


NATIONAL HUMAN GENOME RESEARCH INSTITUTE *Division of Intramural Research*




**Next-Generation Sequencing
Technologies**

Elliott H. Margulies, Ph.D.

Genome Technology Branch
National Human Genome Research Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES | NATIONAL INSTITUTES OF HEALTH | genome.gov/DIR

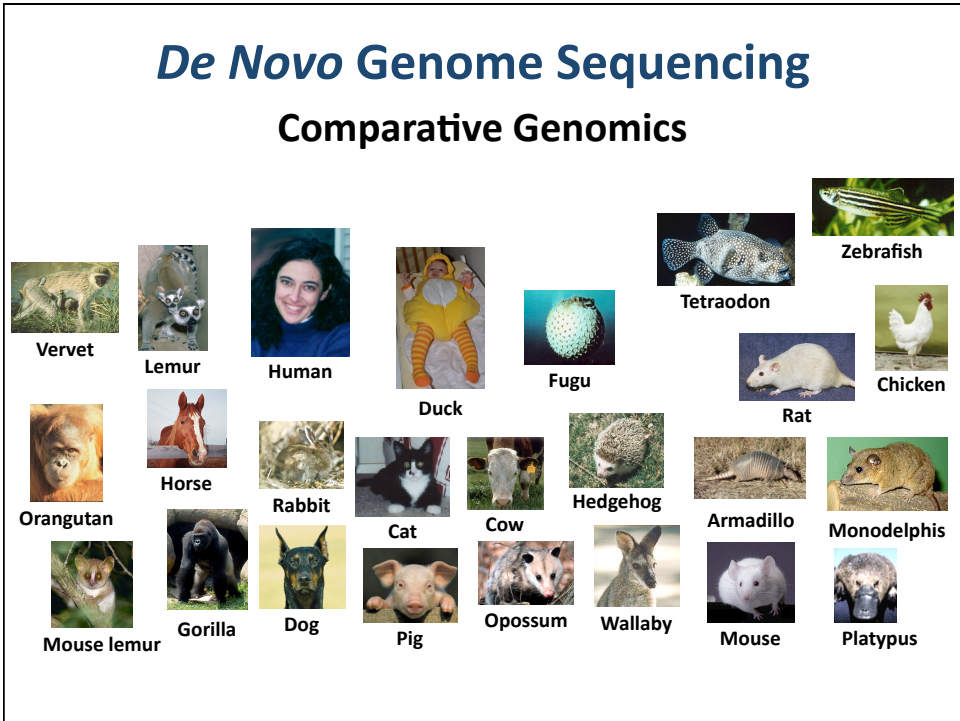


Overview

Background

Technologies

Applications



Variation Detection

nature

Vol 456 | 6 November 2008 | doi:10.1038/nature07485

ARTICLES

DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome

Timothy J. Ley^{1,2,3,4*}, Elaine Brian H. Dunford-Shore⁵, St Dan C. Koboldt¹, Craig Poh Tracie Miner², Lucinda Fult Nathan Sander², Xiaoli Shi Rhonda E. Ries¹, Jacqueline Jennifer Ivanovich^{1,7}, Sharo Daniel C. Link^{1,4}, Timothy A

doi:10.1038/nature08658

nature

ARTICLES

A comprehensive catalogue of somatic mutations from a human cancer genome

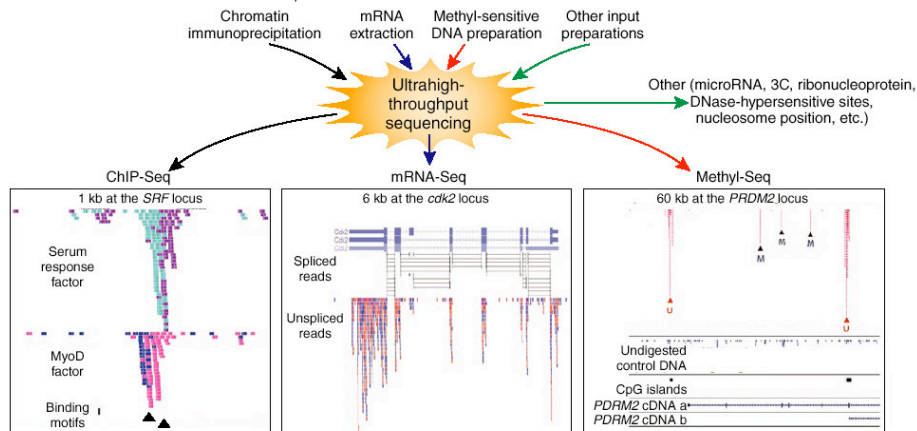
Erin D. Pleasance^{1*}, R. Keira Cheetham^{2*}, Philip J. Stephens¹, David J. McBride¹, Sean J. Humphray², Chris D. Greenman¹, Ignacio Varela¹, Meng-Lay Lin¹, Gonzalo R. Ordóñez, Graham R. Bignell¹, Kai Ye¹, Julie Alipaz⁴, Markus J. Bauer², David Beare¹, Adam Butler¹, Richard J. Carter², Lina Chen¹, Anthony J. Cox², Sarah Edkins¹, Paula I. Kokko-Gonzales², Niall A. Gormley², Russell J. Grocock², Christian D. Haudenschild², Matthew M. Hims², Terena James², Mingming Jia¹, Zoya Kingsbury², Catherine Leroy¹, John Marshall¹, Andrew Menzies¹, Laura J. Mudie¹, Zemin Ning¹, Tom Royce¹, Ole B. Schulz-Trieglaff², Anastassia Spiridou², Lucy A. Stebbings¹, Lukasz Szajkowski², Jon Teague¹, David Williamson², Lynda Chin², Mark T. Ross², Peter J. Campbell¹, David R. Bentley², P. Andrew Futreal¹ & Michael R. Stratton^{1,7}

