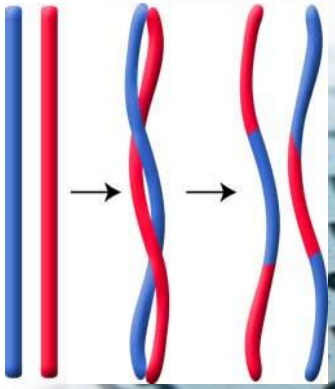
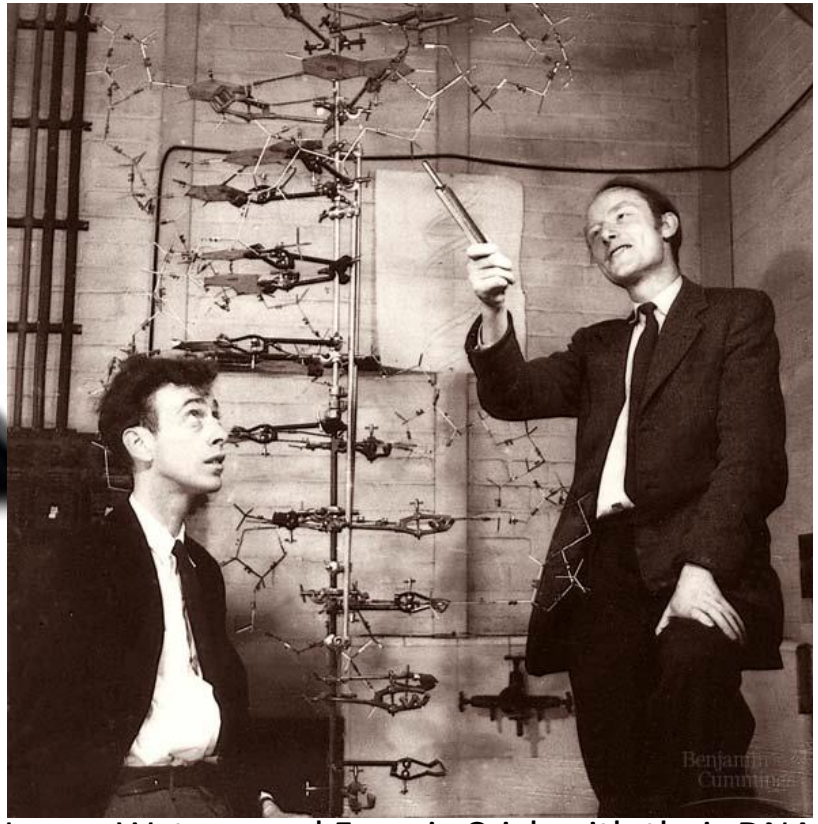


NEXT GENERATION SEQUENCING
Technologies and Applications

Hsuan Liu, Ph.D.(劉軒)
E-Mail: liu-hsuan@mail.cgu.edu.tw

60 DNA 1953 - 2013

Nature No.4356 Apr 25, 1953,



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

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

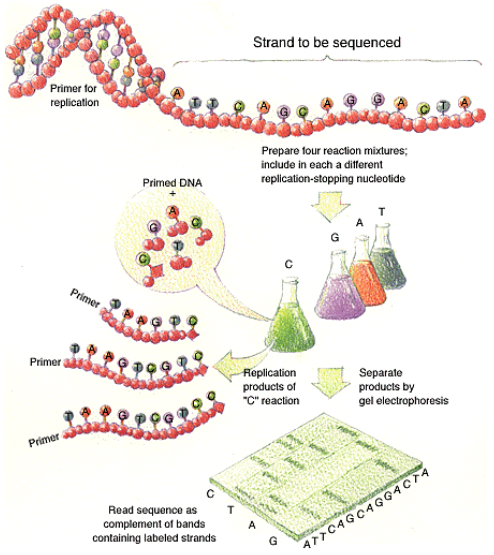
Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

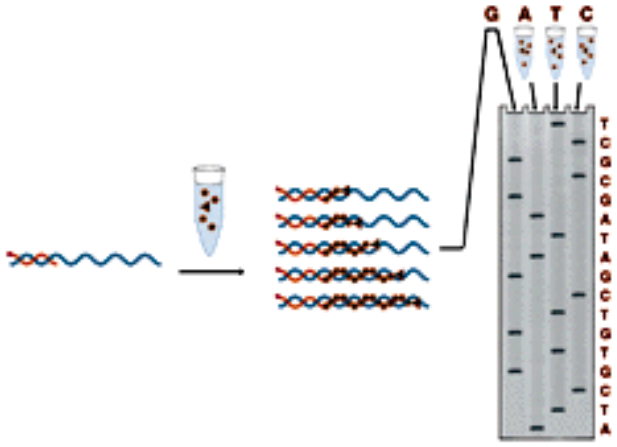
DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage φX174)

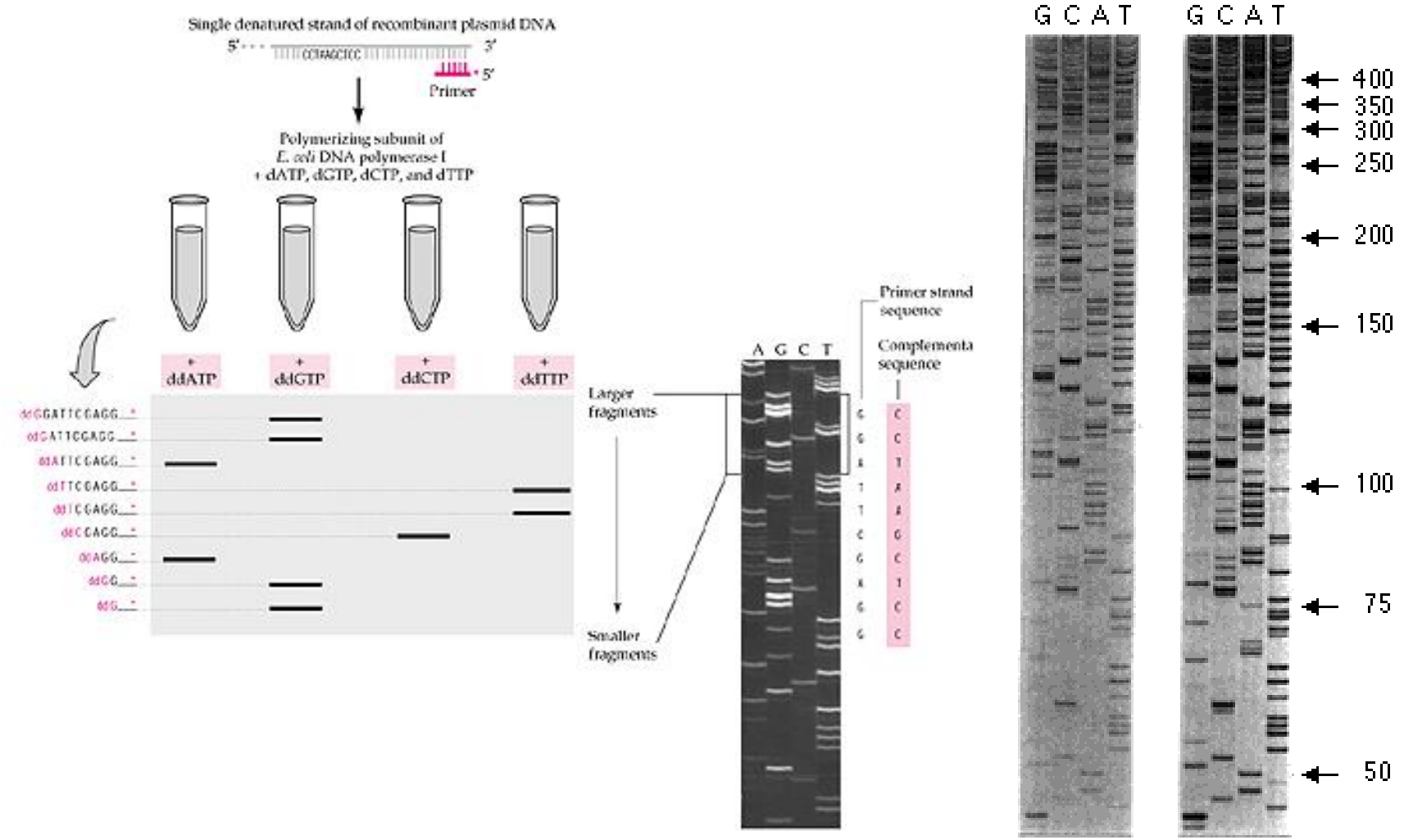
F. SANGER, S. NICKLEN, AND A. R. COULSON
Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977





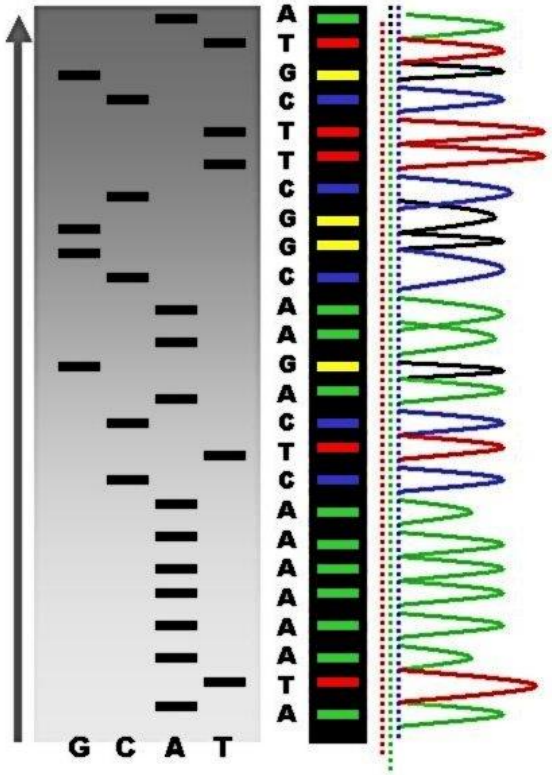
Sanger Dideoxy Sequencing

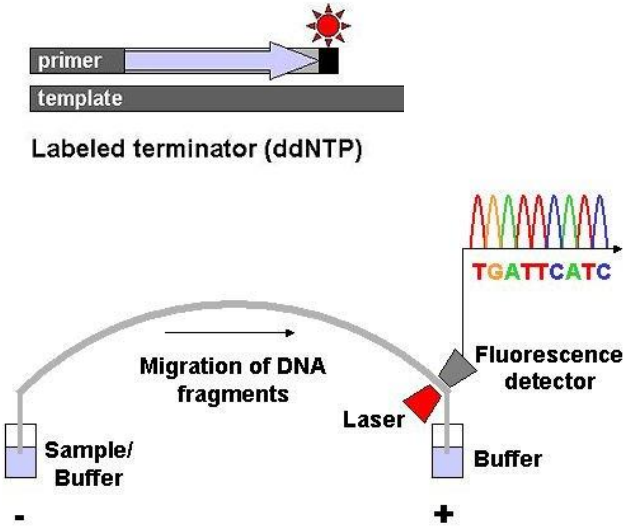


Sanger, F. et al. Nature 24, 687–695 (1977).

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

DNA sequencing: (dideoxyl termination)



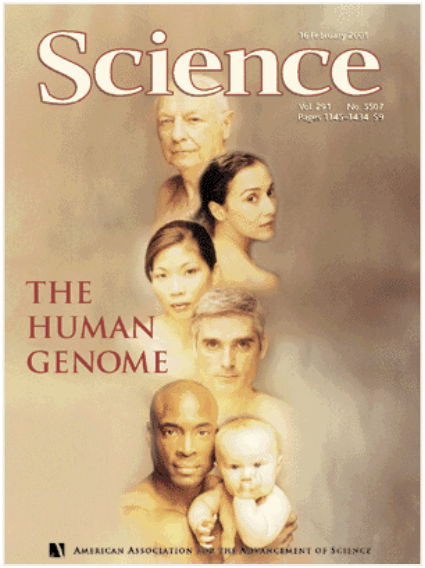


Stats:

- read lengths up to 1,000 bp
- accuracy 99.999%
- costs $0.50 per kilobase
- low throughput

February 2001: Completion of the Draft Human Genome

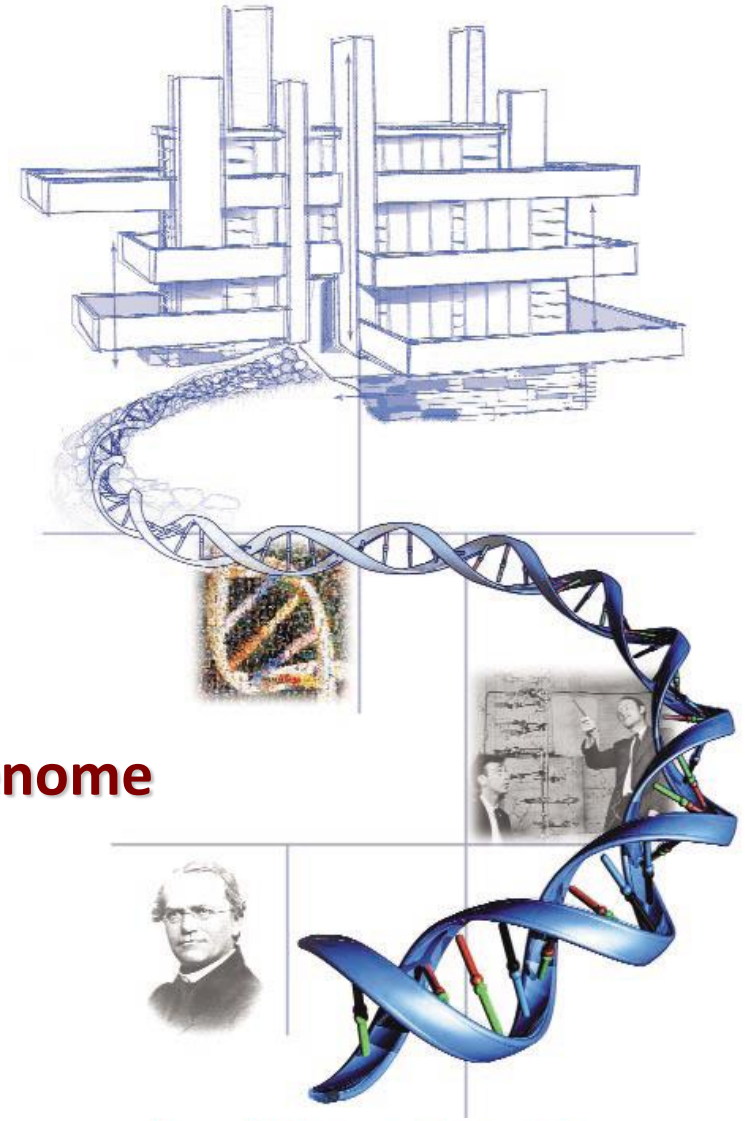
>10 years to finish
USD 3 billion



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



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[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]

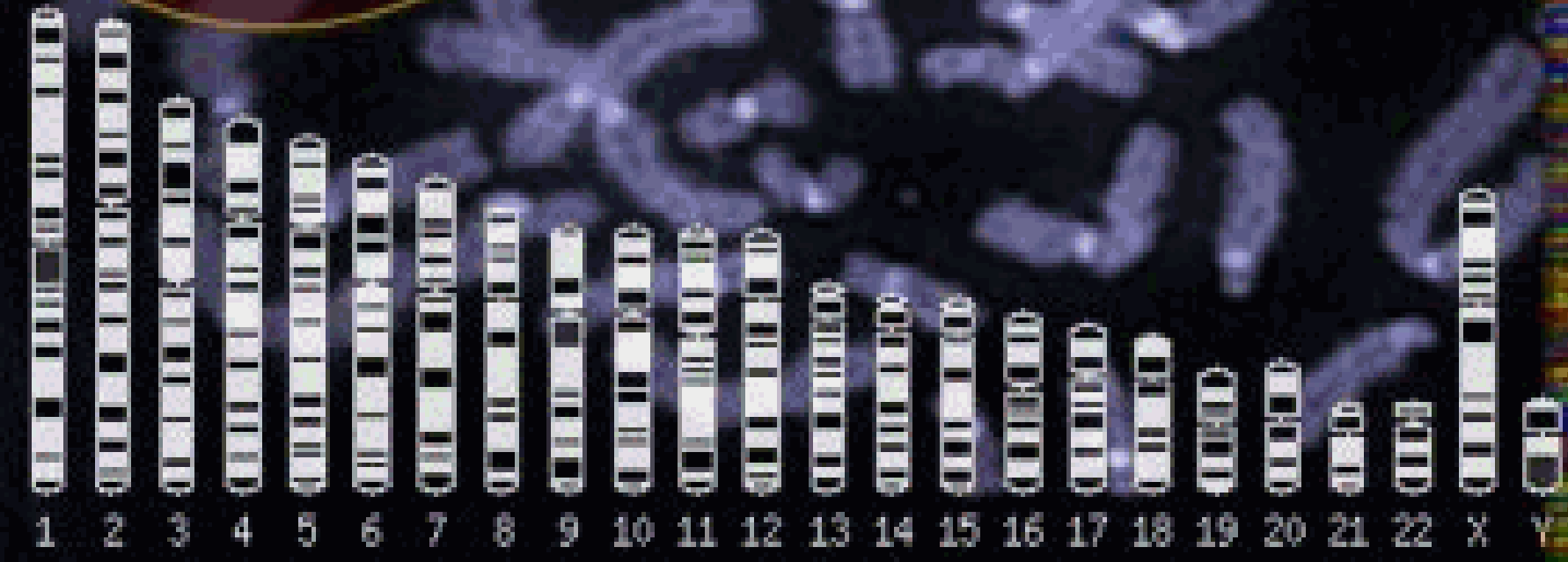
* **Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings):** Richa Agarwala[26], L. Aravind[26], Jeffrey A. Bailey[27], Alex Bateman[2], Serafim Batzoglou[1], Ewan Birney[28], Peer Bork[29,30], Daniel G. Brown[1], Christopher B. Burge[31], Lorenzo 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 Kaspryzk[39], 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[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[39], 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]

