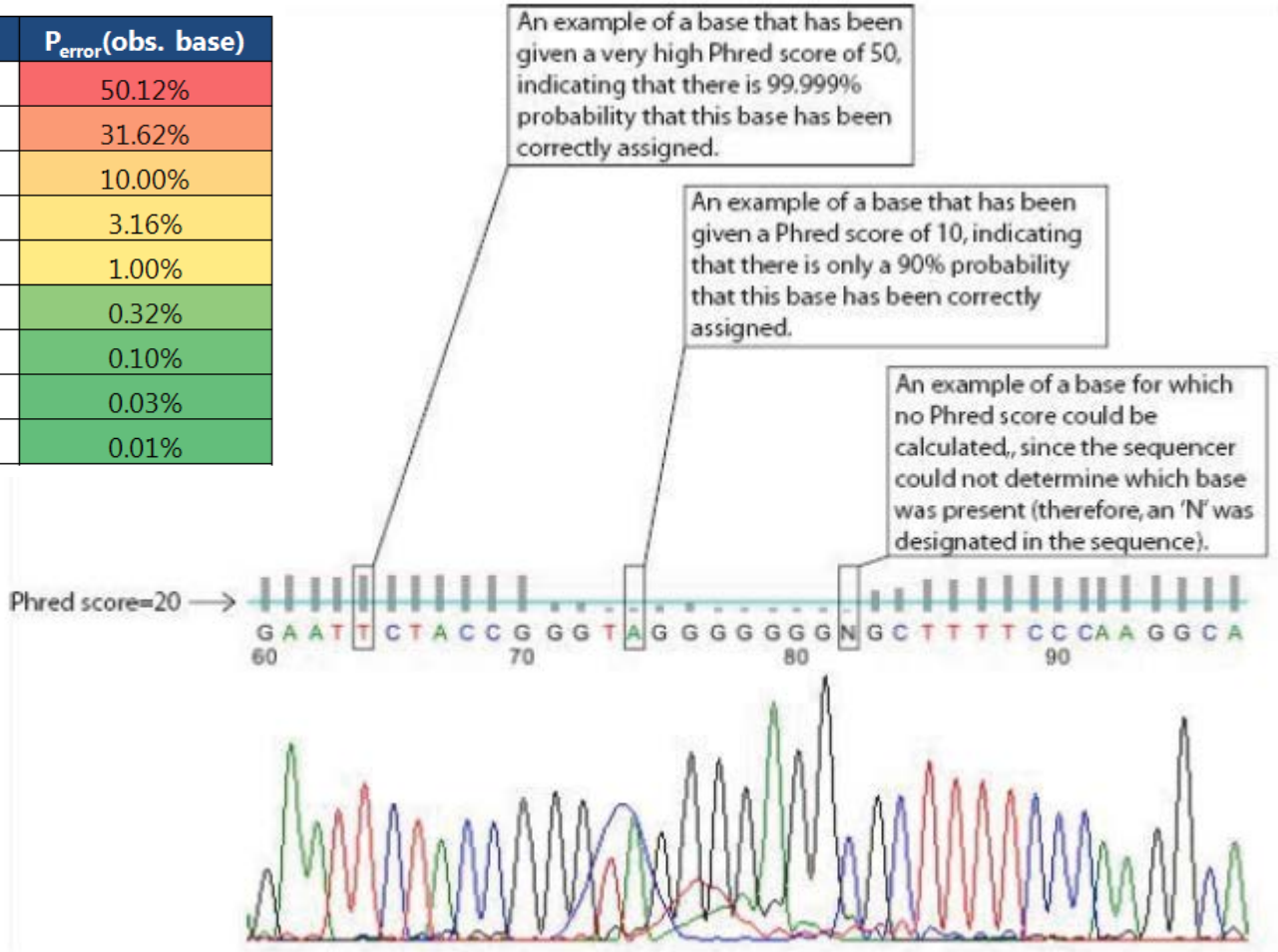
Base-Calling of Automated Sequencer Traces Using *Phred*. II. Error Probabilities

Brent Ewing and Phil Green[1]