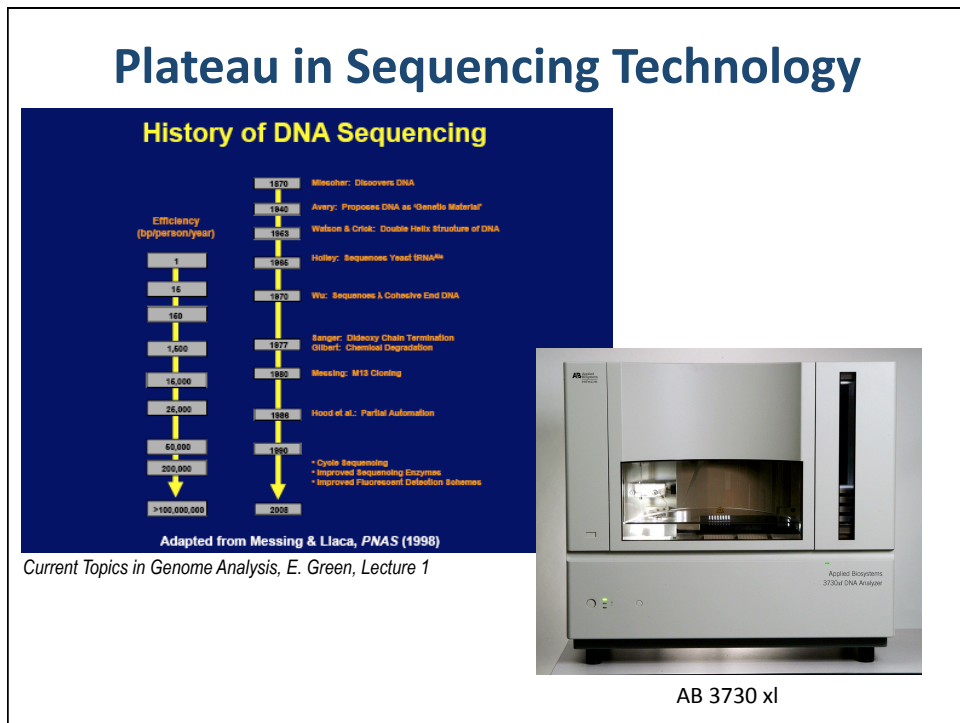
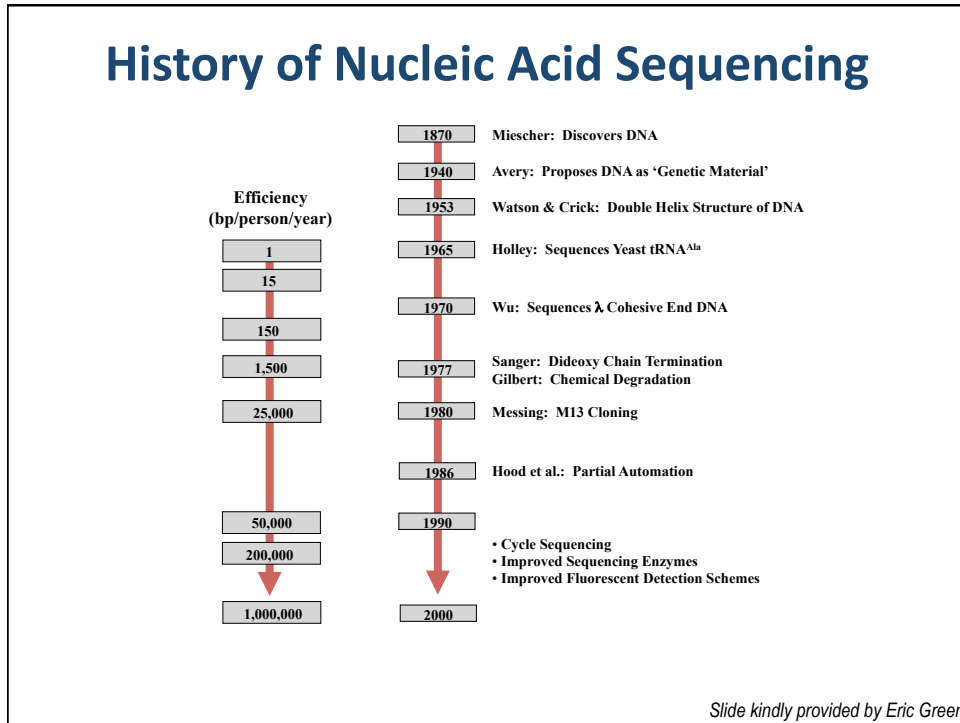
“Counting Experiments”

NATURE METHODS | VOL.5 NO.1 | JANUARY 2008 | 19

Sequence census methods for functional genomics

Barbara Wold & Richard M Myers

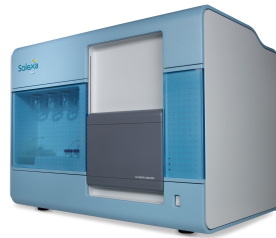




New Sequencing Technologies



454
Pyrosequencing



Genome Analyzer
Reversible Terminator Chemistry



Applied Biosystems
SOLiD
Ligation-based extension

Even Newer than New...

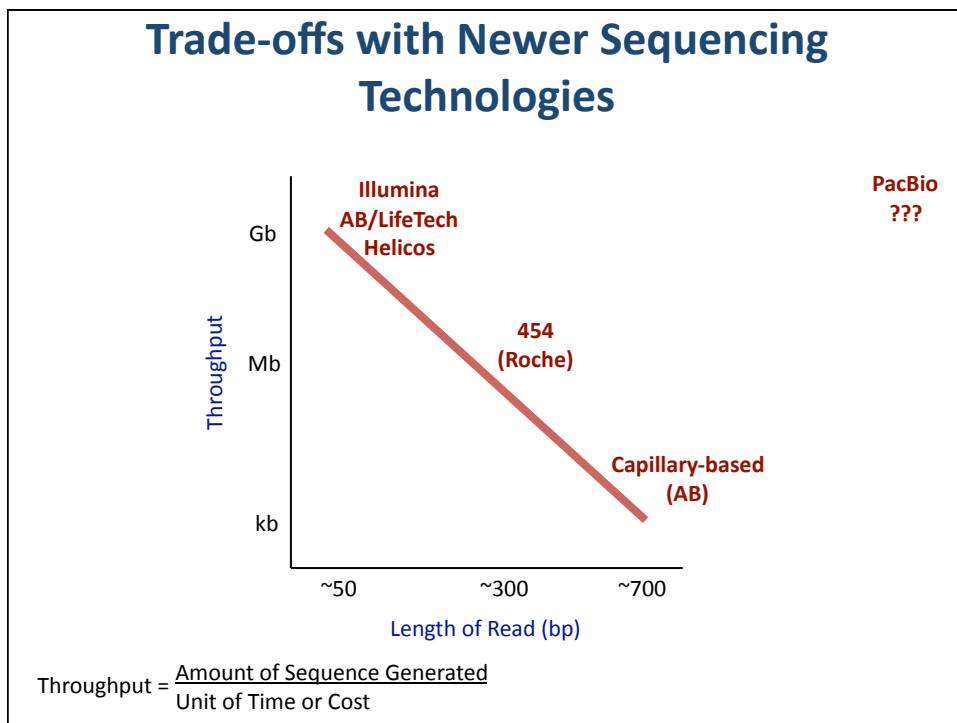
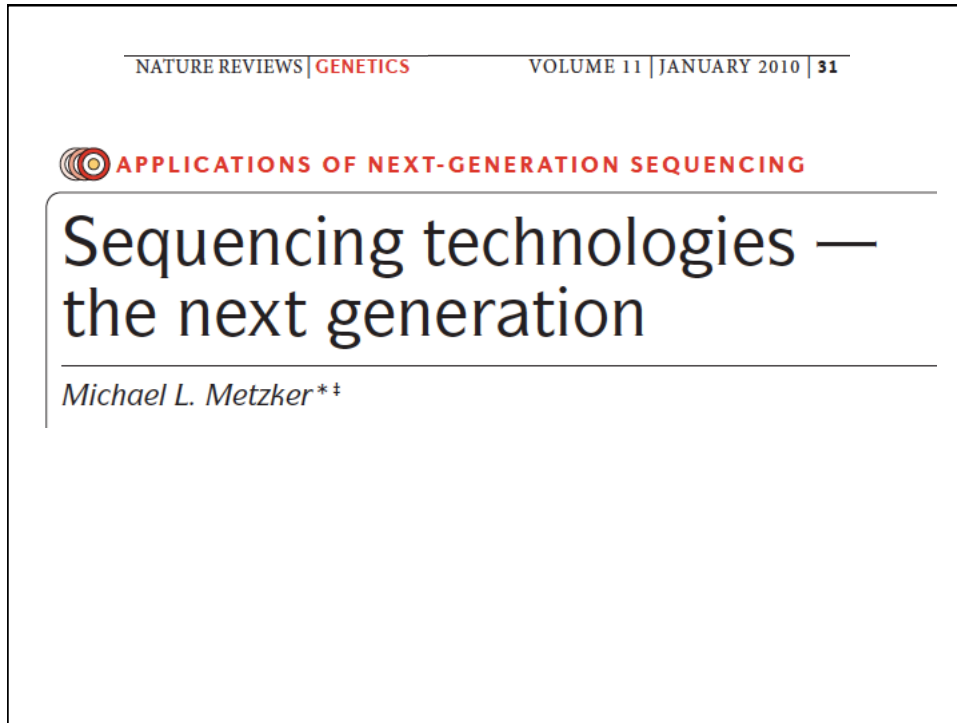
True Single Molecule Sequencing