Human Genome Project

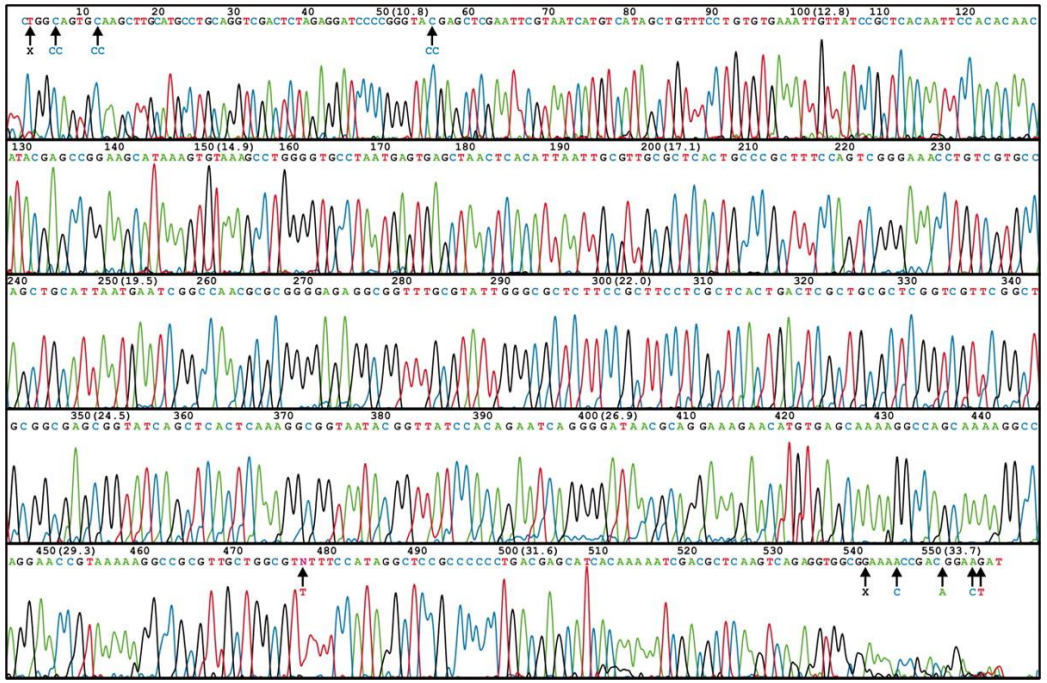
3G

3 billion base pairs



ABI 3730 XL DNA Sequencer





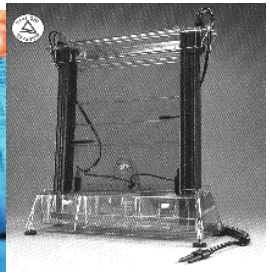
96/384 DNA sequencing in 2 hrs, approximately 600-1000 readable bps per run.

1-4 MB bps/day

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

Sequencing Evolution





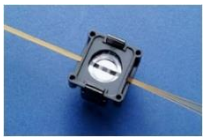
- <1,000 bases per day







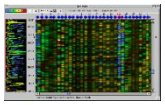




~ 20 years







- >10,000 bases per day

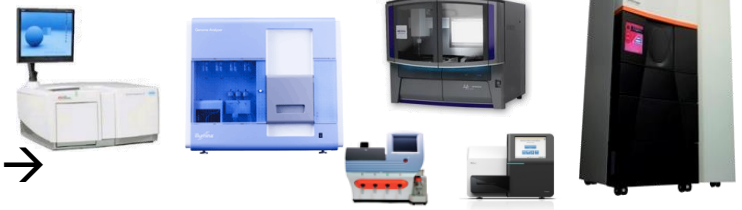
- >1,000,000 bases per day

Human Genome Project



Sequencing Revolution

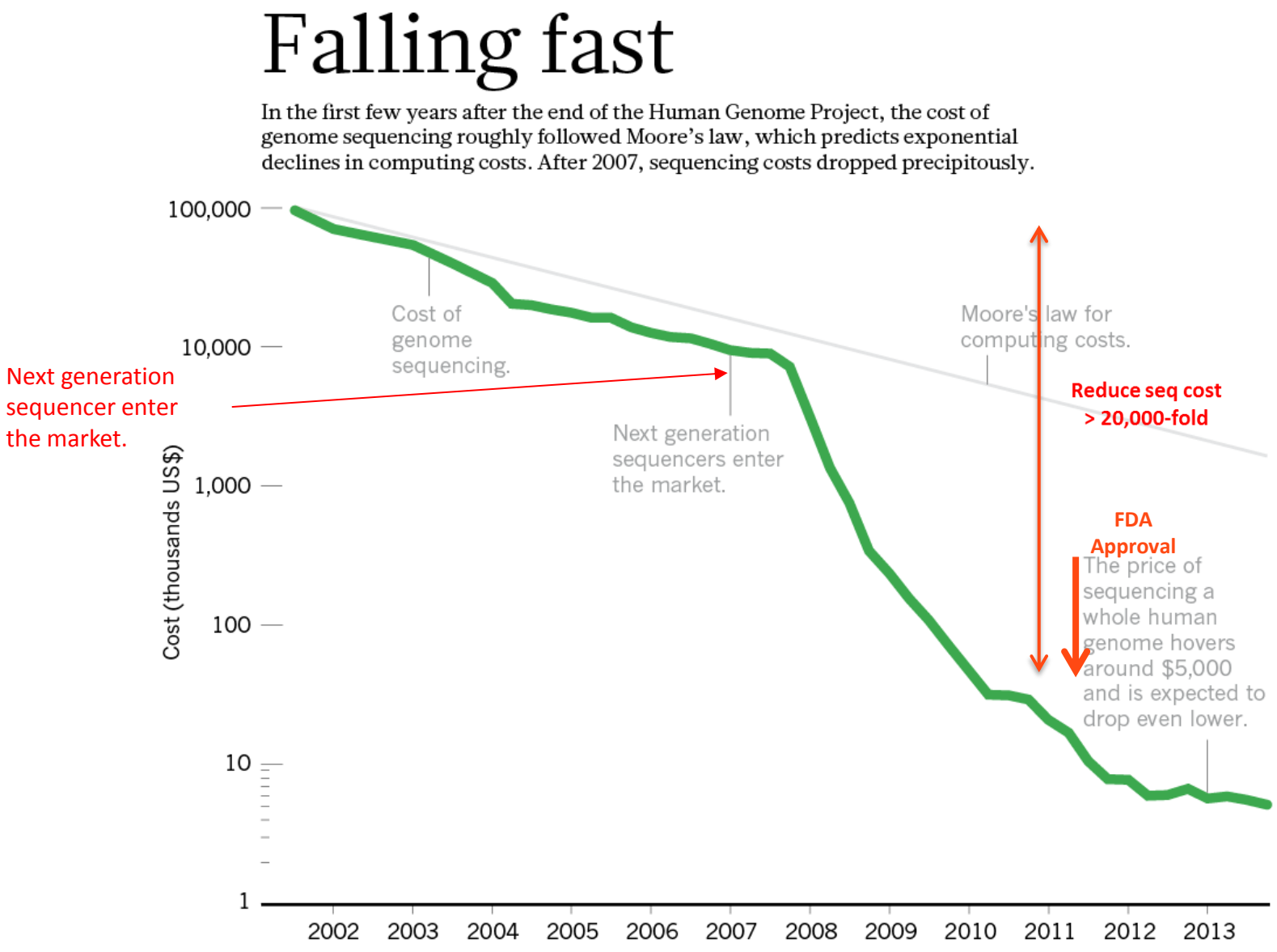




2005 →

- >200,000,000 bases per day

Cost of Genome Sequencing



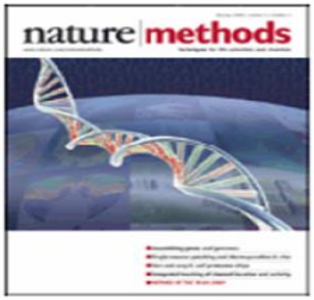
Next generation sequencer enter the market.

NGS Trends

- **2005-2010: Bigger and cheaper**
  - Since 2005 NGS "arms race" has been to
    - produce more reads, longer reads
    - Higher throughput at lower cost
      - 454 GS -> GS FLX Titanium
      - Illumina GAII -> HiSeq
      - AB SOLiDv1,2,3 and 4

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

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: Illumina (2012)



HiSeq 2500/1500

Two run modes. High output or rapid run.

Flexibility to batch process multiple samples with high output in a single run, or get rapid results with fewer samples for time-critical studies.

	HiSeq 2500		HiSeq 1500	
Run Mode	High Output	Rapid Run*	High Output	Rapid Run*
Output (2 × 100 bp)	600 Gb	120 Gb	300 Gb	60 Gb
Run Time (2 × 100 bp)	~11 days	~27 hours	~8.5 days	~27 hours
Cluster Generation	cBot	On board	cBot	On board
Paired-end Reads	6 Billion	1.2 Billion	3 Billion	600 Million
Single Reads	3 Billion	600 Million	1.5 Billion	300 Million
Maximum Read Length**	2 × 100 bp	2 × 150 bp	2 × 100 bp	2 × 150 bp
Quality Scores***	> 85% (2 x 50 bp) > 80% (2 x 100 bp)			

HiSeq 2000/1000

One run mode. High output.

Generate the highest output and number of reads for batch processing multiple samples in a single run.

	HiSeq 2000	HiSeq 1000
Output (2 × 100 bp)	600 Gb	300 Gb
Run Time (2 × 100 bp)	~11 days	~8.5 days
Cluster Generation	cBot	cBot
Paired-end Reads	6 Billion	3 Billion
Single Reads	3 Billion	1.5 Billion
Maximum Read Length**	2 × 100 bp	2 × 100 bp
Quality Scores***	> 85% (2 x 50 bp) > 80% (2 x 100 bp)	

Throughput of NGS machines: SOLiD (2013)

life technologies | 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

Sequencing on the chip surface

NEXT GENERATION SEQUENCING

Basis Principles

NGS

Template fragmentation (library) -> **Amplification** -> **Sequencing**

	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)

Solexa Hiseq

illumina
Solexa

Illumina HiSeq 2500



	High Output Mode	Rapid Run Mode	
Read length (bp)	2x100	2x100	2x150
Yield (Gb)	~600	~120	~180
Run time	~11 days	~27 hours	~39 hours
Bases >Q30	>80%	>80%	>75%
Reads passing filter	>90%	>90%	>90%
Number of flow cells	2	2	2
Lanes/flow cell	8	2	2
Cluster generation	cBot	onboard	onboard



MiSeq



Next-Seq



HiSeq X Ten

Illumina Technology

1. DNA library preparation (Fig 1)
2. Templplification (Bride amplification) (Fig 2-6)
3. Sequencing

1. Prepare genomic DNA sample

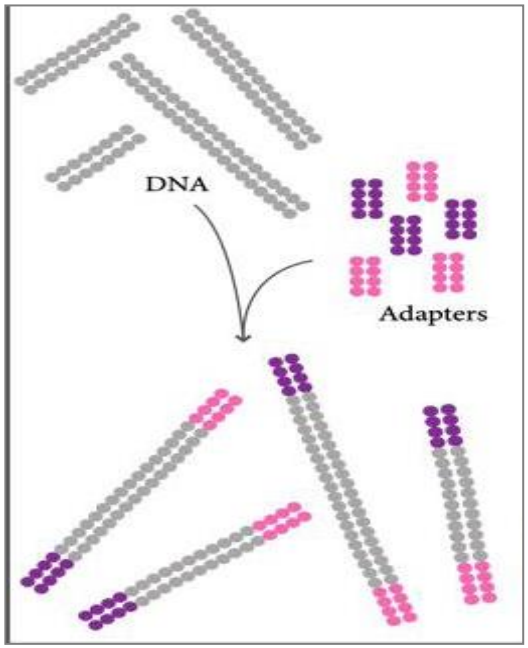
Randomly fragment genomic DNA and ligate adapters to both end of fragments

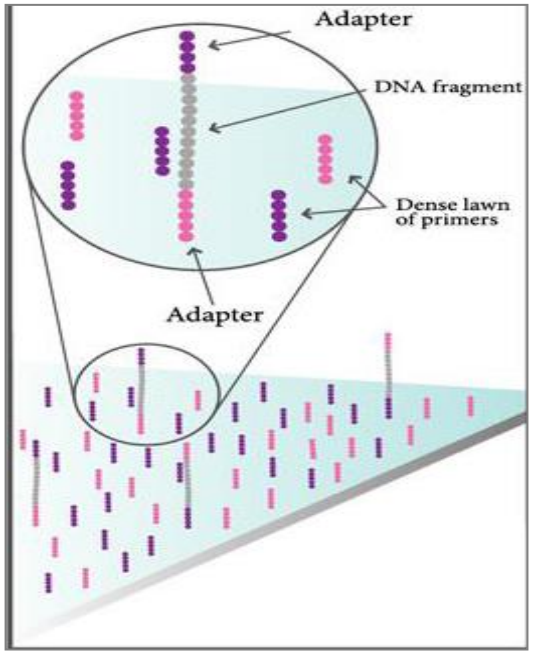
2. Attach DNA to surface

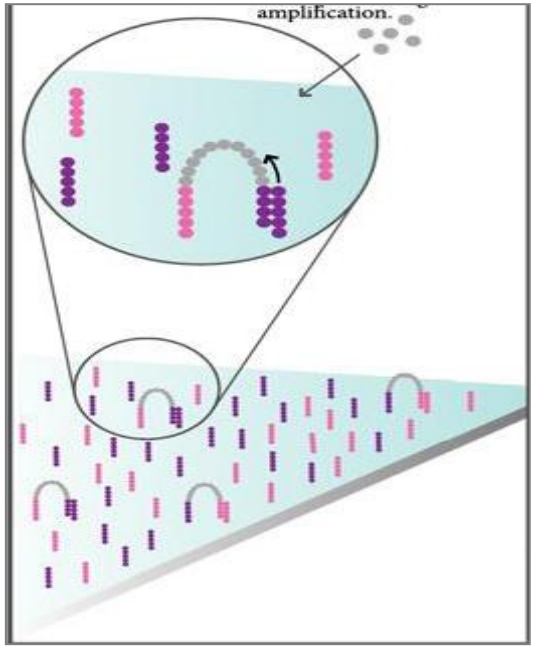
Bind single stranded fragments randomly to the inside surface of the flowcell channels

3. Bridge amplification

Add unlabelled nucleotides and enzyme to initiate solid-phase bridge amplification







Illumina Technology (continue)

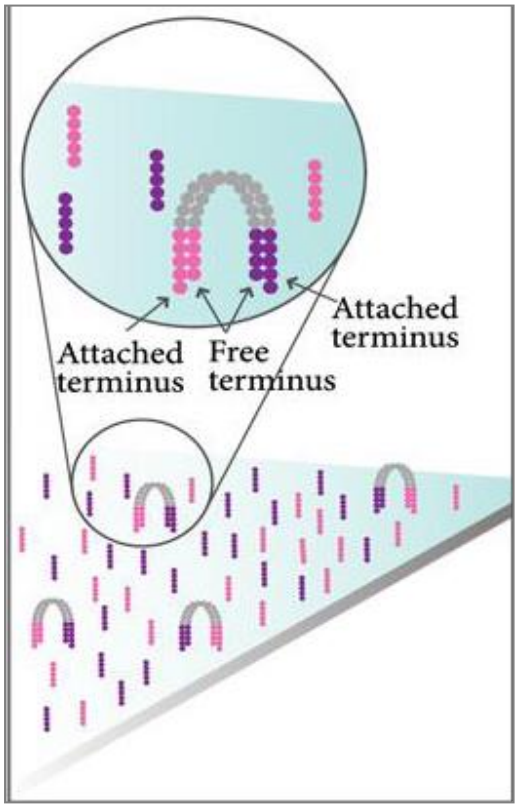
1. DNA library preparation (Fig 1)
2. Templplification (Bride amplification) (Fig 2-6)
3. Sequencing