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA

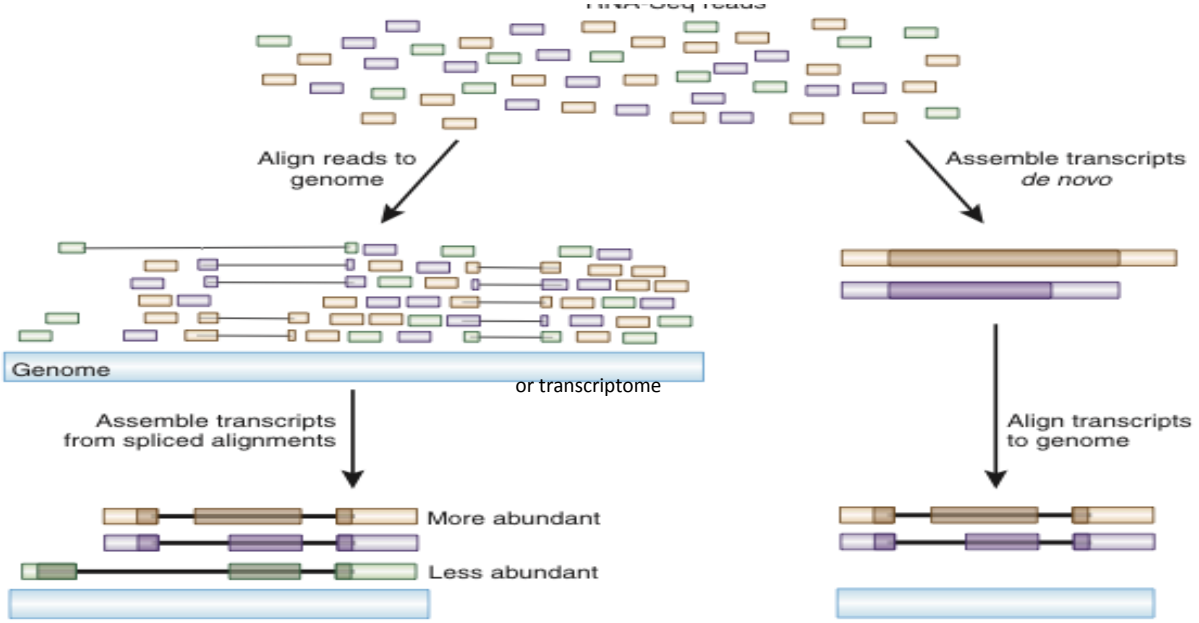
Accessing Quality: phred scores

Base Quality	P_{error}(obs. base)
3	50.12%
5	31.62%
10	10.00%
15	3.16%
20	1.00%
25	0.32%
30	0.10%
35	0.03%
40	0.01%



http://en.wikipedia.org/wiki/Phred_quality_score

Analysis Strategies: Reference Sequence *Alignment* (Mapping) *vs* De novo *Assembly*



Process	Software & Algorithms	Website
Preprocessing step	homemade script	(N/A)
(1.1) Alignment	MAQ	http://maq.sourceforge.net
	BWA	http://bio-bwa.sourceforge.net/bwa.shtml
	BWA-SW (SE only)	http://bio-bwa.sourceforge.net/bwa.shtml
	PERM	http://code.google.com/p/perm/
	BOWTIE	http://bowtie-bio.sourceforge.net
	SOAPv2	http://soap.genomics.org.cn
	MOSAIK	http://bioinformatics.bc.edu/marthlab/Mosaik
	NOVOALIGN	http://www.novocraft.com/
(1.2) *De novo* Assembly	VELVET	http://www.ebi.ac.uk/%7Ezerbino/velvet
	SOAPdenovo	http://soap.genomics.org.cn
	ABYSS	http://www.bcgsc.ca/platform/bioinfo/software/abyss

SRA: Sequence Read Archive

Collection of sequence data from next-generation sequencing technology for different organisms

http://www.ncbi.nlm.nih.gov/sra/ and http://www.ncbi.nlm.nih.gov/Traces/sra/

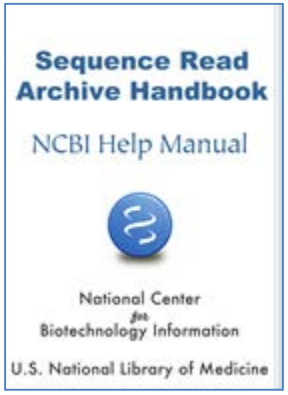
National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