Helicos
HeliScope



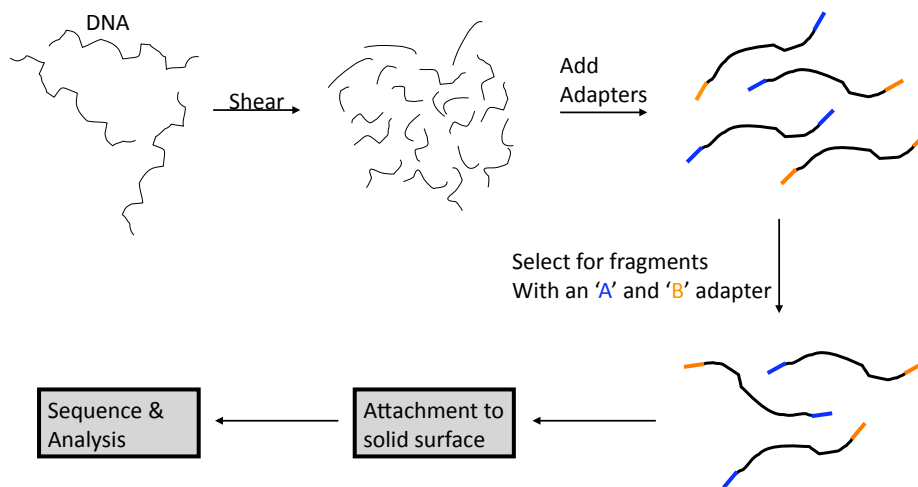
PACIFIC
BIOSCIENCES
SMRT Technology

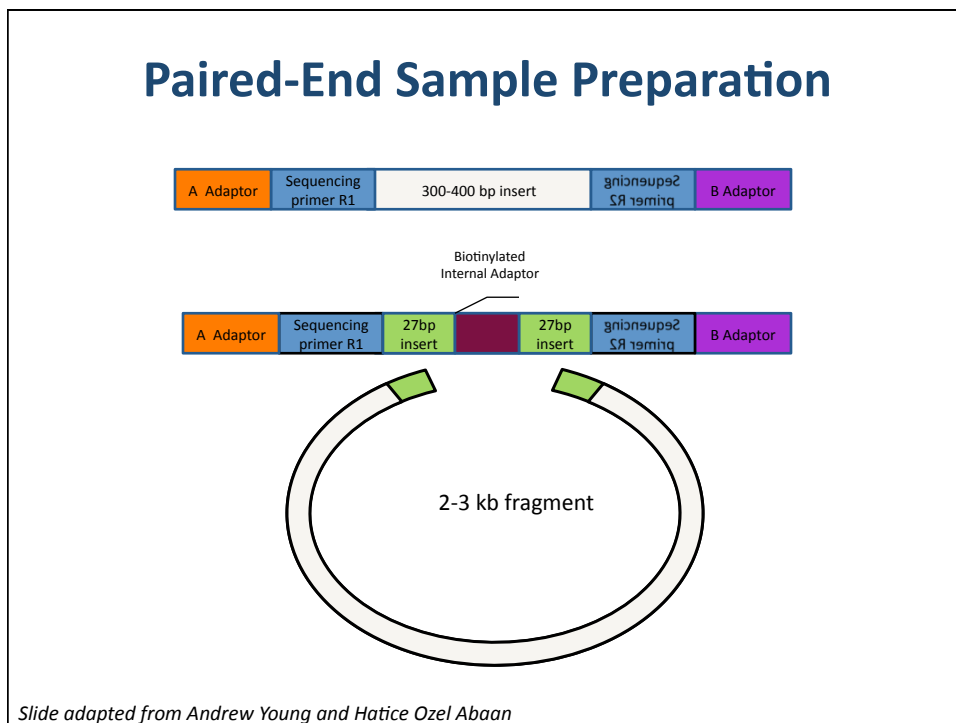
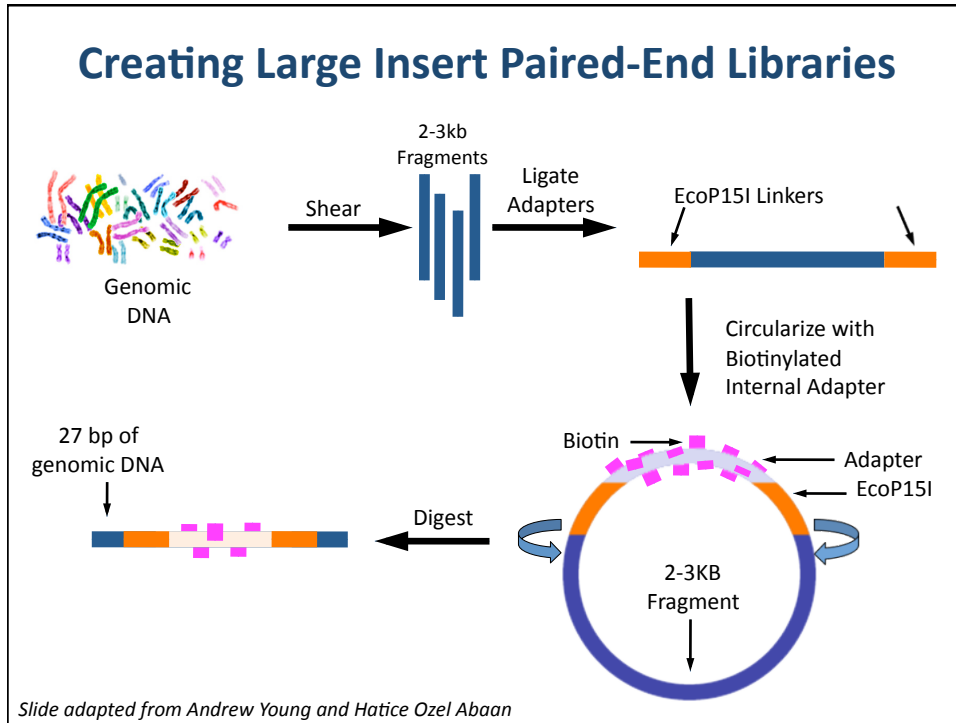