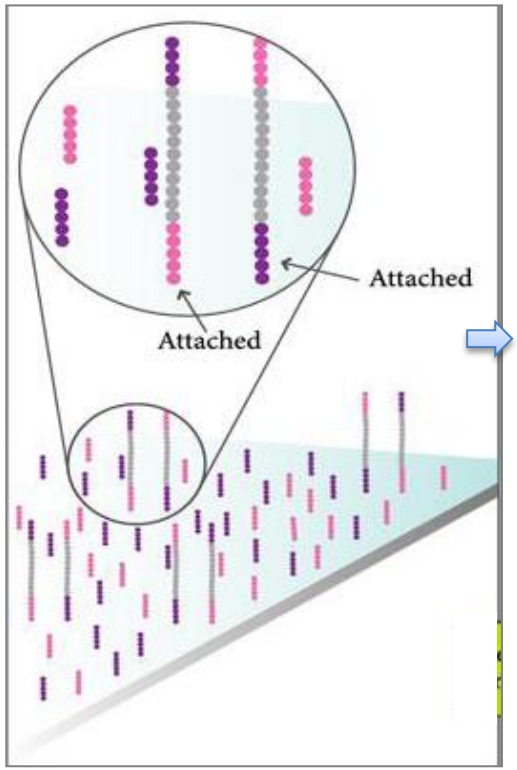
4. Fragments become double stranded

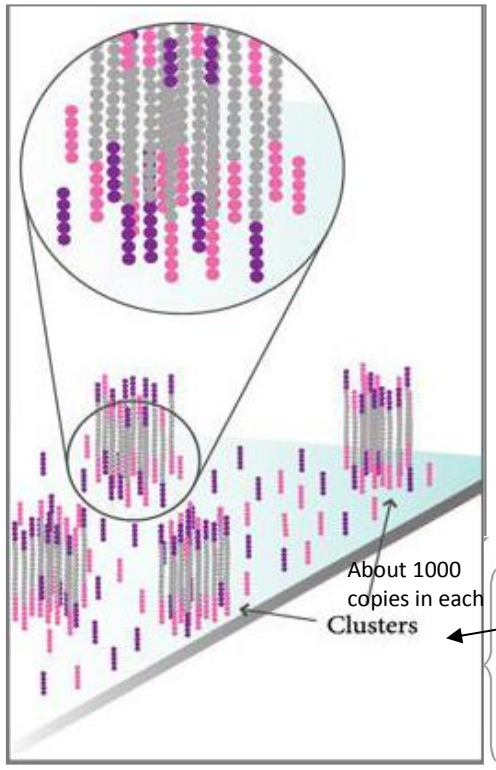
5. Denature the double stranded molecules

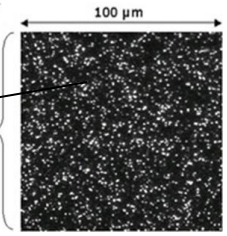
6. Completion of amplification

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



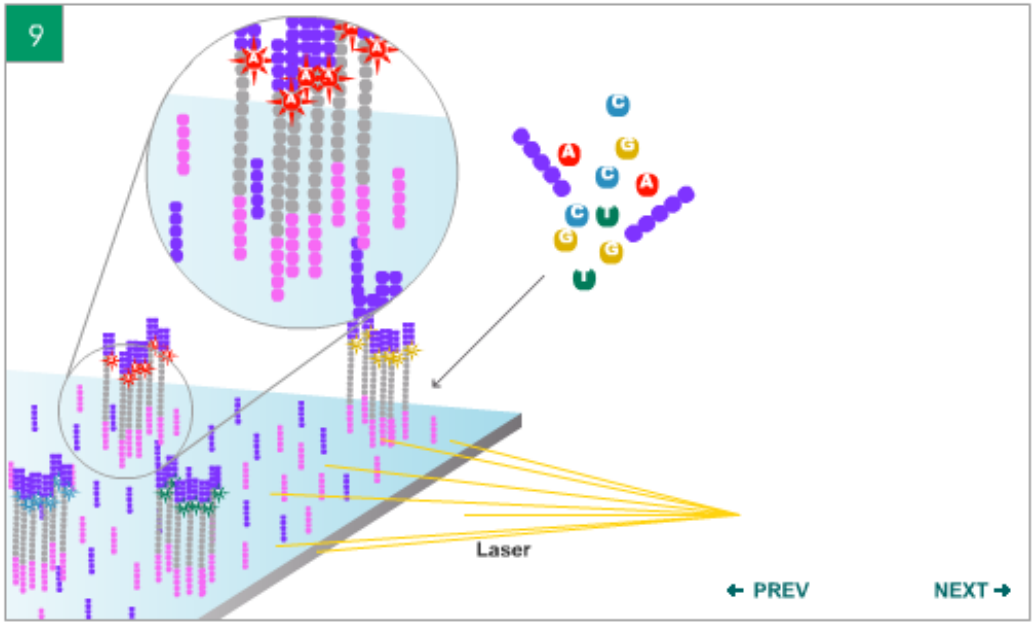


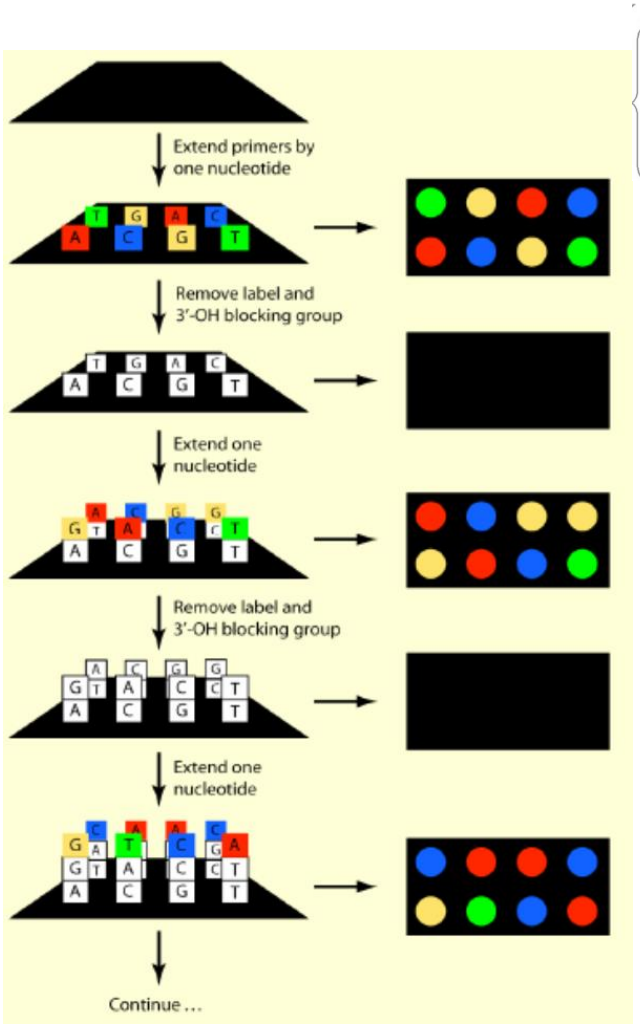




Illumina Technology (continue)

1. DNA library preparation (Fig 1)
2. Templplification (Bride amplification) (Fig 2-6)
3. Sequencing (sequencing-by-synthesis)

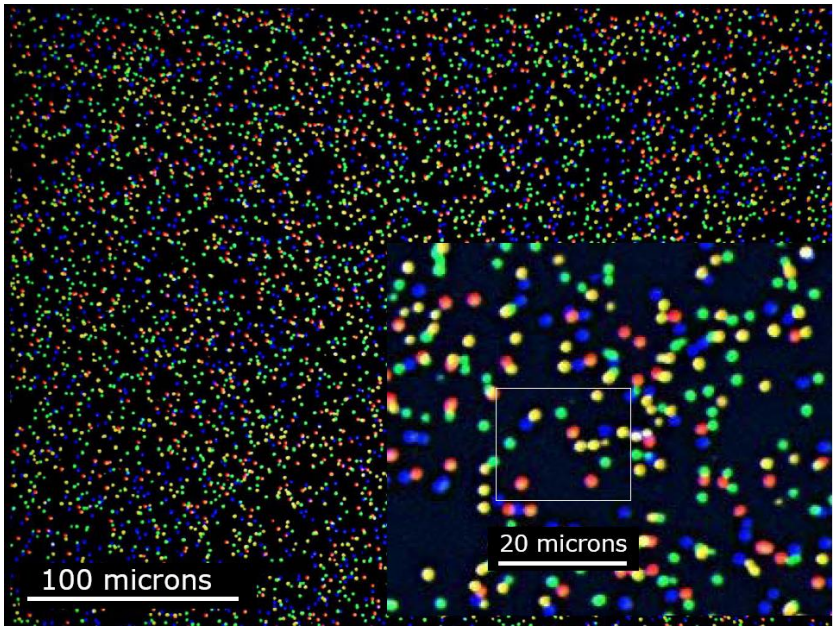




Illumina Technology (continue)

illumina

Sequencing on the chip surface

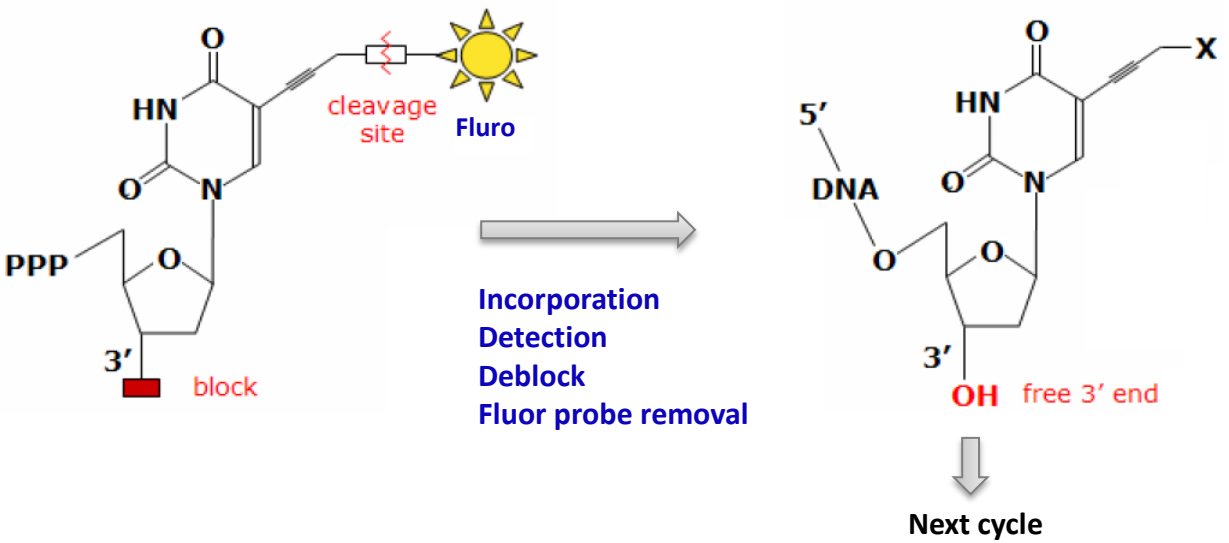




The identity of each base of a cluster is read off from sequential images

Illumina Technology (continue)

Chemistry of Illumina DNA Sequencer



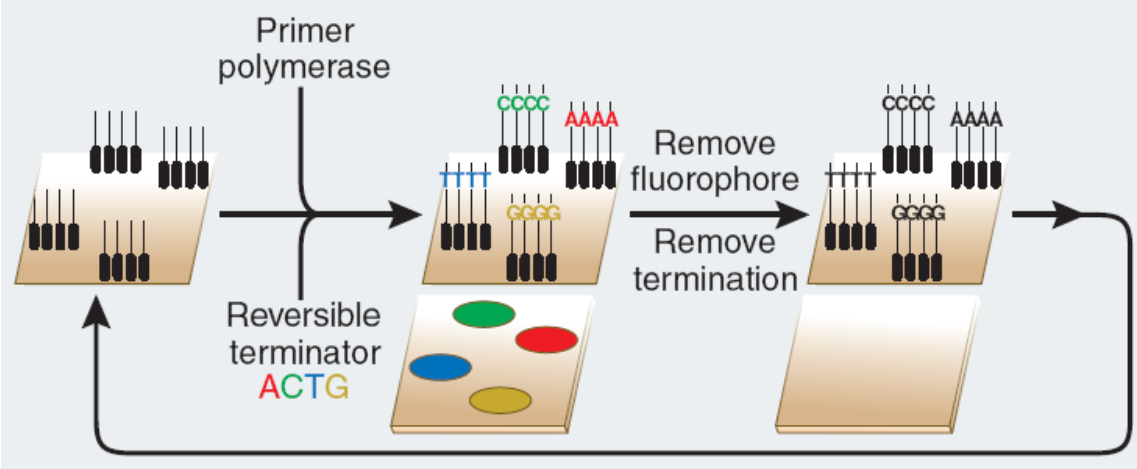
Solexa technology (sequencing-by-synthesis)

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.

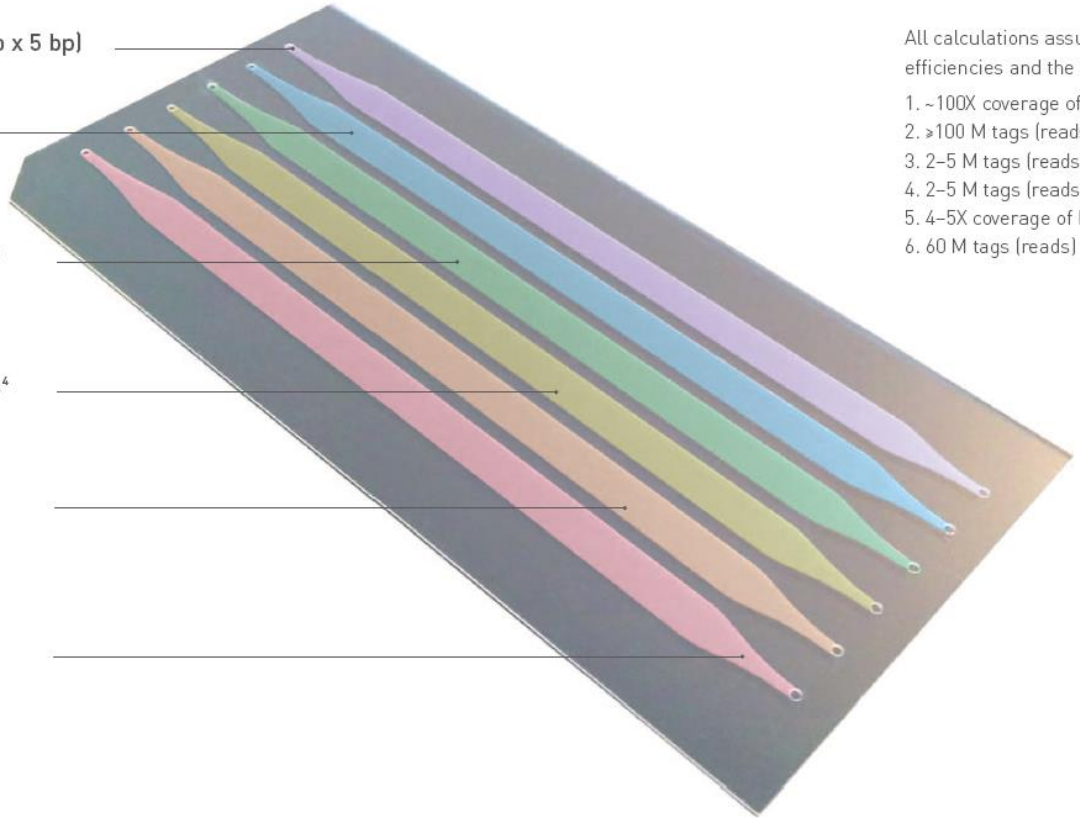


applied biosystems®
by life technologies™



SOLiD M5500XL

- ~2 Exome samples[1] (50 bp x 5 bp)
- 1 mRNA profiles[2] (50 bp x 35 bp)
- 20–50 ChIP-Seq samples[3] (35 bp)
- 20–50 microRNA samples[4] (35 bp)
- 1 Structural variation/ inversion experiment[5] (2 x 60 bp)
- ~2 Methylation samples[6] (enriched) (75 bp)



All calculations assume optimal running efficiencies and the following coverage:

1. ~100X coverage of human exome
2. ≥100 M tags (reads) per sample
3. 2–5 M tags (reads) per sample
4. 2–5 M tags (reads) per sample
5. 4–5X coverage of human genome
6. 60 M tags (reads) per sample

life technologies™

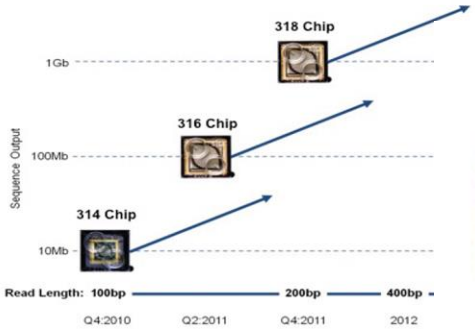
NGS Trends

- **2005-2010: Bigger and cheaper**
  - Since 2005 NGS “arms race” has been to
    - produce more reads, longer reads
    - Higher throughput at lower cost
      - 454 GS -> GS FLX Titanium
      - Illumina GAII -> HiSeq
      - AB SOLiDv1,2,3 and 4
- **2011: Broader application (eg. clinical)**
  - New technologies with some/all of:
    - Moderate throughput
    - Faster run time
    - Higher accuracy
    - Single molecule detection
    - Ion Torrent, Illumina MiSeq. PacBio RS
    - Heavy guns still being developed