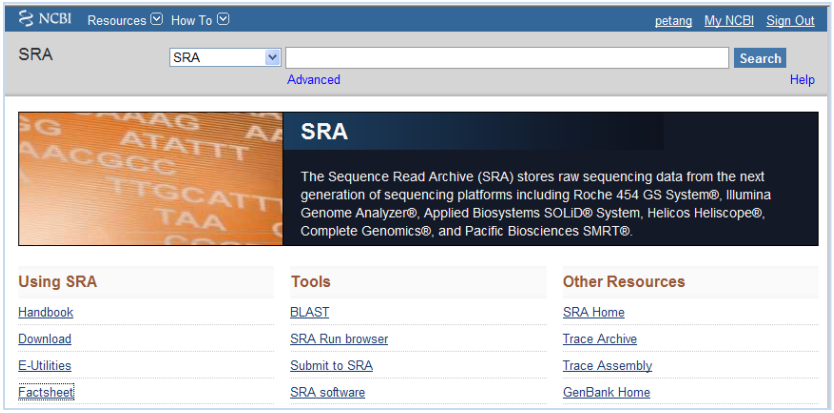
Scope and access

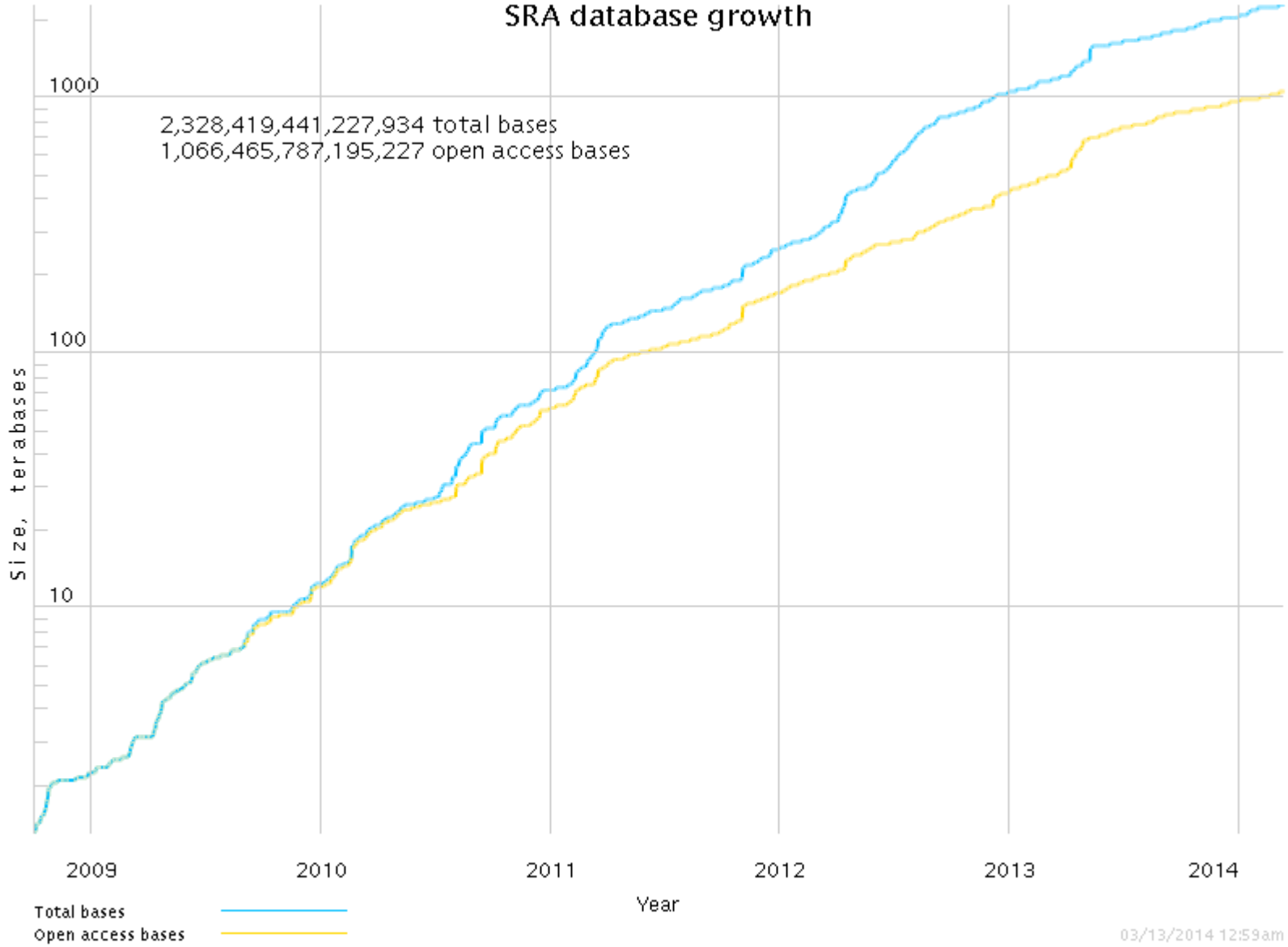
Sequence Read Archive (SRA) is the NCBI database which stores sequence data obtained from next generation sequence (NGS) technology. Through this database, the metadata for those sequences can be queried to locate the sequence dataset for subsequent download and further analysis. Specifically, SRA:

- Archives raw oversampling NGS data for various genomes from several platforms
- Shares NGS data with EMBL and DDBJ
- Serves as a starting point for "secondary analysis"
- Provides access to data from human clinical samples to authorized users who agree to the dataset's privacy and usage mandates

Metadata from SRA can be queried from Entrez SRA page and the SRA project list and sequence data can be browsed, searched and downloaded from its homepage at www.ncbi.nlm.nih.gov/sra/ and www.ncbi.nlm.nih.gov/Traces/sra/, respectively. For sequence-based searches against certain subsets of SRA reads (long reads from the Roche 454 platform) using BLAST, a link is listed under the "Specialized BLAST" section on the BLAST homepage: blast.ncbi.nlm.nih.gov/ .






SRA database growth
2,328,419,441,227,934 total bases
1,066,465,787,195,227 open access bases
1000
100
10
Size, terabases
2009
2010
2011
2012
2013
2014
Year
Total bases
Open access bases
03/13/2014 12:59am

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Display Settings: Abstract

Send to:

Proc Natl Acad Sci U S A. 2011 Jan 25;108(4):1513-8. doi: 10.1073/pnas.1017351108. Epub 2010 Dec 27.

High-quality draft assemblies of mammalian genomes from massively parallel sequence data.

Gnerre S[1], Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB.

Author information

Abstract

Massively parallel DNA sequencing technologies are revolutionizing genomics by making it possible to generate billions of relatively short (~100-base) sequence reads at very low cost. Whereas such data can be readily used for a wide range of biomedical applications, it has proven difficult to use them to generate high-quality de novo genome assemblies of large, repeat-rich vertebrate genomes. To date, the genome assemblies generated from such data have fallen far short of those obtained with the older (but much more expensive) capillary-based sequencing approach. Here, we report the development of an algorithm for genome assembly, ALLPATHS-LG, and its application to massively parallel DNA sequence data from the human and mouse genomes, generated on the Illumina platform. The resulting draft genome assemblies have good accuracy, short-range contiguity, long-range connectivity, and coverage of the genome. In particular, the base accuracy is high (≥99.95%) and the scaffold sizes (N50 size = 11.5 Mb for human and 7.2 Mb for mouse) approach those obtained with capillary-based sequencing. The combination of improved sequencing technology and improved computational methods should now make it possible to increase dramatically the de novo sequencing of large genomes. The ALLPATHS-LG program is available at http://www.broadinstitute.org/science/programs/genome-biology/crd.

PMID: 21187386 [PubMed - indexed for MEDLINE] PMCID: PMC3029755 Free PMC Article

Related information

- Related Citations
- Assembly
- BioSample
- MedGen
- Nucleotide
- References for this PMC Article
- Related Project
- SRA
- Free in PMC
- Cited in PMC

SRA

SRA

Advanced

Display Settings: Summary, 20 per page

Results: 1 to 20 of 319

1. DNA replication time of the human genome S phase Replicate 1
   1 ILLUMINA (Illumina HiSeq 2000) run: 104.7M spots, 20.9G bases, 12.8Gb downloads
   Accession: SRX352985
2. NA12878_S
   1 ILLUMINA (Illumina HiSeq 2000) run: 23M spots, 4.6G bases, 3.2Gb downloads
   Accession: SRX352183
3. NA12878_G1
   1 ILLUMINA (Illumina HiSeq 2000) run: 13.2M spots, 2.6G bases, 1.8Gb downloads
   Accession: SRX352182
4. Illumina Platinum Genomes NA12878
   1 ILLUMINA (Illumina HiSeq 2000) run: 791.4M spots, 159.4G bases, 69.4Gb downloads
   Accession: SRX335485
5. NA12878.all.ILLUMINA.bwa.CEU.high_coverage.20120522-reload
   1 ILLUMINA (Illumina HiSeq 2000) run: 1.4G spots, 290.2G bases, 148.8Gb downloads
   Accession: SRX333017