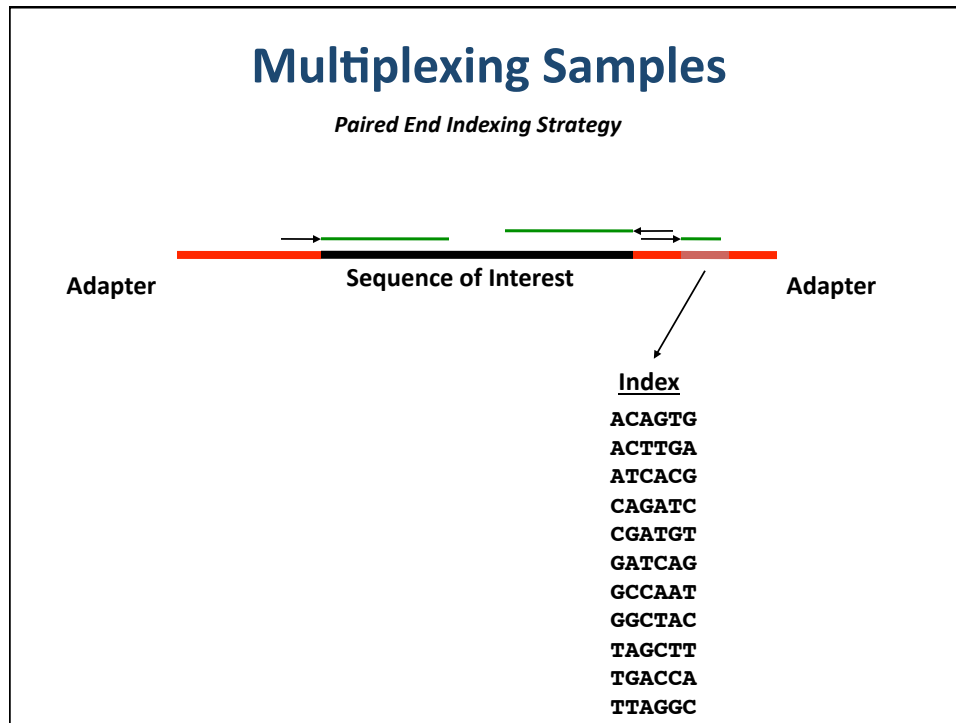
Template/Library Preparation Methods

“Single Molecule” Sequencing


Really a clonal amplification of a single DNA molecule







454 Sequencing Technology



doi:10.1038/nature03959 Nature 31st July 2005 nature

ARTICLES

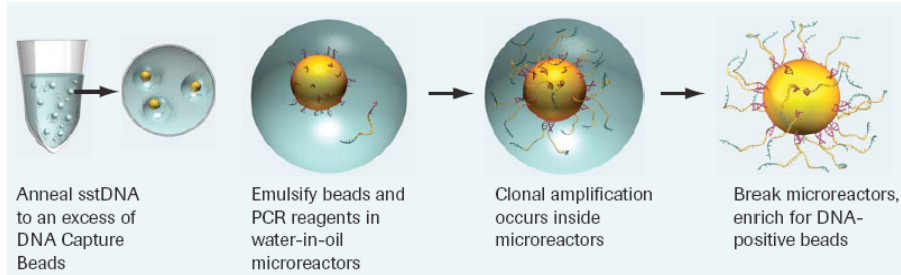
Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies^{1*}, Michael Egholm^{1*}, William E. Altman¹, Said Attiya¹, Joel S. Bader¹, Lisa A. Bembgen¹, Jan Berka¹, Michael S. Braverman¹, Yi-Ju Chen¹, Zhoutao Chen¹, Scott B. Dewell¹, Lei Du¹, Joseph M. Fierro¹, Xavier V. Gomes¹, Brian C. Godwin¹, Wen He¹, Scott Helgesen¹, Chun He Ho¹, Gerard P. Irzyk¹, Szilveszter C. Jando¹, Maria L. I. Alenquer¹, Thomas P. Jarvie¹, Kshama B. Jirage¹, Jong-Bum Kim¹, James R. Knight¹, Janna R. Lanza¹, John H. Leamon¹, Steven M. Lefkowitz¹, Ming Lei¹, Jing Li¹, Kenton L. Lohman¹, Hong Lu¹, Vinod B. Makhijani¹, Keith E. McDade¹, Michael P. McKenna¹, Eugene W. Myers², Elizabeth Nickerson¹, John R. Nobile¹, Ramona Plant¹, Bernard P. Puc¹, Michael T. Ronan¹, George T. Roth¹, Gary J. Sarkis¹, Jan Fredrik Simons¹, John W. Simpson¹, Mithreyan Srinivasan¹, Karrie R. Tartaro¹, Alexander Tomasz¹, Kari A. Vogt¹, Greg A. Volkmer¹, Shally H. Wang¹, Yong Wang¹, Michael P. Weiner¹, Pengguang Yu¹, Richard F. Begley¹ & Jonathan M. Rothberg²

454 LIFE SCIENCES

Slide (though slightly modified) courtesy of Elaine Mardis, Wash U., St. Louis

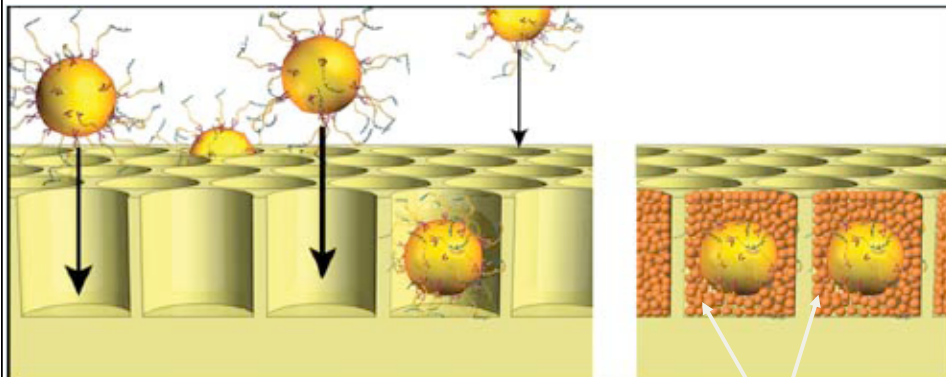
Emulsion PCR (Template Prep)



Each bubble in the emulsion will potentially contain a different fragment.