Ion Torrent Proton

Semiconductor-based Sequencing Technology: 2-4 hr run time
Sequencing Capacity is Built on “Chip”

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



Ion Torrent PGM



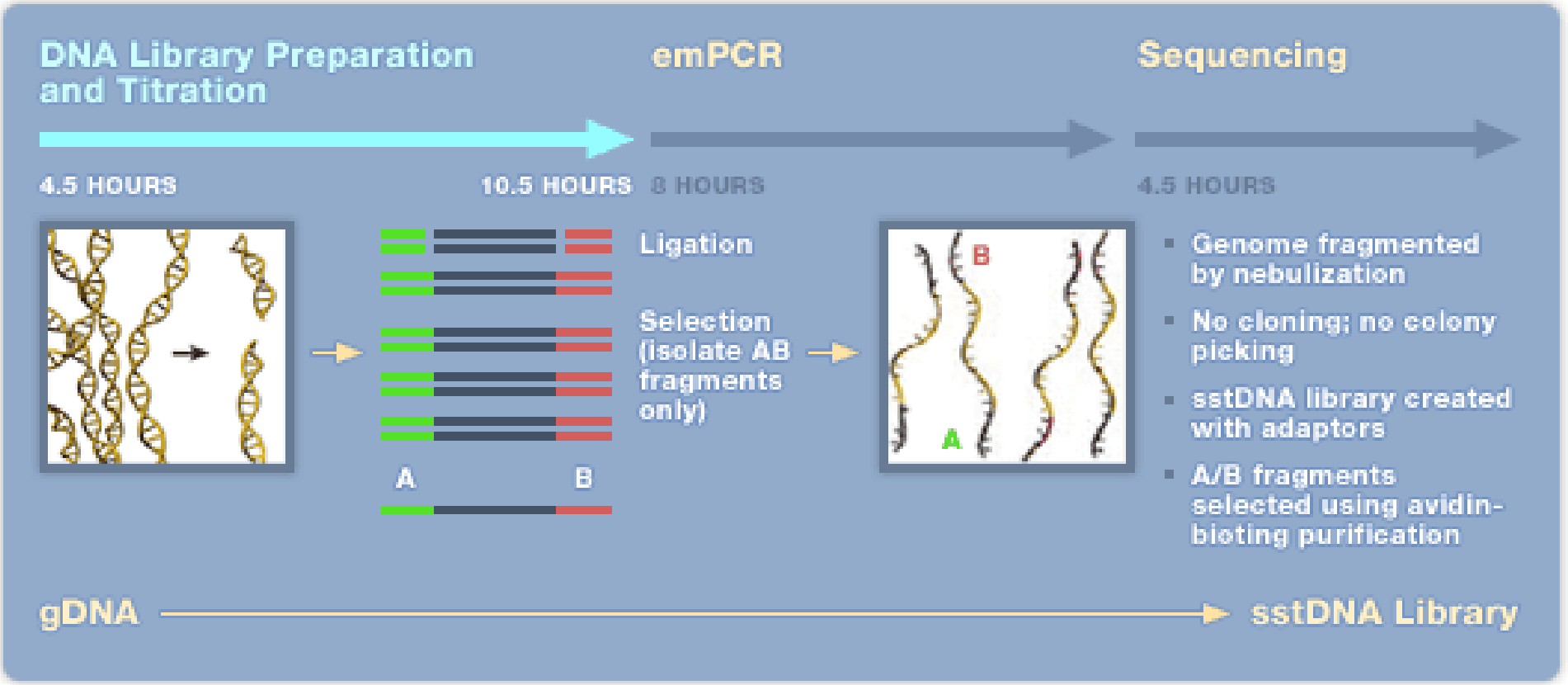
Ion Proton



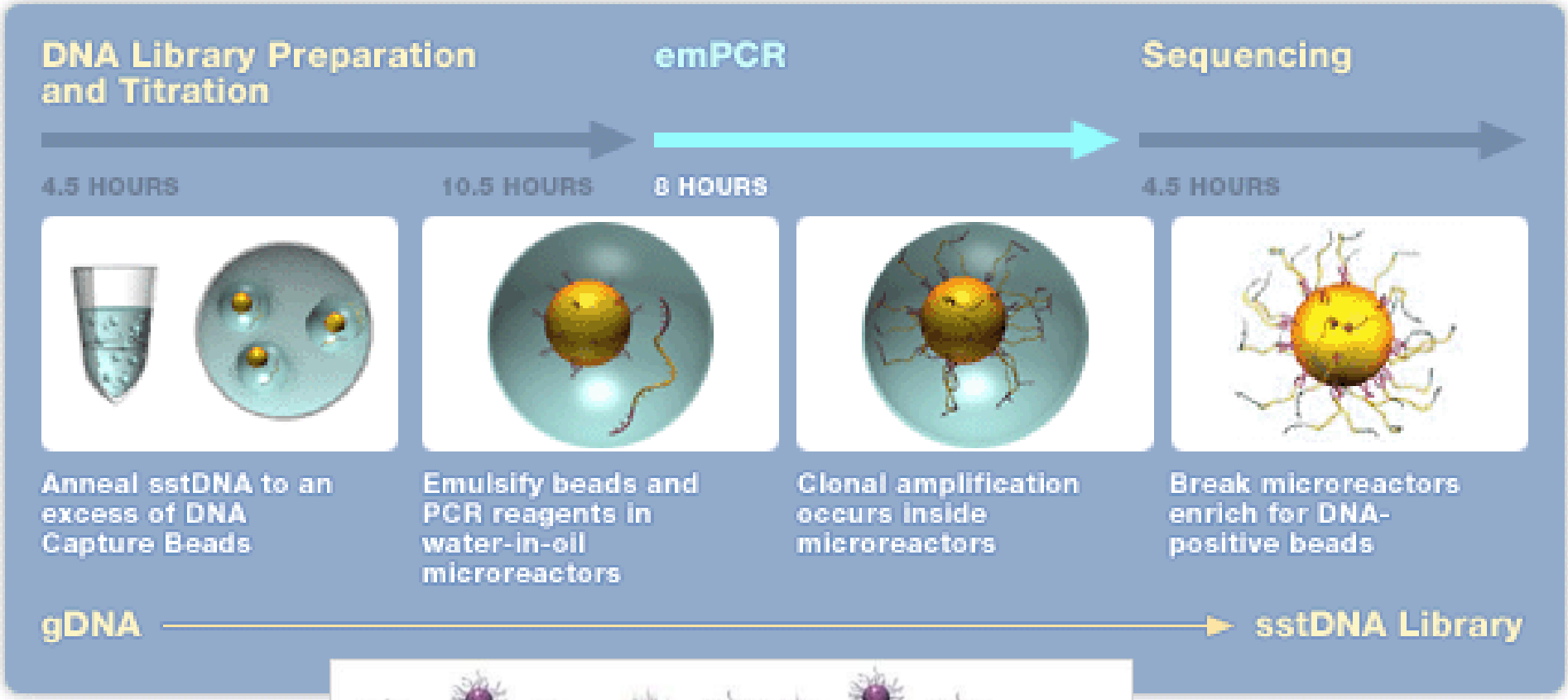
Ion Proton I
165 million wells
Up to 200 bases
~10Gb

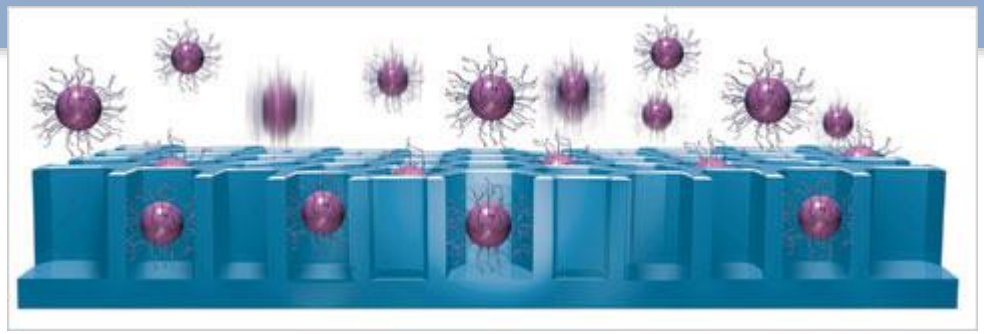
Ion Proton II
660 million wells
400 bases?
~60Gb

Step 1. DNA Library Preparation



Step 2. Emulsion PCR (emPCR)

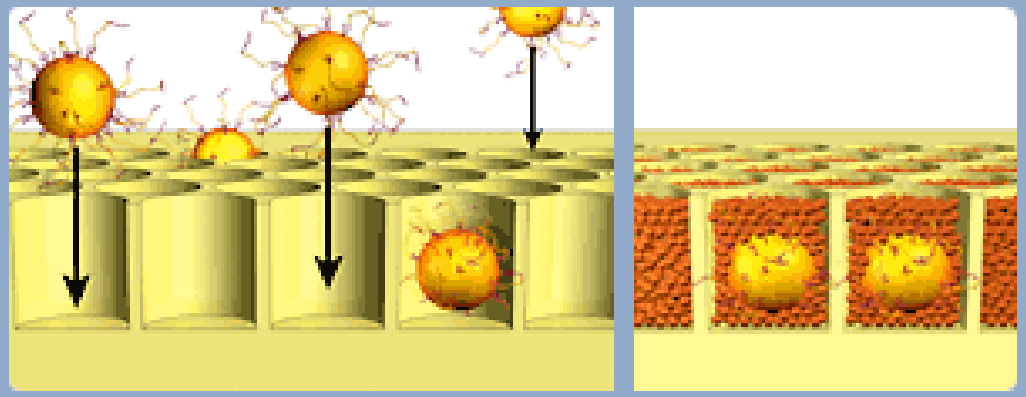




Step 3. Sequencing

DNA Library Preparation and Titration → **emPCR** → **Sequencing**

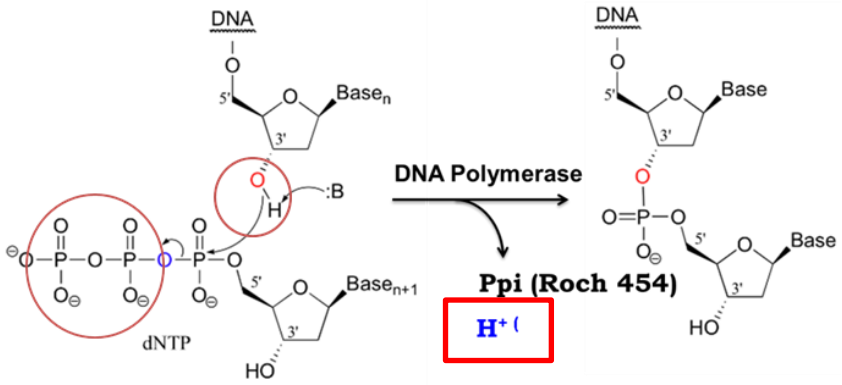
4.5 HOURS | 10.5 HOURS | 8 HOURS | 4.5 HOURS

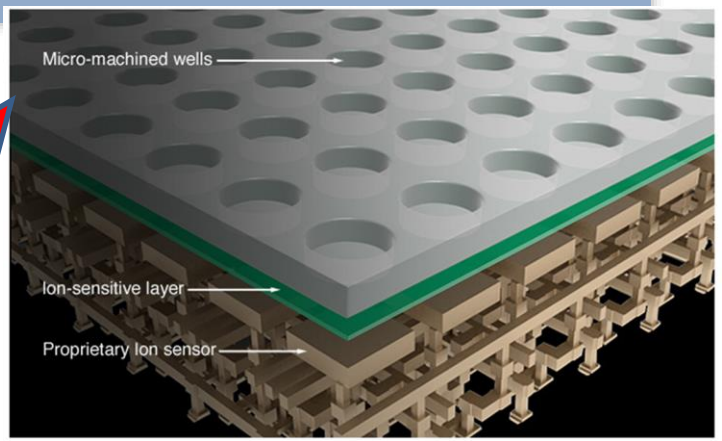


- Well diameter: average of 44µm
- 200,000 reads obtained in parallel
- A single cloned amplified sstDNA bead is deposited per well

Amplified sstDNA library beads → Quality filtered bases

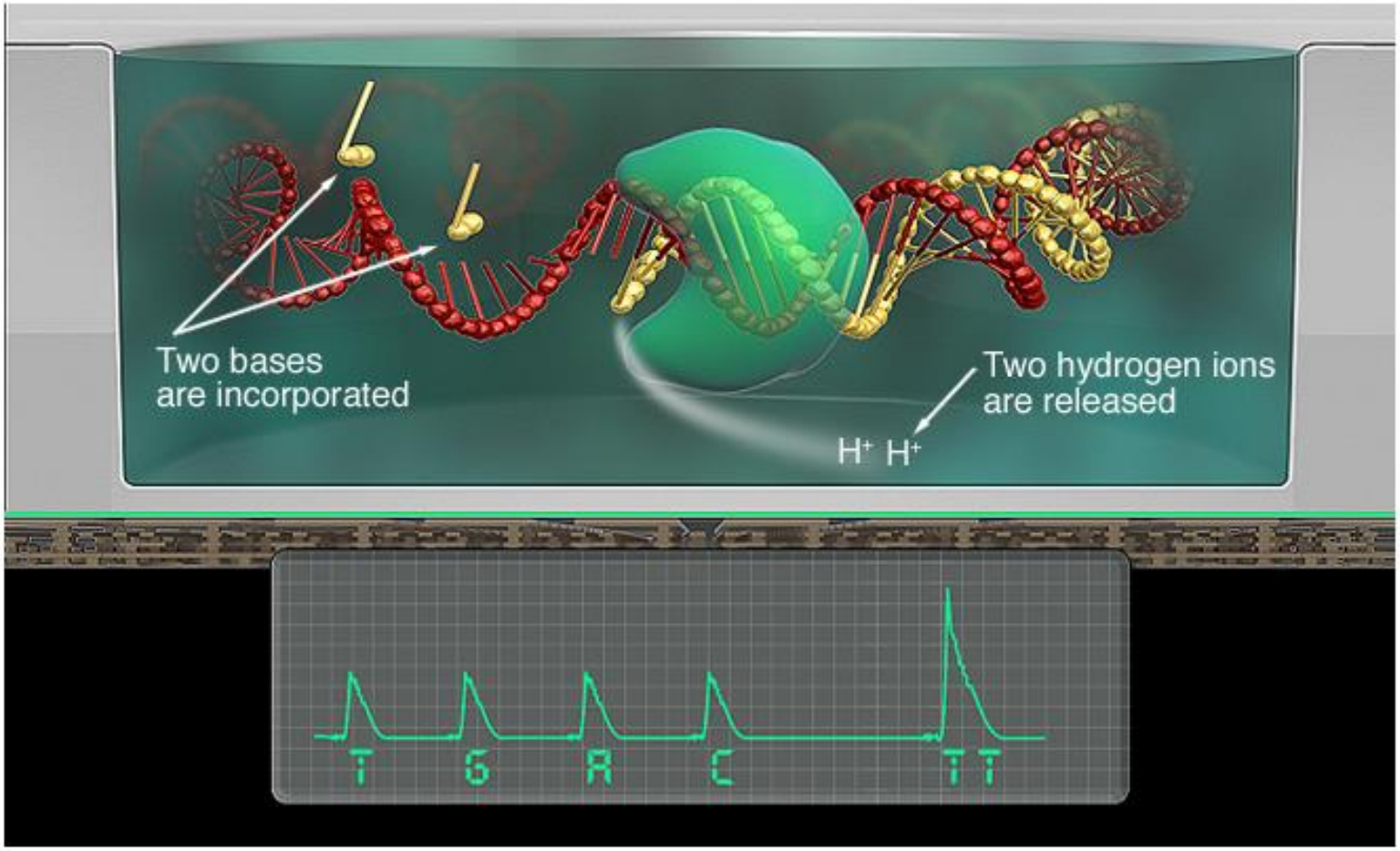
Chemistry in DNA Polymerization





Ion Torrent PGM

ion torrent
by life technologies



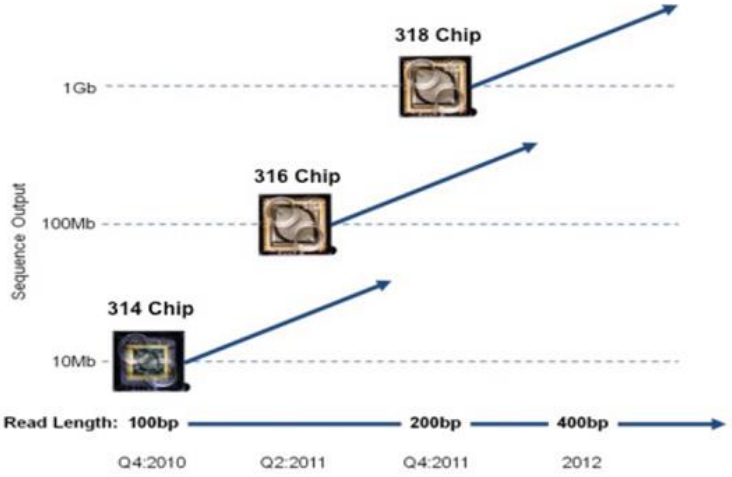
Ion Torrent Proton

Semiconductor-based Sequencing Technology: 2-4 hr run time
Sequencing Capacity is Built on “Chip”