SRX352985: **DNA replication time of the human genome S phase Replicate 1**
1 ILLUMINA (Illumina HiSeq 2000) run: 104.7M spots, 20.9G bases, 12.8Gb downloads

Accession: SRX352985
Experiment design: n/a
Submitted by: Harvard Medical School
Study summary: SRP013016 • Differential and sex-specific relationship of DNA replication timing to different forms of human mutation and variation • SRP013016 • All experiments (less...)

Abstract: Understanding the factors that affect the generation of mutations is important in medical, cancer and evolutionary genetics. DNA replication timing is known to affect the distribution of mutations across the human genome. We generated a high resolution map of the human genome's replication timing program and analyzed its relationship to different forms of human mutation and genome variation. We show that the replication program profoundly shapes the distribution of substitution mutations, with a stronger influence on transversion relative to transition mutations. In addition, we find distinct relationships between replication timing and different kinds of structural mutations: mutations mediated by non-allelic homologous recombination are enriched in the early replicating parts of the genome, while structural mutations arising from other mechanisms are enriched in late replicating regions. An early-replication bias is also observed for meiotic recombination hotspots. Both recombination and structural mutations show gender differences in their relationships to DNA replication timing, with a stronger early-replication bias observed in females. Together, our results indicate a differential and sex-specific relationship of DNA replication timing to the distribution of different forms of mutation and variation across the human genome.

Center Project:

Sample: SAMN00801888 • GM12878 female unaffected International hapmap project - ceph (plate i) utah residents with ancestry from northern and western europe sample (less...)

Organism: Homo sapiens
Attributes:
cell line: GM12878
dna id: NA12878
sex: Female
cell type: B-Lymphocyte
family id: 1463
family role: mother
race: Caucasian
ethnicity: UTAH/MORMON
collection: NIGMS Human Genetic Cell Repository
tag: Cytochrome p450, subfamily iic, polypeptide 19
tag: Cyp2c19
tag: International hapmap project - ceph utah residents with ancestry from northern and western europe
tag: International hapmap project - ceph (plate i) utah residents with ancestry from northern and western europe
tag: Ceph/utah pedigree 1463

Library: Replicate 1 (less...)
Strategy: WGS
Source: GENOMIC
Selection: size fractionation
Layout: PAIRED

Platform: Illumina (less...)
Instrument model: Illumina HiSeq 2000

Spot descriptor:



Total: 1 run, 104.7M spots, 20.9G bases, 12.8Gb

#	Run	# of Spots	# of Bases	Size
1.	SRR496477	104,650,130	20.9G	12.8Gb

ID: 501159

http://seqanswers.com/

SEQanswers

the next generation sequencing community

BIOO SCIENTIFIC
AIR™ Genomic DNA Sequencing Kit
AIR™ Paired-End DNA Sequencing Kit
AIR™ DNA Fragmentation Kit
Discover More

SEQanswers

Welcome, petang.
You last visited: 02-28-2011 at 04:25 PM

User CP | FAQ | Community | Calendar | New Posts | Search | Quick Links | Log Out

Hey there petang, looks like you have not posted on our forums in several weeks, why not take a few moments to ask or answer a question in one of our forums?



Forum	Last Post	Threads	Posts
Introductions New here? Stop in and introduce yourself. Where you are, what you work on, etc.	A message for K8MCre! by fkrueger Today 04:45 AM	496	1,076
General (2 Viewing) Any topic/question that does not fit into the subcategories below. If you're unsure of where to put something, ask in here!	Ensembl human genome naming... by Bert Yesterday 10:32 PM	335	1,379
Core Facilities Dealing with customer samples, data, and the challenges that come with both?	Equpment Cost in setting up... by BTS Yesterday 04:03 AM	10	48
Literature Watch (2 Viewing) Discussion of any scientific study related to high content or next generation genomics. Whole genome association, metagenomics, digital gene expression, etc.	Sequencing all for the... by Joann Today 02:03 AM	847	1,045
Events / Conferences (1 Viewing) Relevant events or scientific conferences that may be of interest to the community. (Mgmt reserves right to limit commercial content without approval.)	Workshop on Molecular... by shandley Today 04:23 AM	116	340