Slide Courtesy of Alice Young, NISC

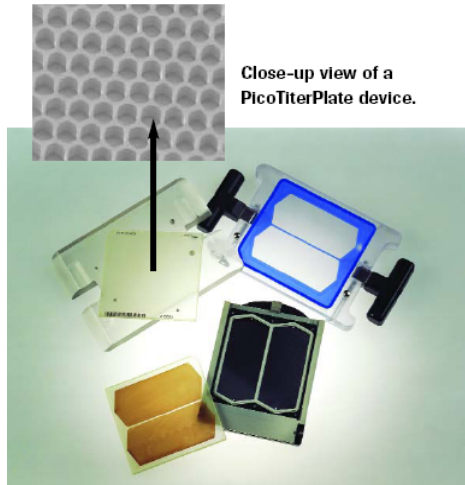
Load PicoTiter Plate



Packing beads and enzyme beads

Slide Courtesy of Alice Young, NISC

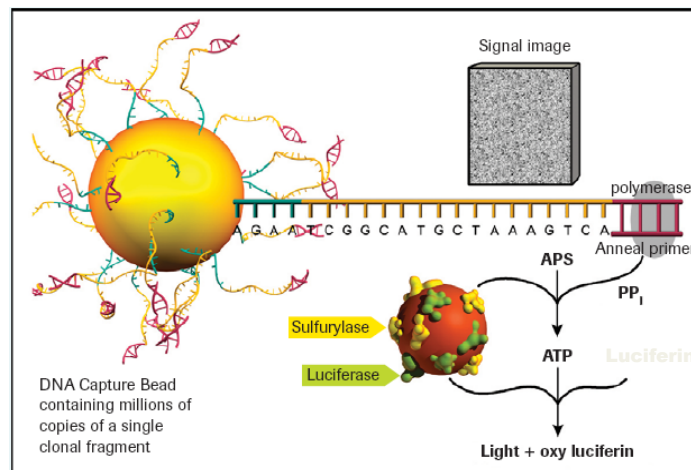
PicoTiter Plate Apparatus



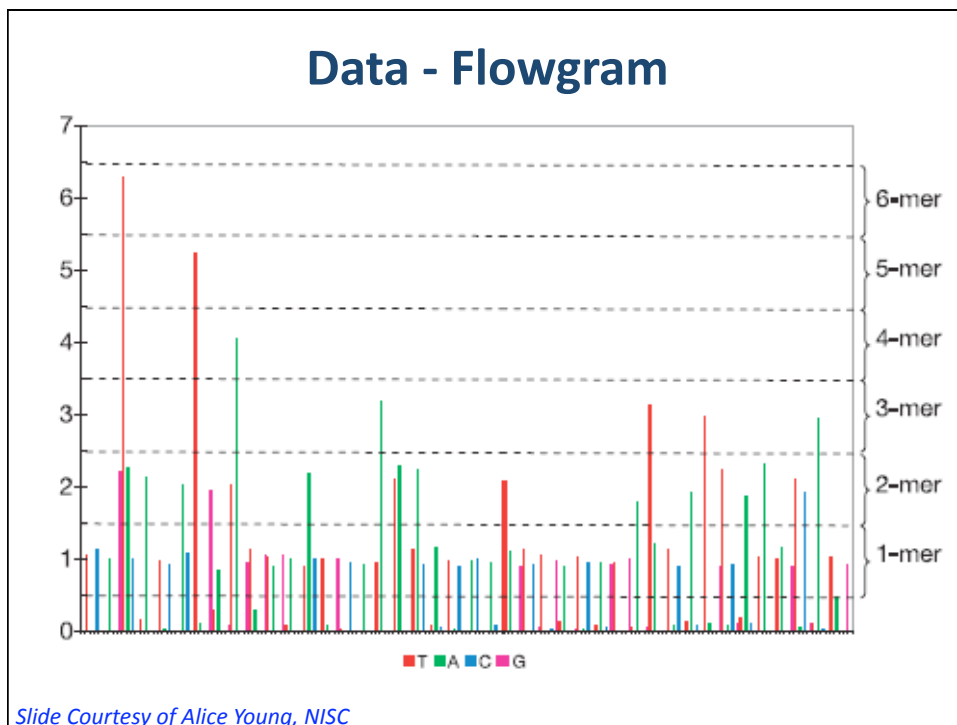
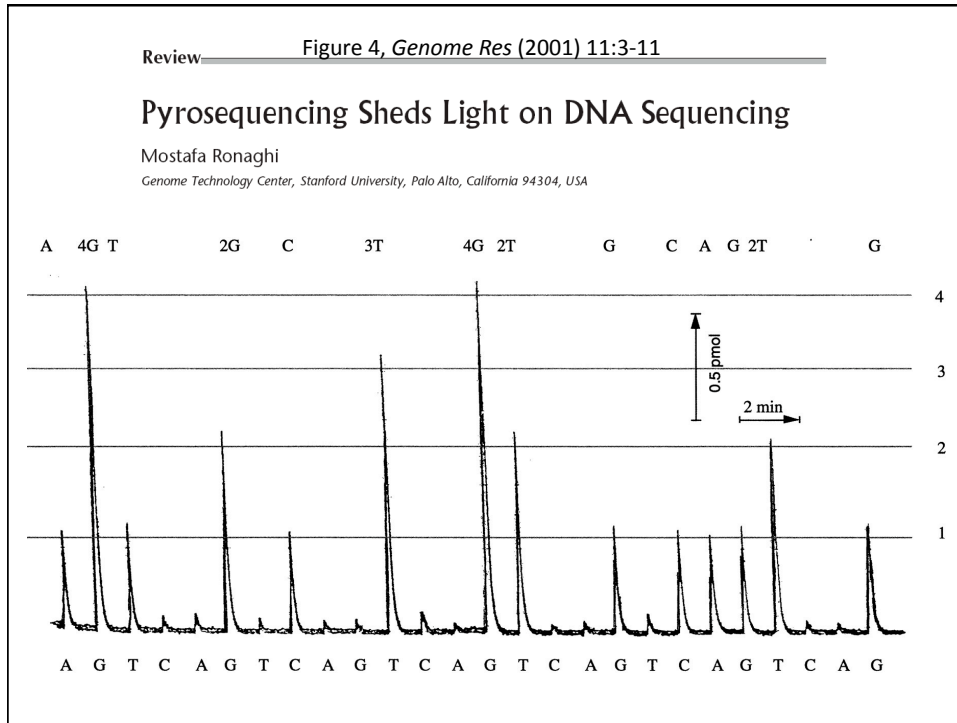
Instead of 96 reads/run, there are hundreds of thousands.

Slide Courtesy of Alice Young, NISC

PyroSequencing



Slide Courtesy of Alice Young, NISC



454 Sequencing Summary

- Run time ~8 hrs
- Produces 100's of Mb of sequence
- Read length ~300-400 bp
- Most "mature" of the next-generation technologies
- Homopolymer runs can be an issue

Applications:

- *de novo* sequencing
- Variation detection
- Gene Expression
- "Metagenomics"



ARTICLES

Nature, 2006 November 16; vol. (7117), 444 330-336

Analysis of one million base pairs of Neanderthal DNA

Richard E. Green¹, Johannes Krause¹, Susan E. Ptak¹, Adrian W. Briggs¹, Michael T. Ronan², Jan F. Simons², Lei Du², Michael Egholm², Jonathan M. Rothberg², Maja Paunovic^{3,4} & Svante Pääbo¹

Science, 2006 November 17; vol. 314, 1113-111

Sequencing and Analysis of Neanderthal Genomic DNA

James P. Noonan,^{1,2} Graham Coop,³ Sridhar Kudaravalli,³ Doug Smith,¹ Johannes Krause,¹ Joe Alessi,¹ Feng Chen,¹ Darren Platt,¹ Svante Pääbo,⁴ Jonathan K. Pritchard,³ Edward M. Rubin^{1,2*}



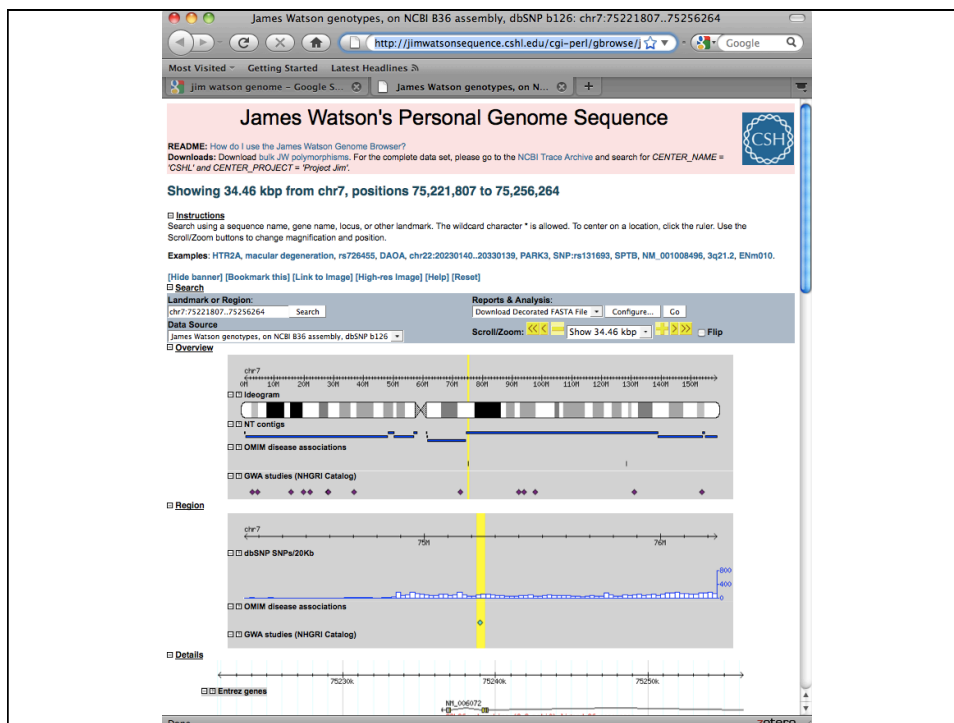
<http://popsci.typepad.com/photos/uncategorized/2007/10/25/laluezafox11r.jpg>