Ion Torrent PGM



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



Ion Proton



Proton P1



165 million

Ion Proton I
165 million wells
Up to 200 bases
~10Gb

Ion Proton II
660 million wells
400 bases?
~60Gb

Ion Torrent Data



Greatest weakness is homopolymer repeats

NGS Trends

- **2005-2010: Bigger and cheaper**
  - Since 2005 NGS “arms race” has been to
    - produce more reads, longer reads
    - Higher throughput at lower cost
      - 454 GS -> GS FLX Titanium
      - Illumina GAII -> HiSeq
      - AB SOLiDv1,2,3 and 4
- **2011: Broader application (eg. clinical)**
  - New technologies with some/all of:
    - Moderate throughput
    - Faster run time
    - Higher accuracy
    - Single molecule detection
    - Ion Torrent, Illumina MiSeq. PacBio RS
    - Heavy guns still being developed

Third-Fourth-Fifth-Generation Sequencers

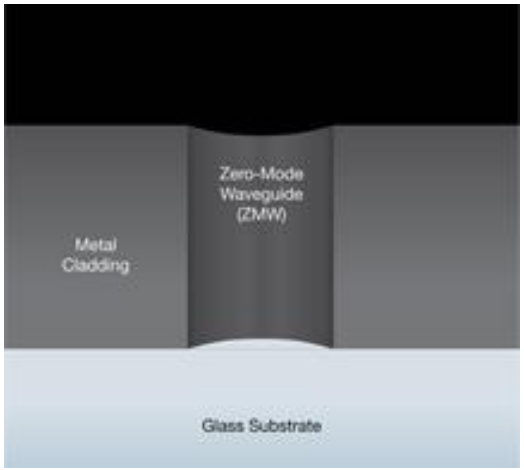
NEW GENERATION SEQUENCING

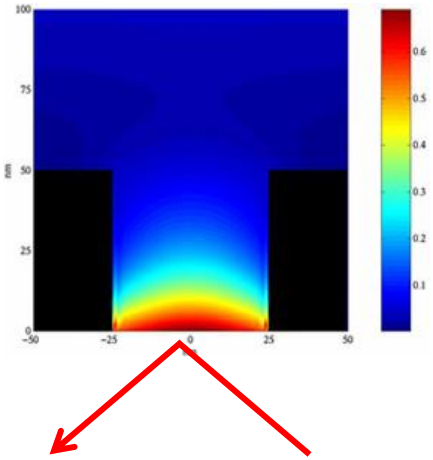
NGS

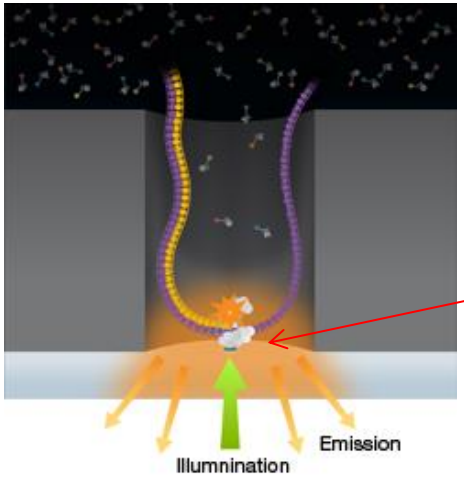
Single Molecule Real-Time Sequencing

Pacific Biosciences
Single Molecule Real Time (SMRT)

www.pacificbiosciences.com







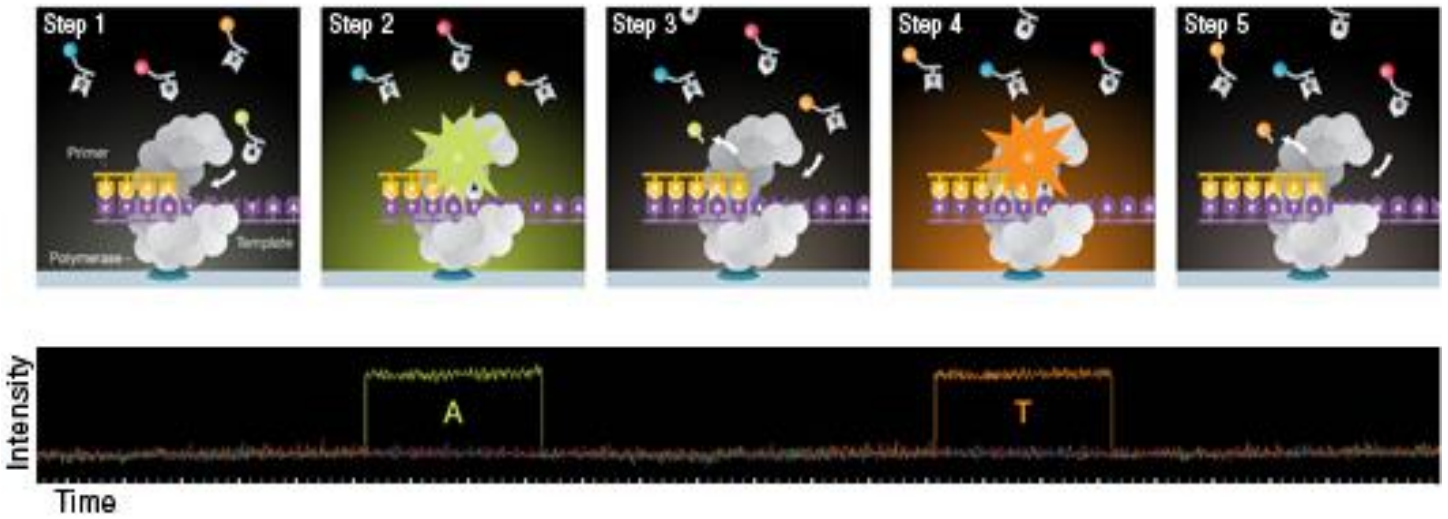
- Read Length >5kb
- 86% accuracy

Eid *et al.* Science November 2008

Pacific Biosciences

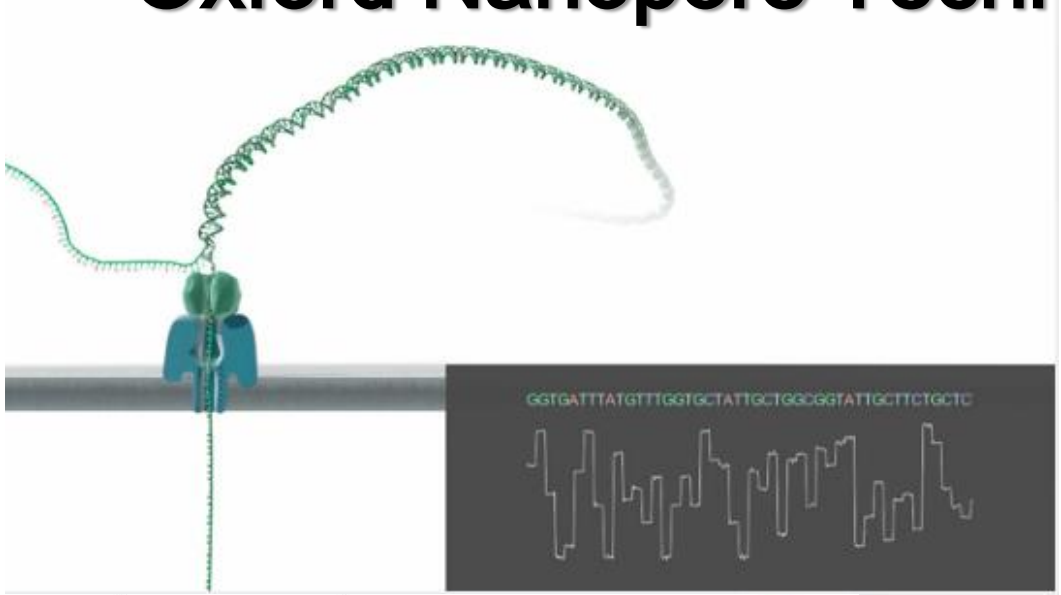
Single Molecule Real Time (SMRT)

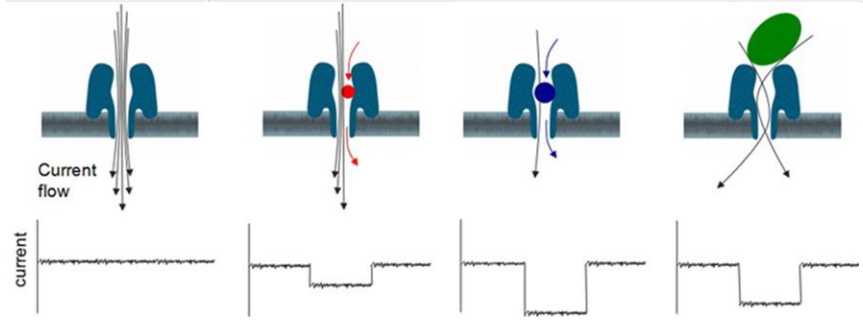
www.pacificbiosciences.com



Labeled nucleotides rapidly diffuse | Labeled “A” is held during incorporation | Fluorescent label is released | Labeled “T” is held during incorporation | . . .

Oxford Nanopore Technologies





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.

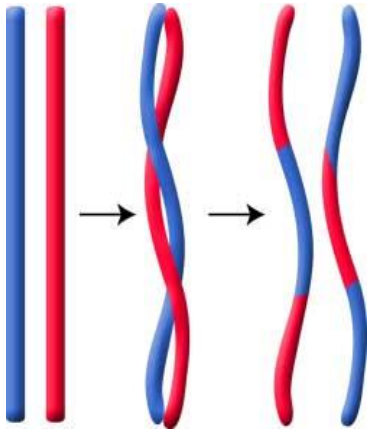
In a talk at the Advances in Genome Biology and Technology conference this morning, chief technology officer Clive Brown presented details of what the British company calls its "new generation" of single-strand sequencing technology, featuring accurate long reads of single-stranded DNA molecules, including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



NEXT GENERATION SEQUENCING

Applications & Analysis

Biomarker Discovery
Translational Bioinformatics
NGS
Biological Pathways
Evolution
Bioinformatics
Cancer Genome
Disease
Visualization
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

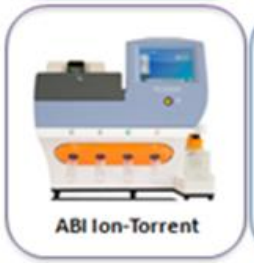


Applications on Biomedical Sciences


ABI SOLiD 5500XL


Illumina Hiseq 2000


Roche 454 FLX


ABI Ion-Torrent

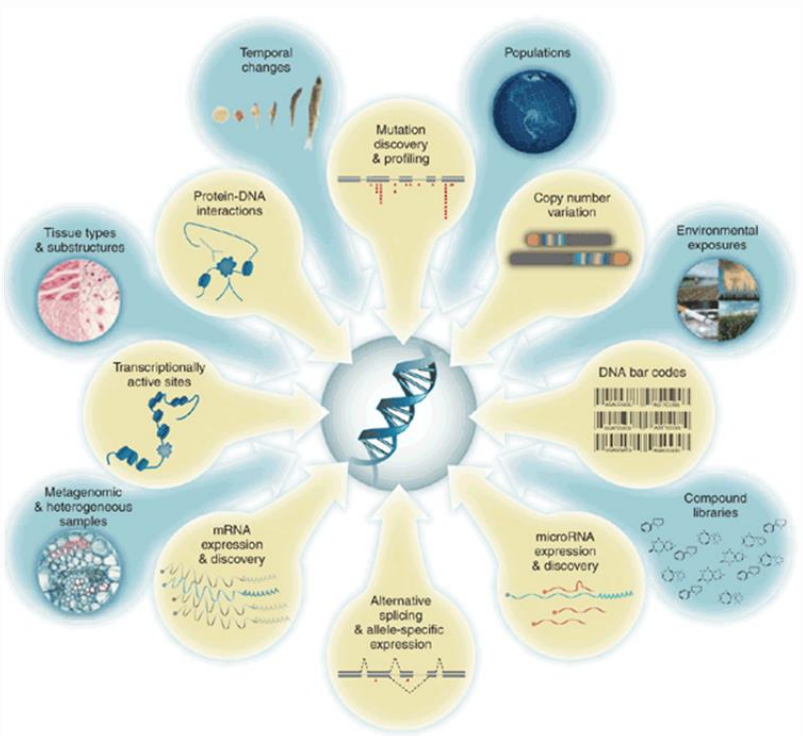

Illumina Miseq


ABI 3730XL

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

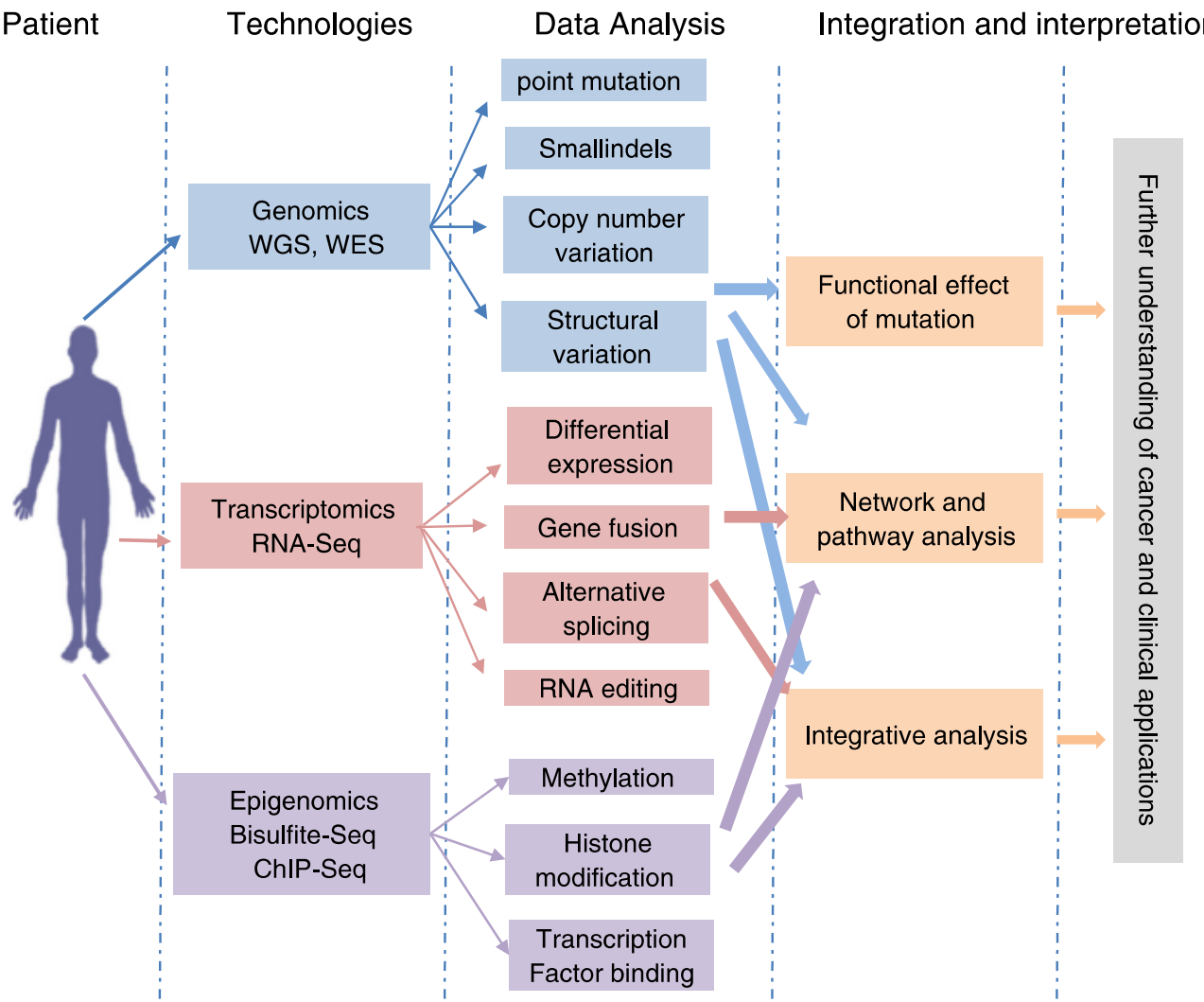
RNA Small RNA Sequencing
Transcriptome Sequencing
De novo Transcriptome Sequencing
Meta-Transcriptome Sequencing

BIOINFORMATICS



IBM X3850

Application of NGS sequencing in cancer research



Genome sequencing (DNA-seq)

WGS : Whole genome sequencing

- human genome : $3*10^9$ bp,
- 30X coverage,
- $ 200, 000~350,000)

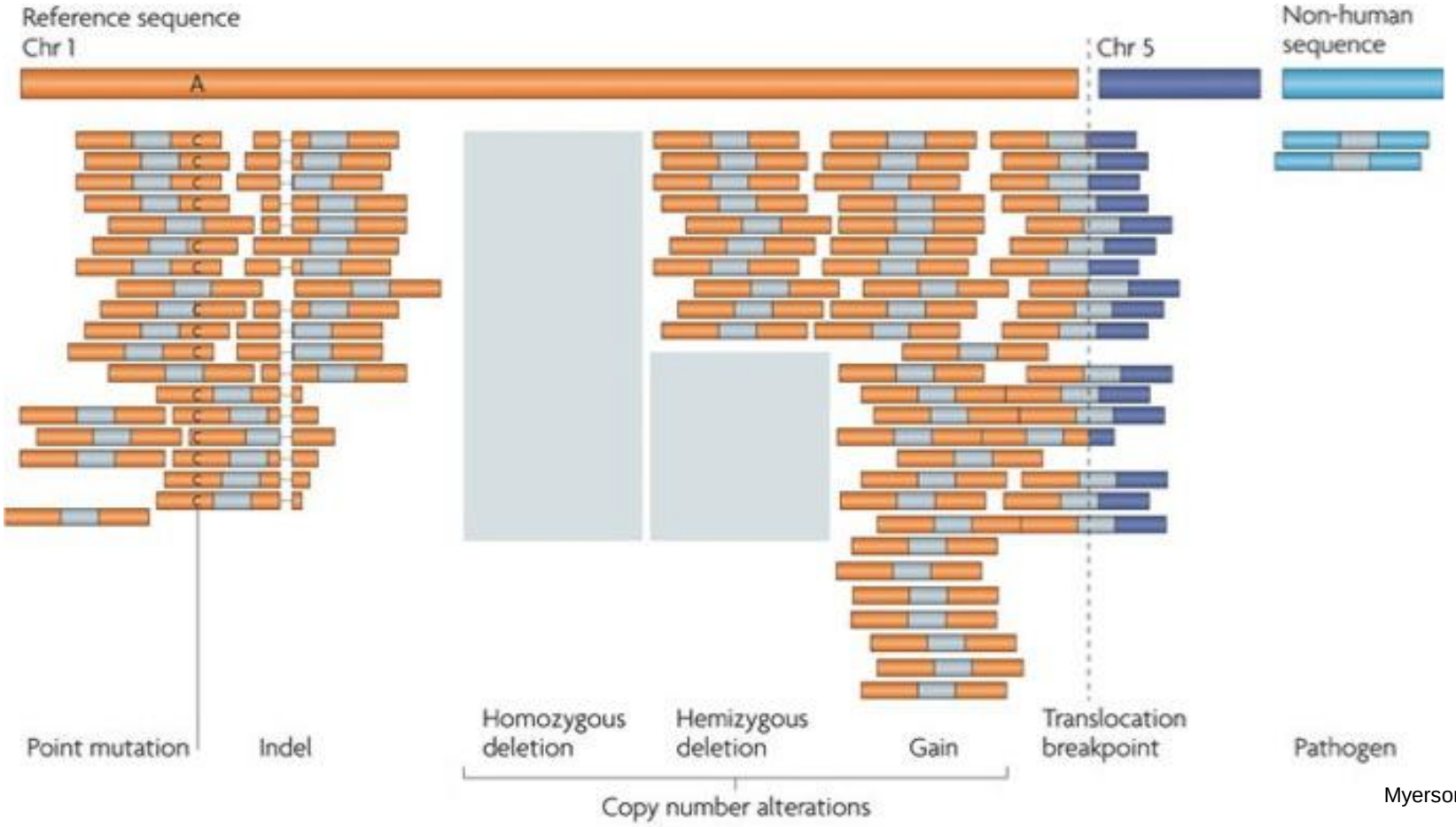
WES: Whole Exome sequencing (coding region)

-1 human whole exome: 1% of the human whole genome
-100X coverage
- $50*10^6$ bp,
- $ 40,000-80,000

Target re-sequencing

-Cancer panel (45 -1000 cancer related genes), Colorectal cancer panel, Lung cancer panel, BRCA1/2 panel, inherited disease panel……
- 500-1000 X coverage
- much cheaper than WGS and WES

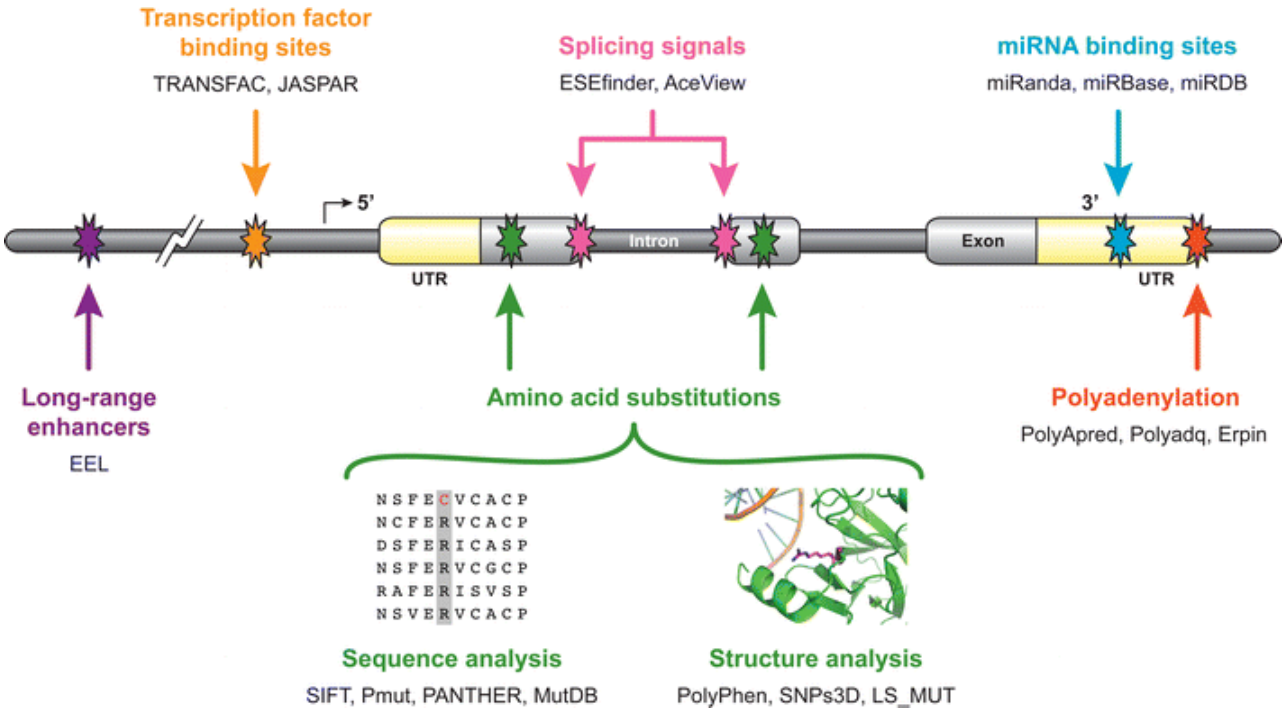
At DNA Level



Myerson et al 2010

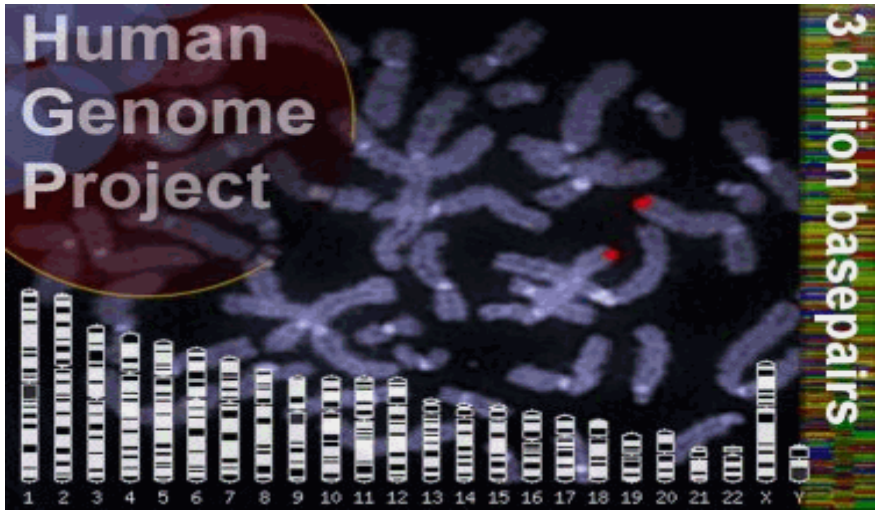
Annotation/Interpretation

Driver vs passenger?



Source: Lee W, Yue P, Zhang Z. Hum Genet. 2009

How much does it cost for a Human Genome?

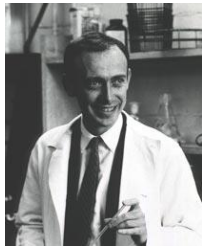




Human Reference Genome

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

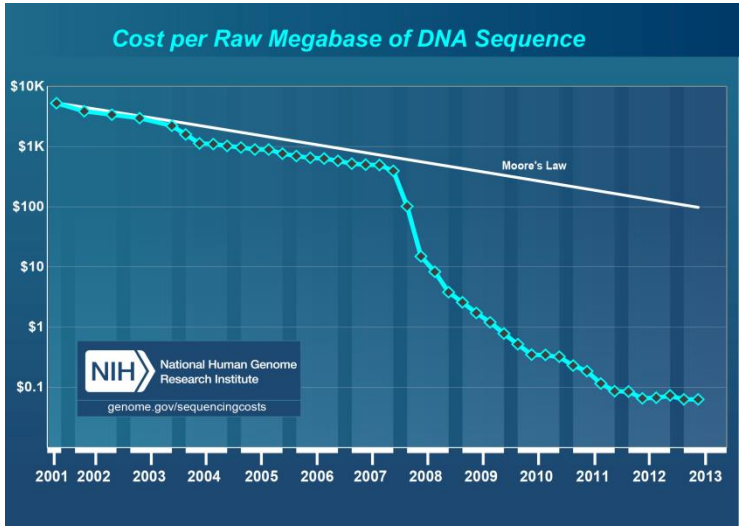


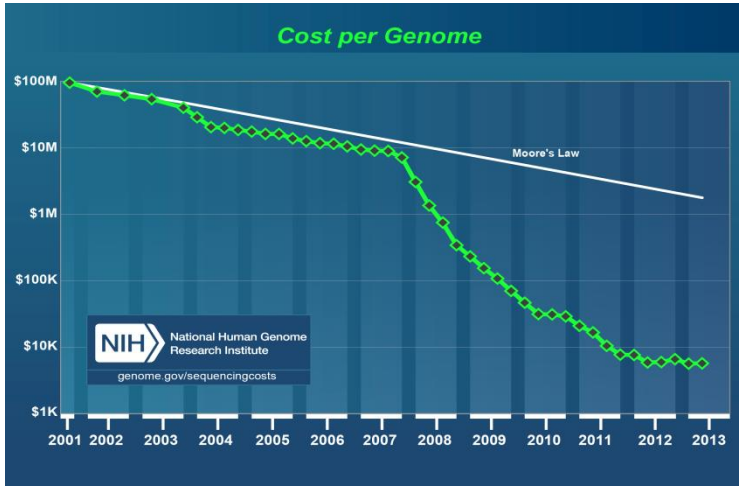
Personal Genome

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

USD 1,000 Human Genome?

Date	Cost per Mb	Cost per Genome
Sep-01	$5,292.39	$95,263,072
Sep-02	$3,413.80	$61,448,422
Oct-03	$2,230.98	$40,157,554
Oct-04	$1,028.85	$18,519,312
Oct-05	$766.73	$13,801,124
Oct-06	$581.92	$10,474,556
Oct-07	$397.09	$7,147,571
Oct-08	$3.81	$342,502
Oct-09	$0.78	$70,333
Oct-10	$0.32	$29,092
Oct-11	$0.09	$7,743
Oct-12	$0.07	$6,618
Jan-13	$0.06	$5,671

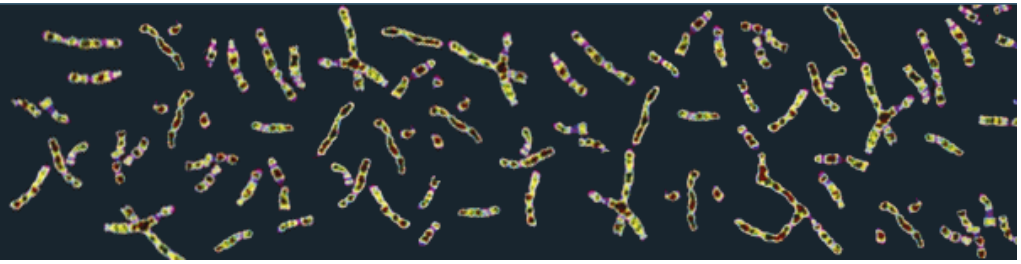




Source - NHGRI : http://www.genome.gov/sequencingcosts/

1000 Genomes

A Deep Catalog of Human Genetic Variation

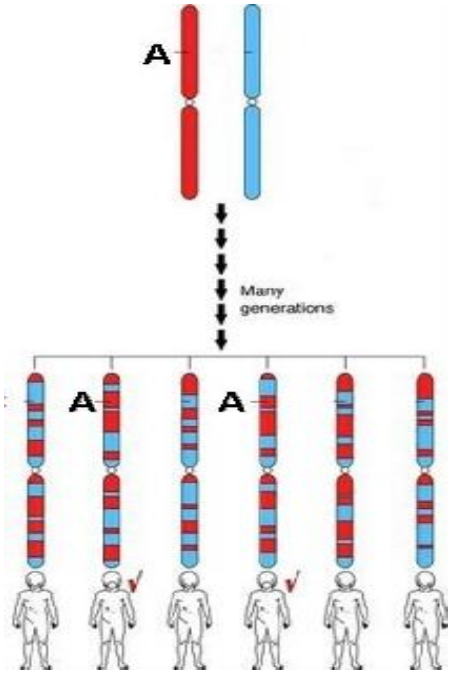


The 1000 Genomes Project is an international research consortium formed to create the most detailed and medically useful picture to date of human genetic variation. The project involves sequencing the genomes of approximately 1200 people from around the world and receives major support from the Wellcome 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).