nature Vol 452 | 7 April 2008 | doi:10.1038/nature06884

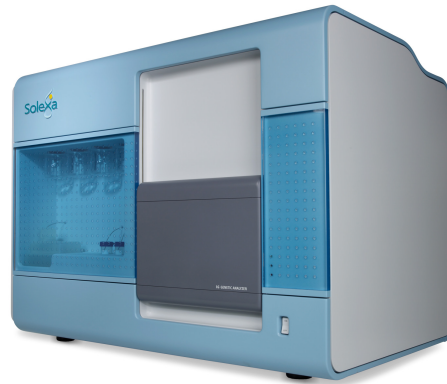
LETTERS

The complete genome of an individual by massively parallel DNA sequencing

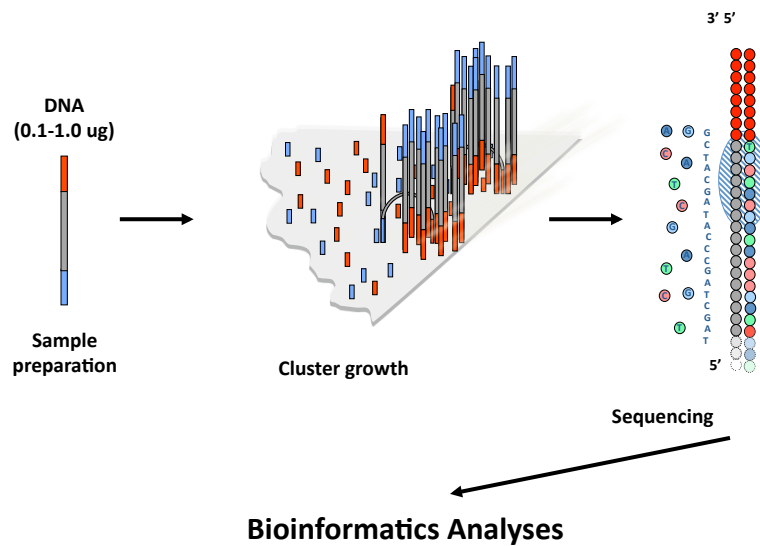
David A. Wheeler^{1*}, Maithreyan Srinivasan^{2*}, Michael Egholm^{2*}, Yufeng Shen^{1*}, Lei Chen¹, Amy McGuire³, Wen He², Yi-Ju Chen², Vinod Makhijani², G. Thomas Roth², Xavier Gomes², Karrie Tartaro²†, Faheem Niazi², Cynthia L. Turcotte², Gerard P. Irzyk², James R. Lupski^{4,5,6}, Craig Chinault⁴, Xing-zhi Song¹, Yue Liu¹, Ye Yuan¹, Lynne Nazareth¹, Xiang Qin¹, Donna M. Muzny¹, Marcel Margulies², George M. Weinstock^{1,4}, Richard A. Gibbs^{1,4} & Jonathan M. Rothberg²†



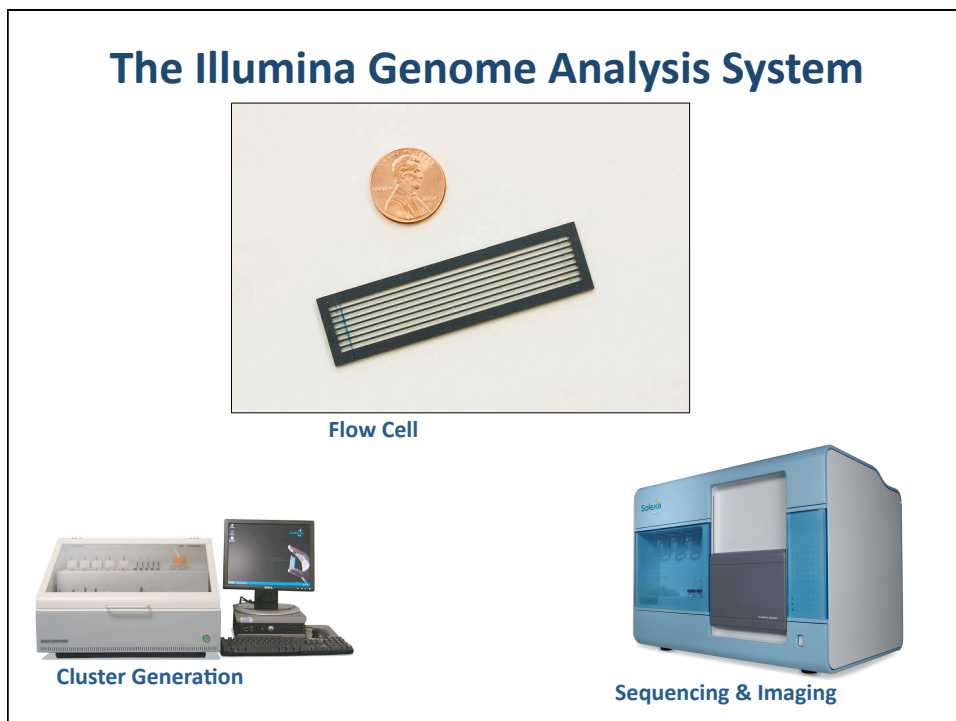
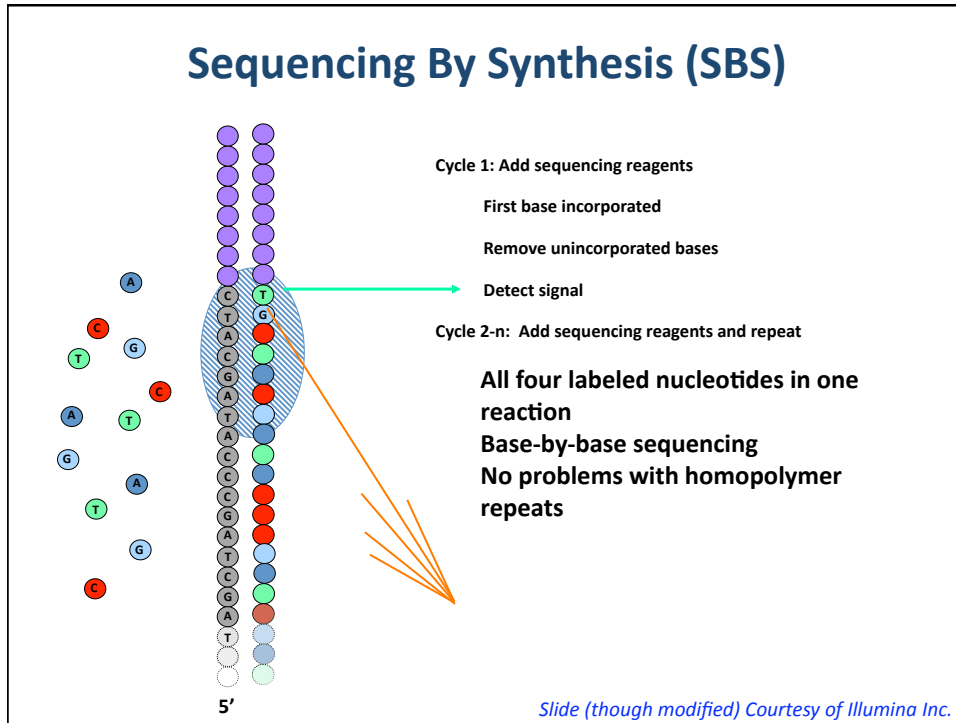
Illumina Genome Analyzer