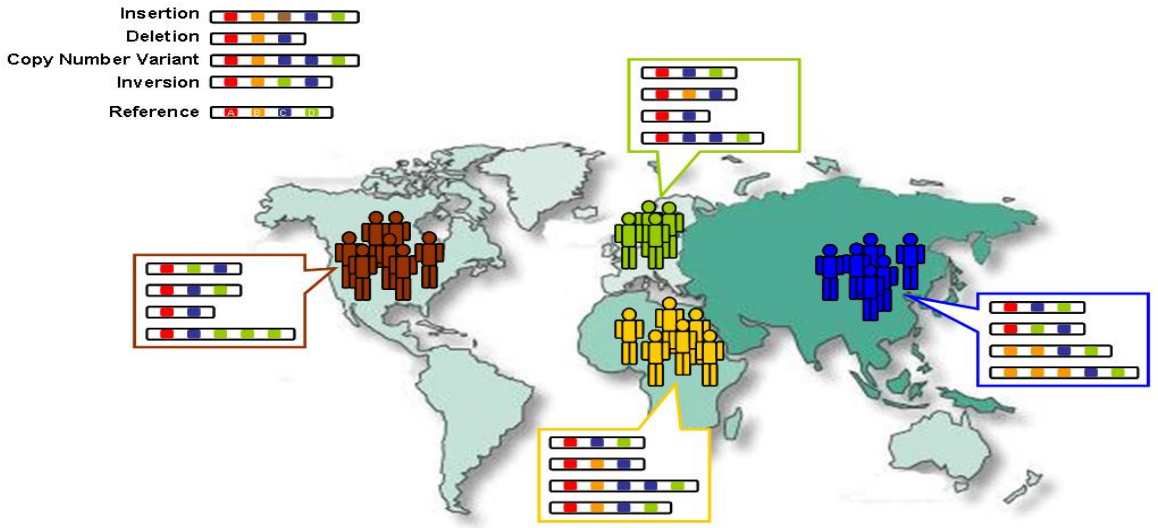
Drawing on the expertise of multidisciplinary research teams, the 1000 Genomes Project will develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data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

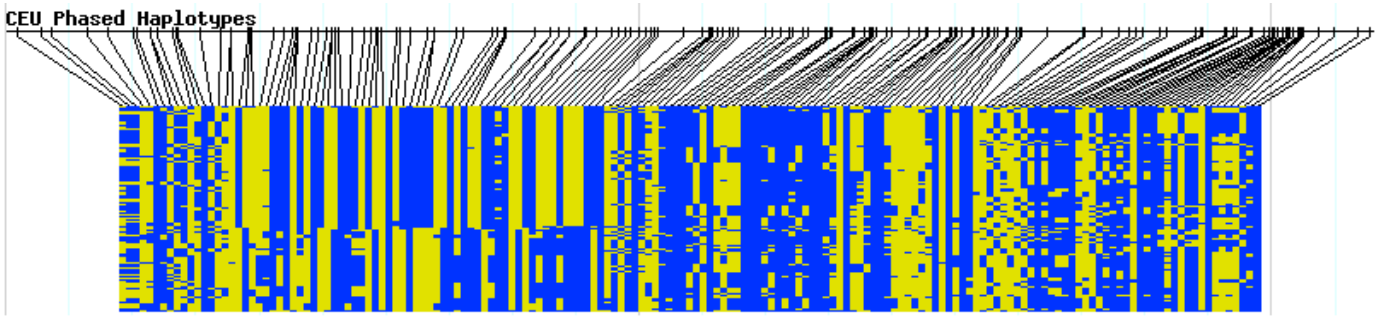
On 4 September 2008, the co-chairs of the analysis group and overall project co-chairs drafted a letter to the NIH Council about 1000 Genomes Project. This letter, available here, reviews the goals, describes the current status, and provide an update on the critical tasks the Analysis Group must accomplish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

The Scientific Goals of the 1000 Genomes Project

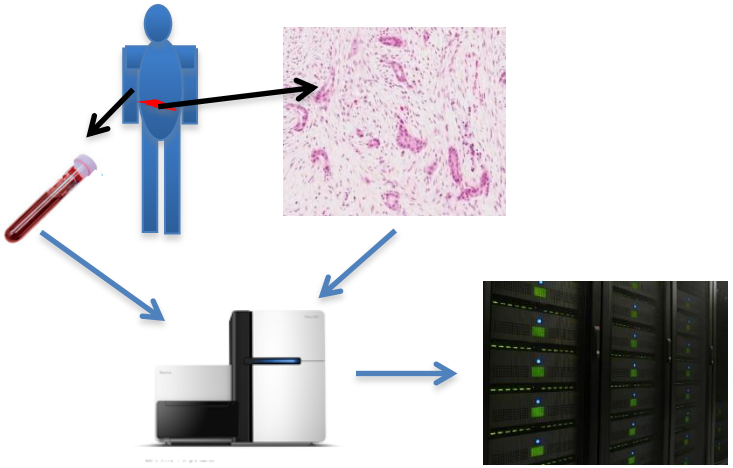


To produce a catalog of variants that are present at 1 percent or greater frequency in the human population across most of the genome, and down to 0.5 percent or lower within genes.

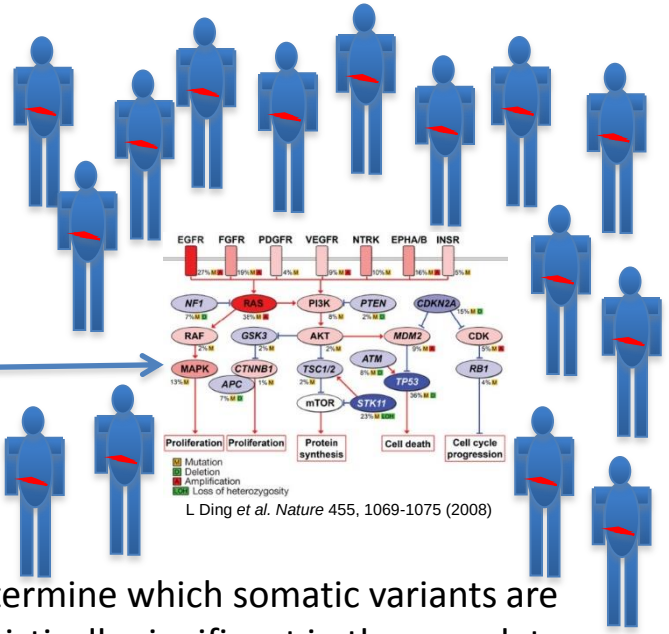




Cancer Genome Projects



Identify all genome variants associated with an individual cancer (somatic and germline) including point mutations, insertions and deletions, copy number changes, structural rearrangements



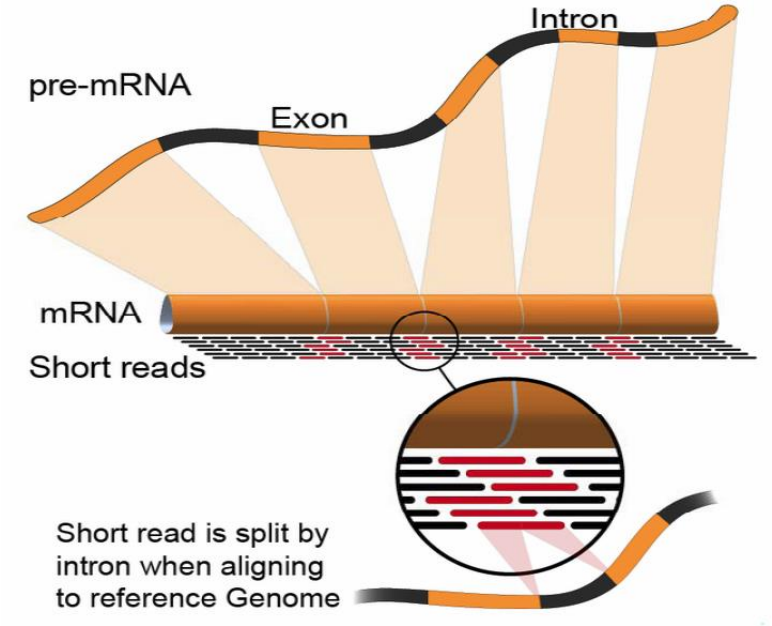
L Ding *et al. Nature* 455, 1069-1075 (2008)

Determine which somatic variants are statistically significant in the complete population of patients with that cancer type and determine which genes and pathways are essential to this tumor type.

Transcriptome sequencing (RNA-seq/R-seq)

Poly-A RNA (mRNA) sequencing (human: 1-2% of total RNA)

Ribominus RNA sequencing (human: 15% of total RNA)

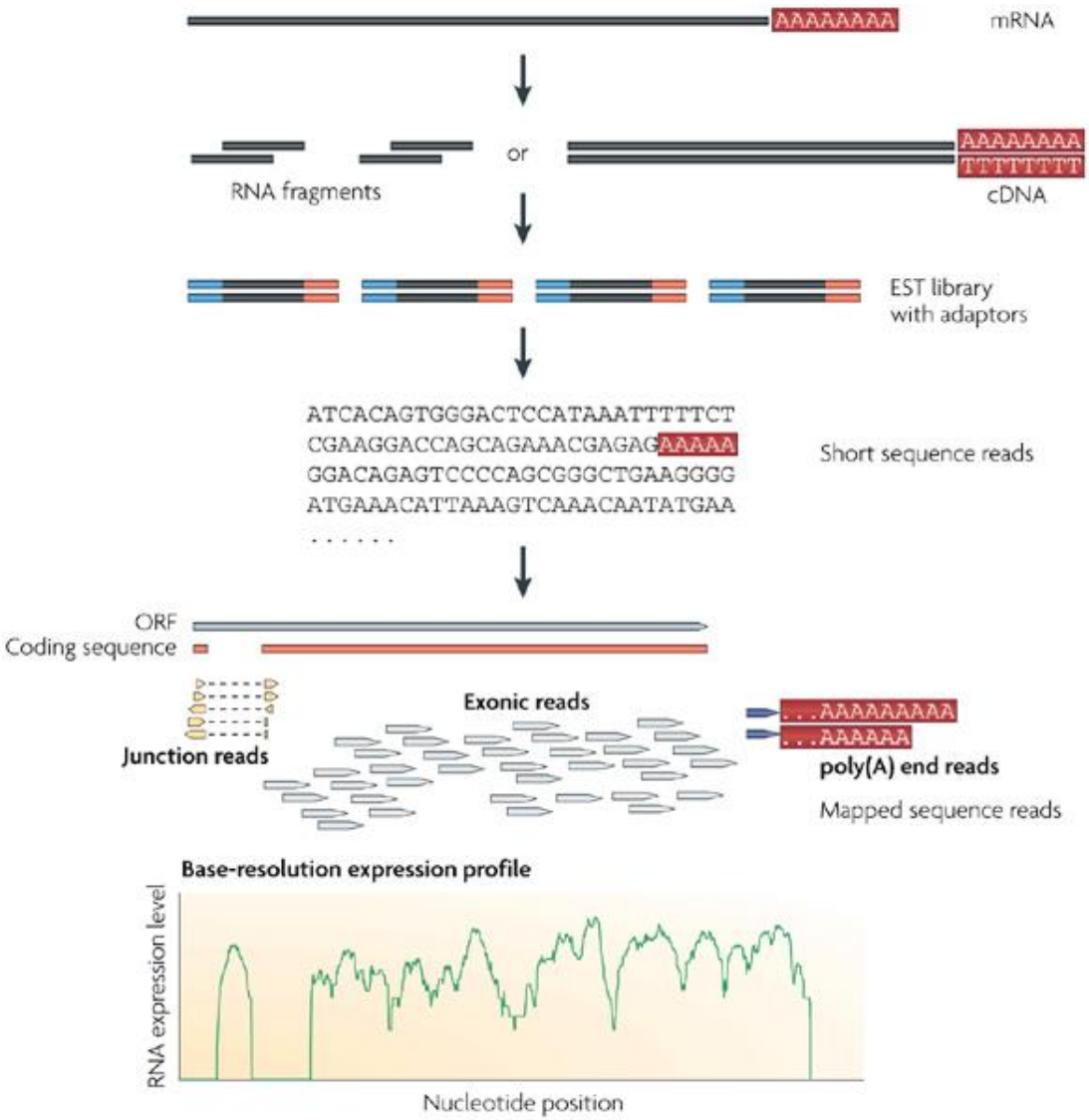
- 

1. Expression level
2. Novel transcripts
3. Alternative splicing
4. RNA editing

Expression: RPKM (reads per kilobase per million reads)

Small RNA-seq 1. expression 2. novel small RNA

Overview of RNA-Seq



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

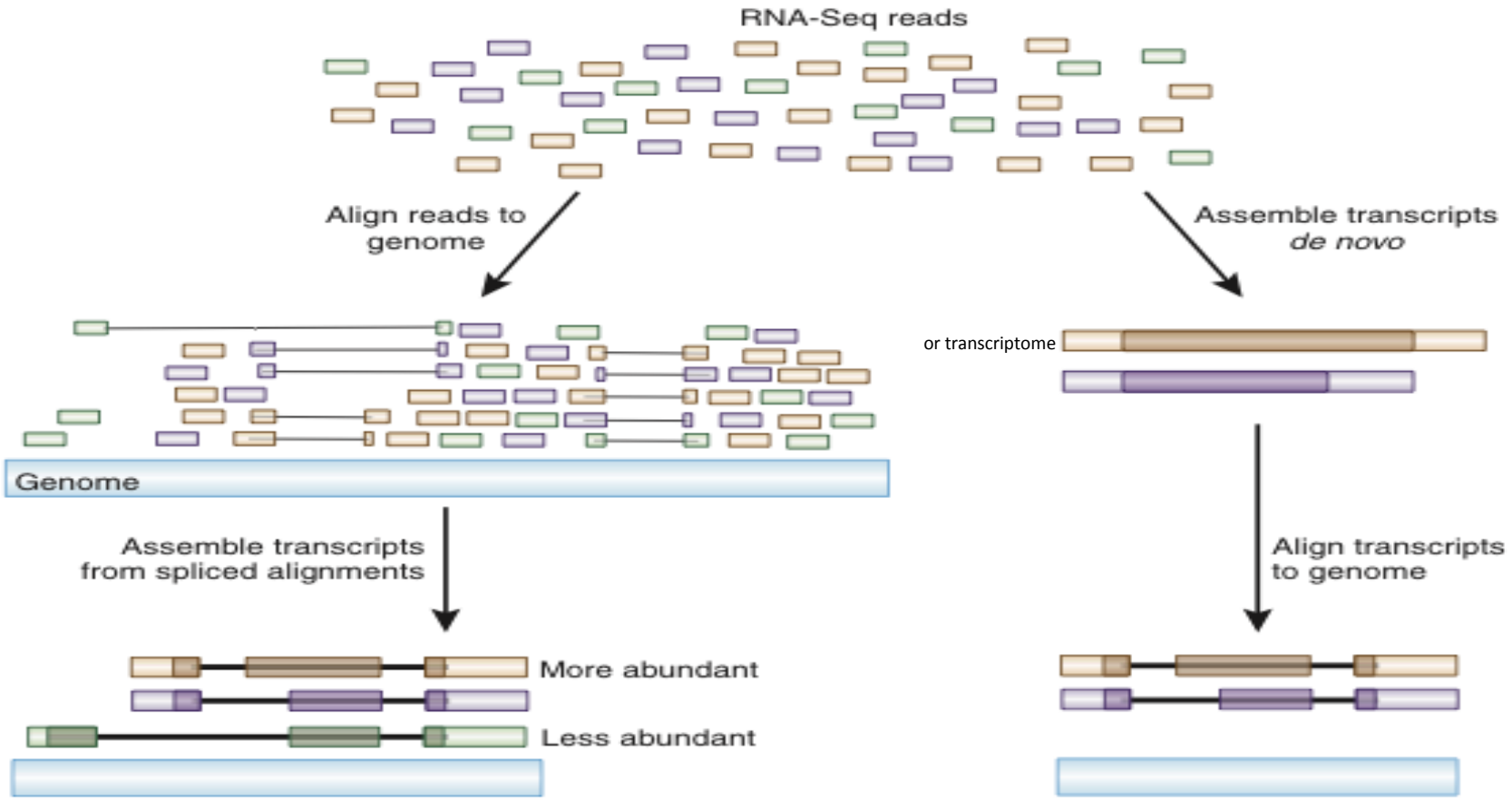
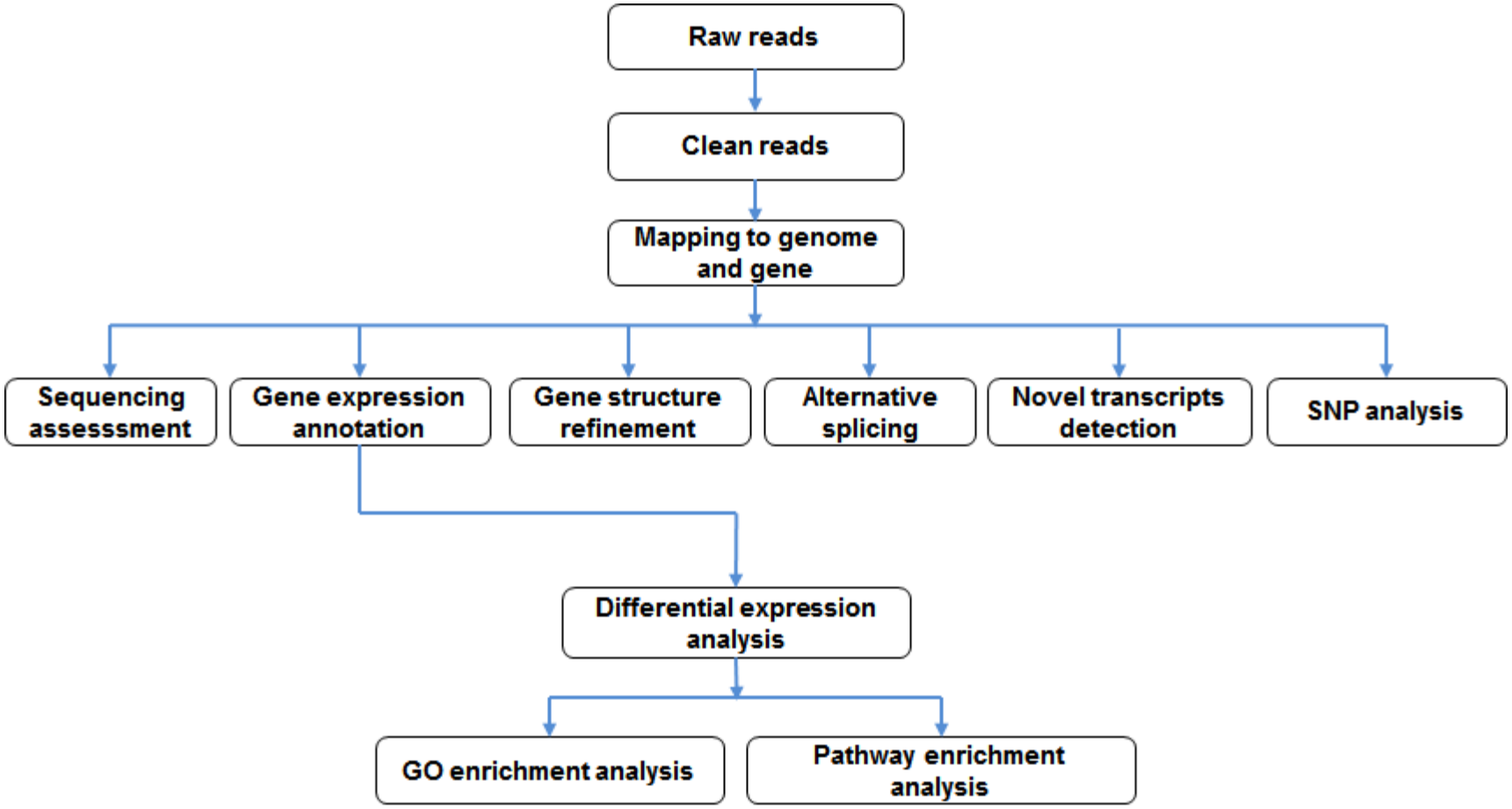