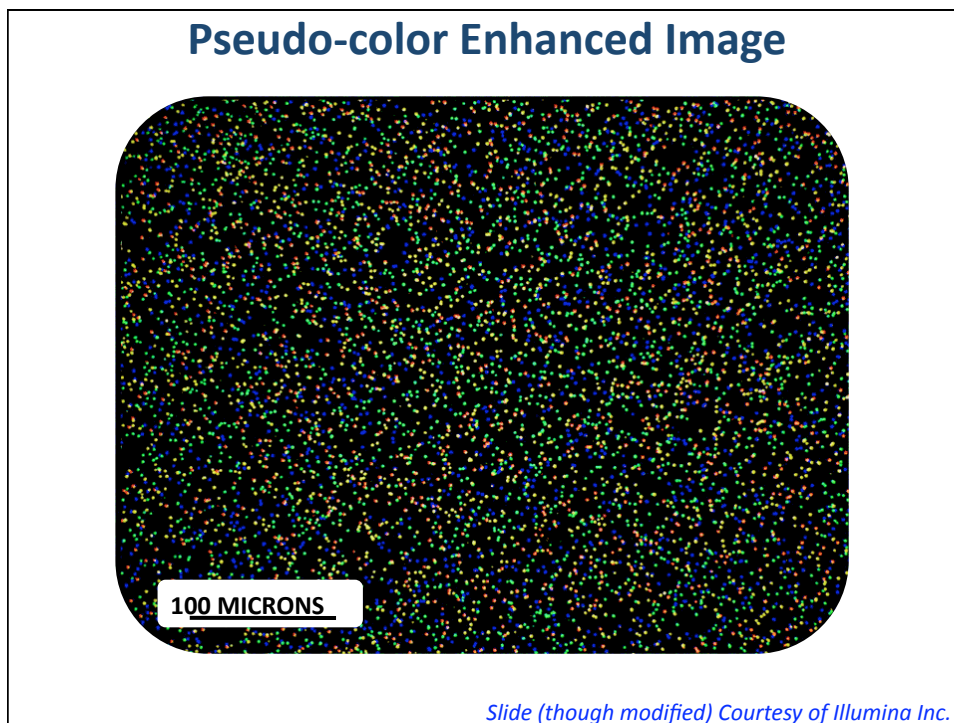
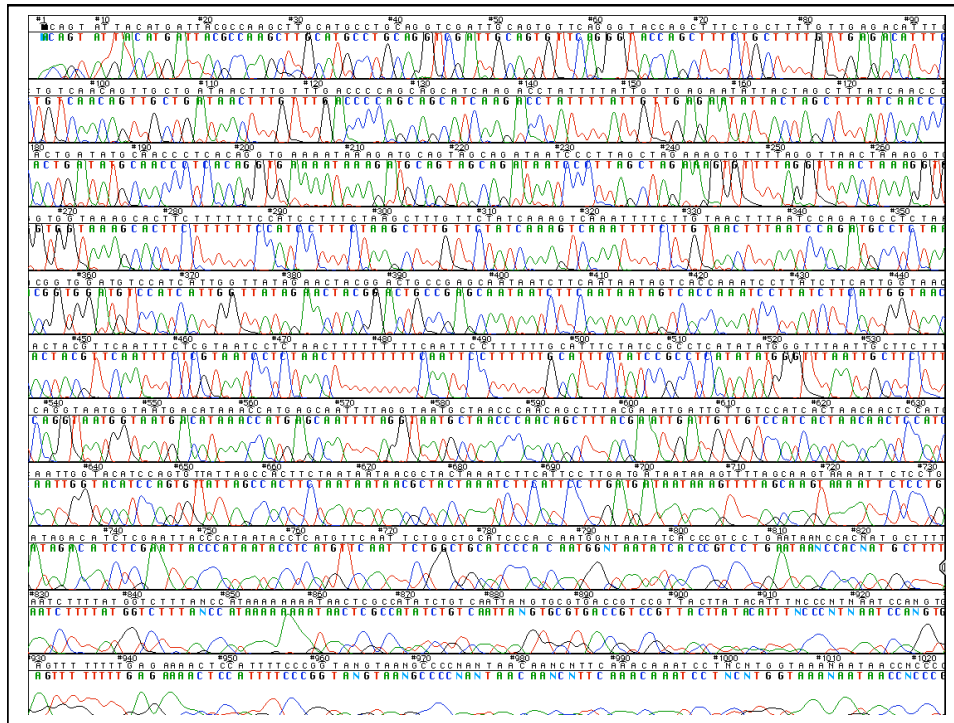


Illumina/Solexa Sequencing

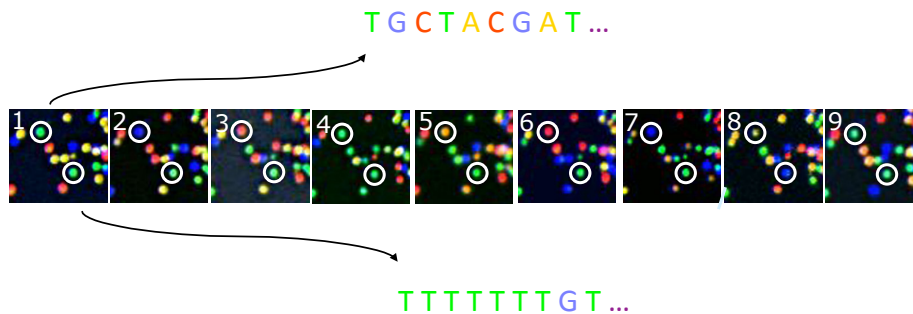


Slide Courtesy of Illumina Inc.





Base Calling from Raw Data



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

Slide (though modified) Courtesy of Illumina Inc.

Illumina Throughput

- **Each lane can sequence 20-30 million molecules**
 - 8 lanes = up to 240 million reads
- **36 bp reads suitable for counting based experiments**
- **Capable of up to 100bp paired end reads**
 - 50 Gigabases of sequence per run



HiSeq2000

- **Same chemistry**
- **Runs 2 flowcells at the same time**
 - Imaging one flowcell – chemistry on the other
- **Flowcells are bigger**
 - More surface area can be scanned
 - Focuses on top and bottom of flowcell
- **Improvements to hardware**
 - Better lasers, cameras, etc.
- **Initial release mid-February at 100-125G per flowcell**
- **8 day runtime for 2 flowcells**
 - Two “whole genomes” in 8 days!