Image from Haas & Zody, 2010

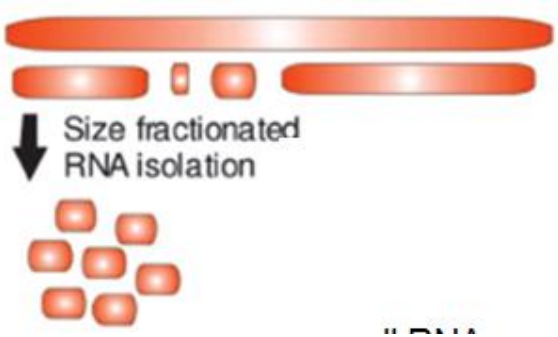
*Assembly is the only option when working with a creature with no genome sequence, alignment of contigs may be to ESTs, cDNAs etc

Bioinformatics Analysis Pipeline for Transcritpome Sequencing with Reference

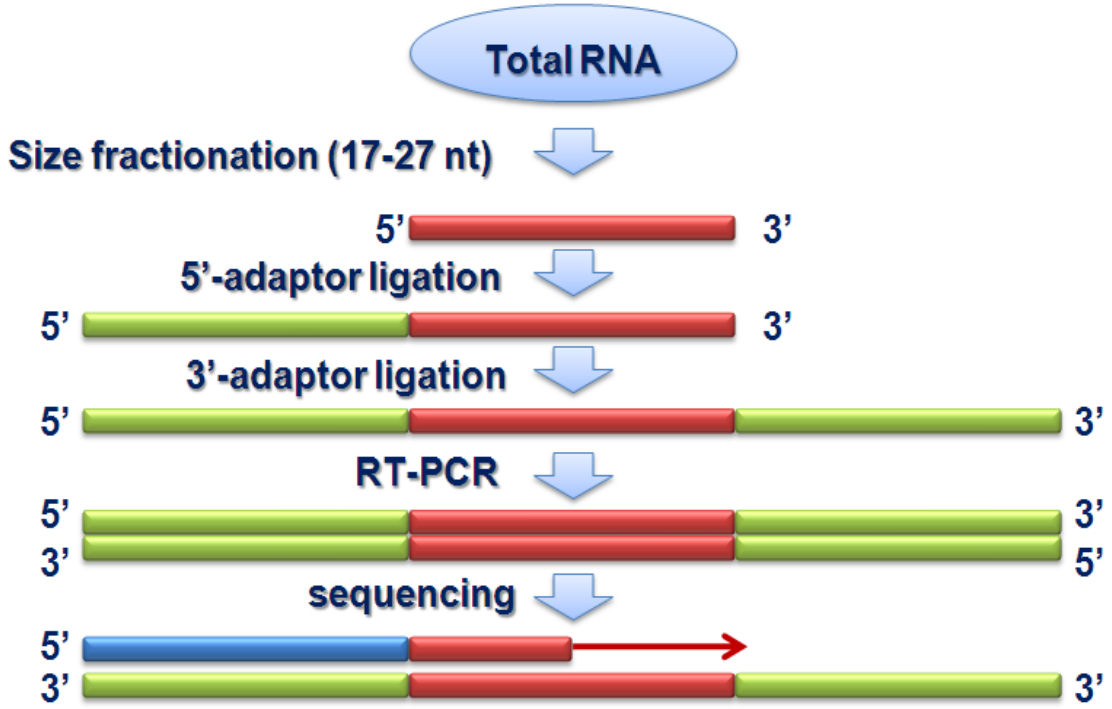


Small RNA Sequencing





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)

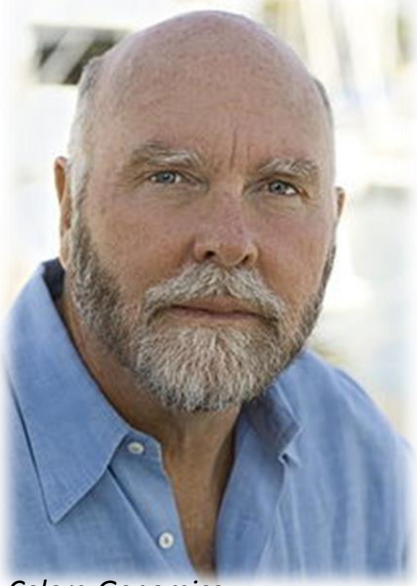
↓

Comparative miRNAomics



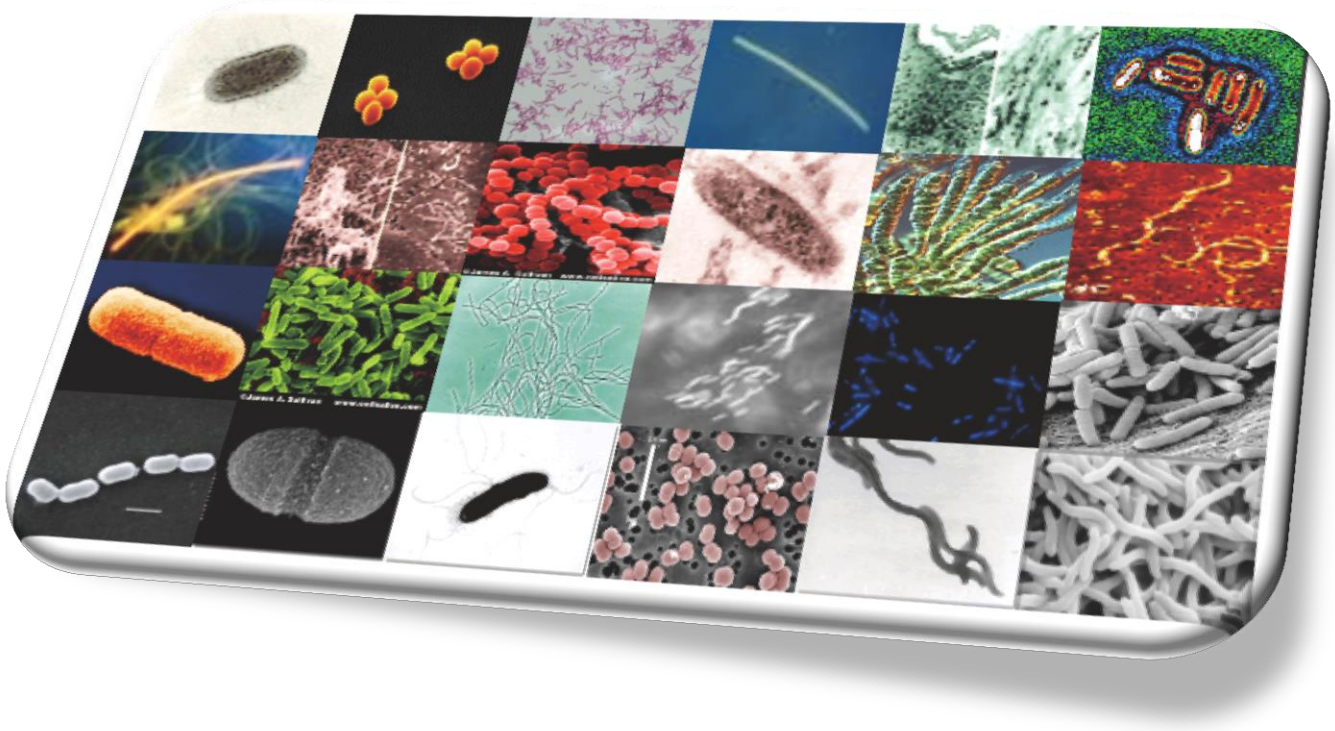
What is Metagenomics?

Craig Venter

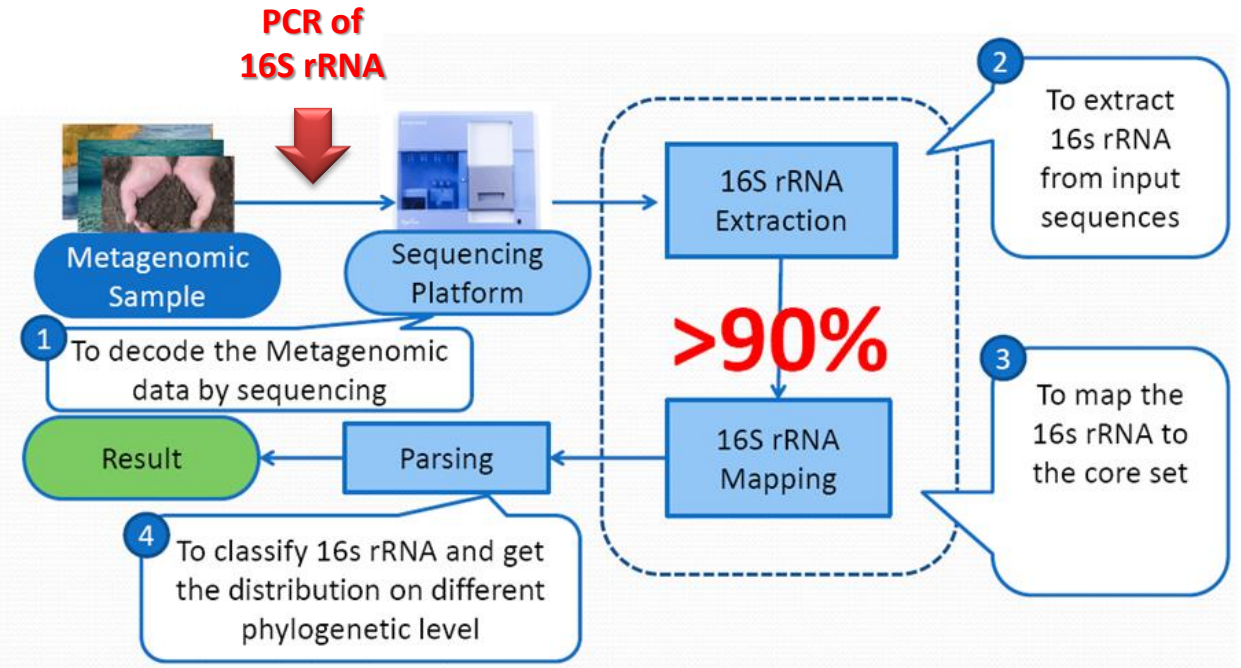


Celera Genomics
The Institute for Genomic Research
The J. Craig Venter Institute

Metagenomics is the study of metagenomes by recovering genetic materials directly from environmental or biological samples.



Classical Metagenomics Approach



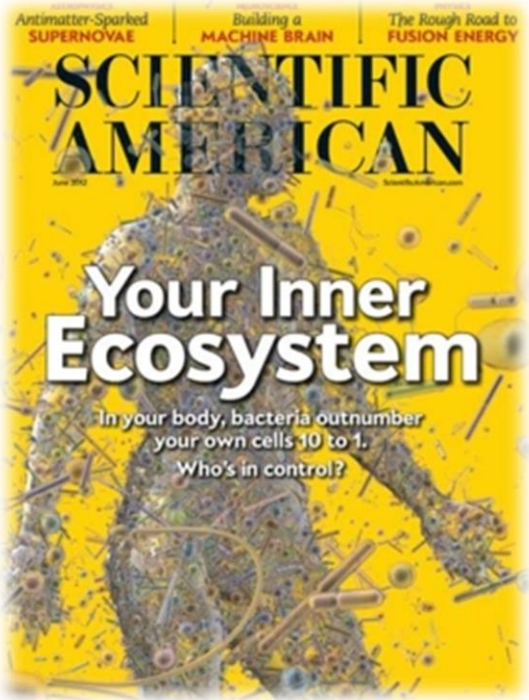
Human Microbiome Projects

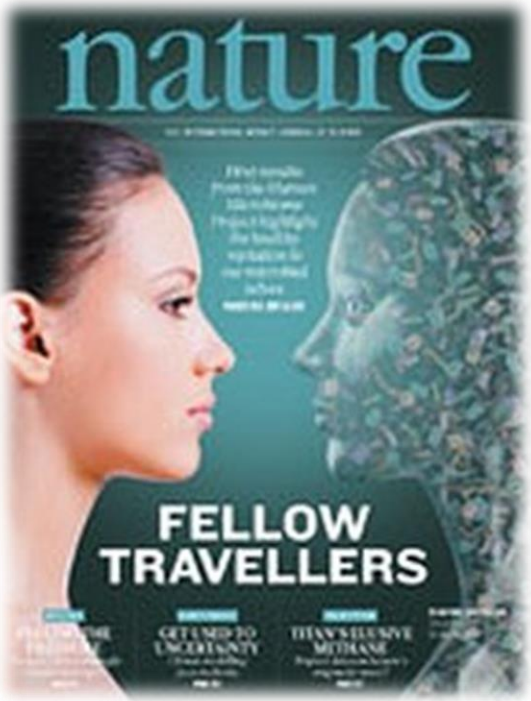
Category	HMP funded	All human associated
Airways	142	212
Blood	40	117
Eye	1	7
Gastrointestinal tract	480	686
Heart	2	2
Lymph Node	1	4
Nose	3	3
Oral	345	388
Skin	292	325
Urogenital tract	286	346
Other	504	3082
All Strains	2096	5172

- 16S rRNA sequence dependent
- 1st PCR (16S rRNA family specific primer sets)
- 2nd PCR (NGS library)
- 454 Sequencing platform
- Database update

THE HUMAN MICROBIOME PROJECT







THE HUMAN MICROBIOME

Bacteria, fungi, and viruses outnumber human cells in the body by a factor of 10 to one. The microbes synthesize key nutrients, fend off pathogens and impact everything from weight gain to perhaps even brain development. The Human Microbiome Project is doing a census of the microbes and sequencing the genomes of many. The total body count is not in but it's believed over 1,000 different species live in and on the body.



600+ SPECIES

in the mouth, pharynx and respiratory system include:

- Streptococcus viridans
- Neisseria sicca
- Candida albicans
- Streptococcus salivarius

25 SPECIES

in the stomach include:

- Helicobacter pylori
- Streptococcus thermophilus

1,000 SPECIES

in the skin include:

- Pityrosporum ovale
- Staphylococcus epidermidis
- Corynebacterium jeikeium
- Trichosporon
- Staphylococcus haemolyticus

500–1,000 SPECIES

in the intestines include:

- Lactobacillus casei
- Lactobacillus reuteri
- Lactobacillus gasseri
- Escherichia coli
- Bacteroides fragilis
- Bacteroides thetaiotaomicron
- Lactobacillus rhamnosus
- Clostridium difficile

60 SPECIES

in the urogenital tract include:

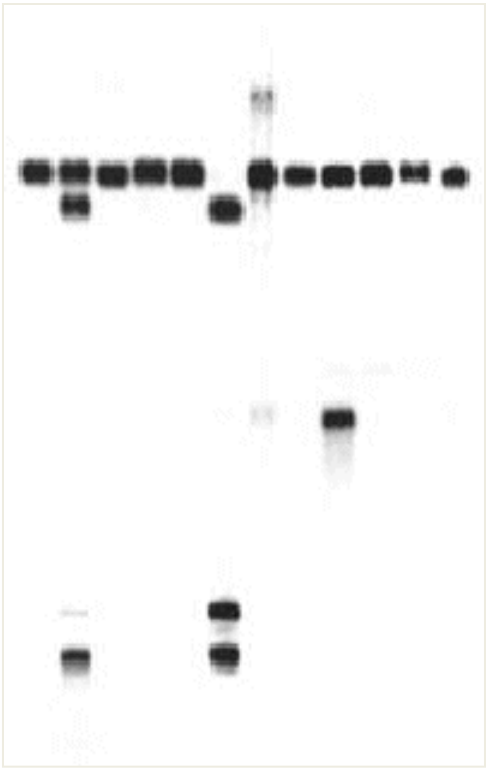
- Ureaplasma parvum
- Corynebacterium aurimucosum

SOURCES: NATIONAL INSTITUTES OF HEALTH, SCIENTIFIC AMERICAN; HUMAN MICROBIOME PROJECT

Dean Tweed • POSTMEDIA NEWS / IMAGE: Fotolia

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