SOLiD from Applied Biosystems (now Life Technologies)



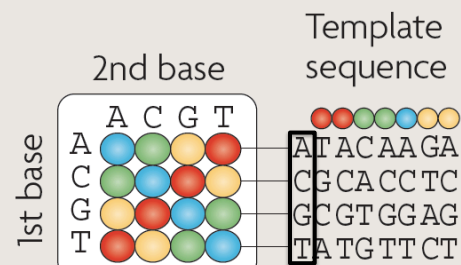
Two-base encoding

1,2-probes

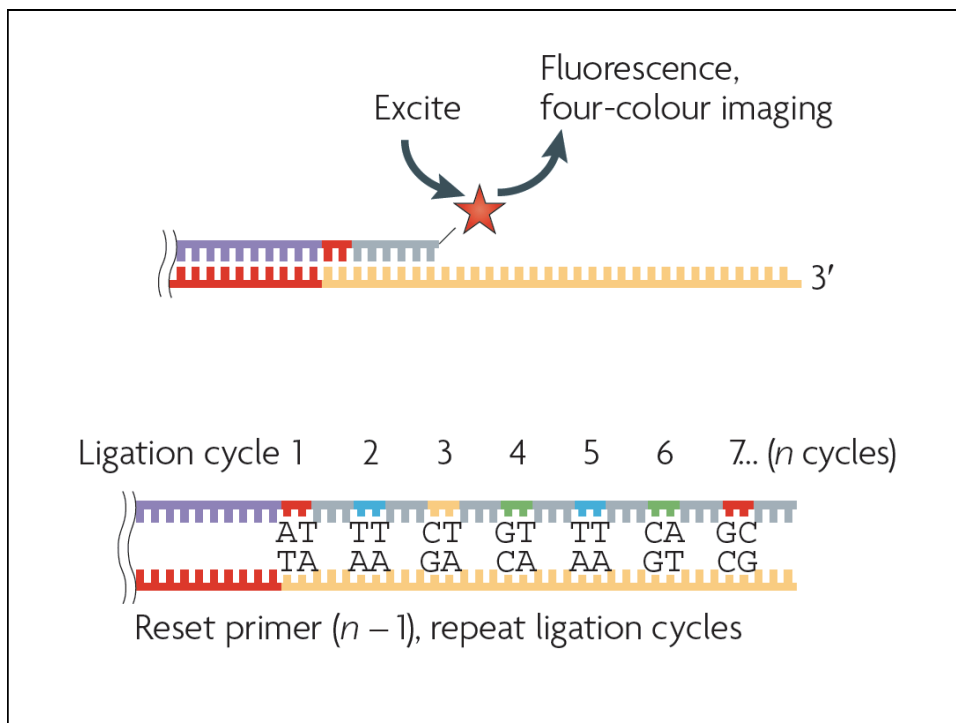
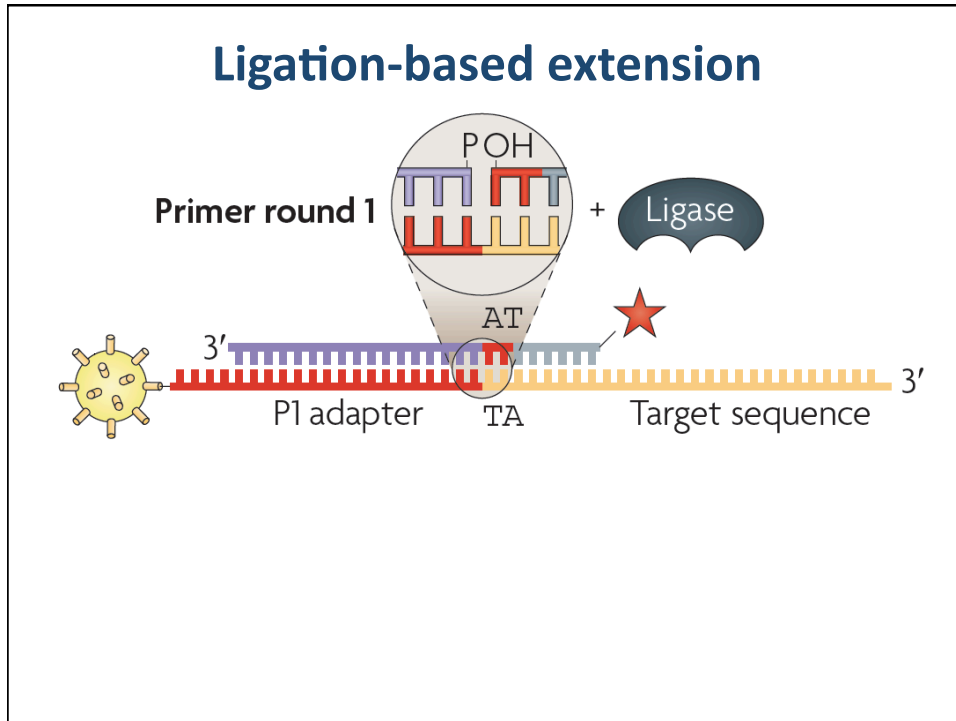
- x, y Interrogation bases
- n Degenerate bases
- z Universal bases

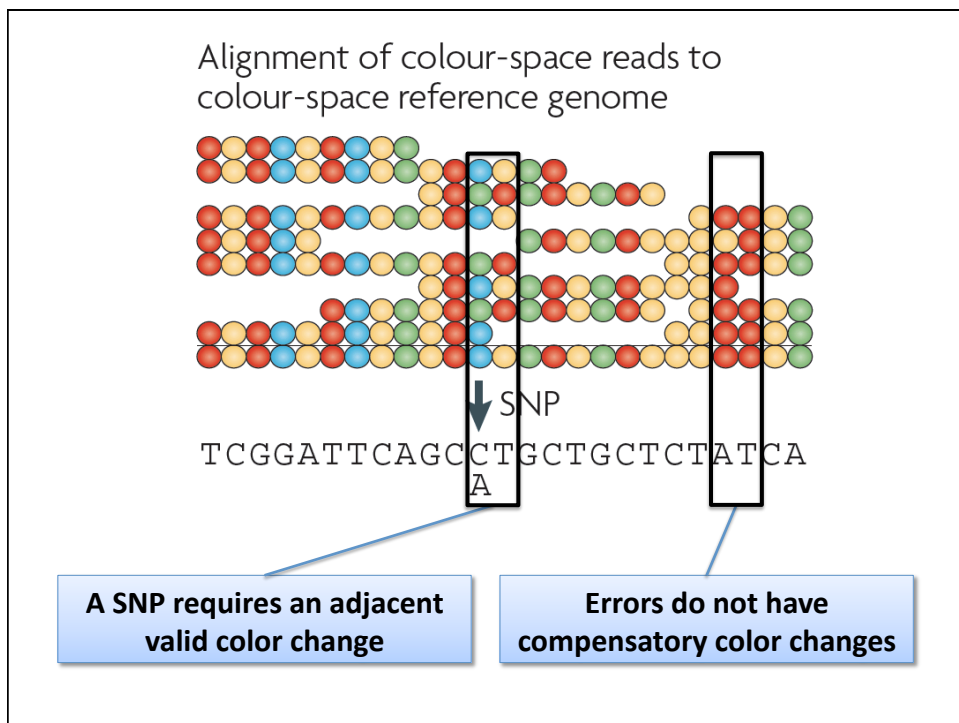
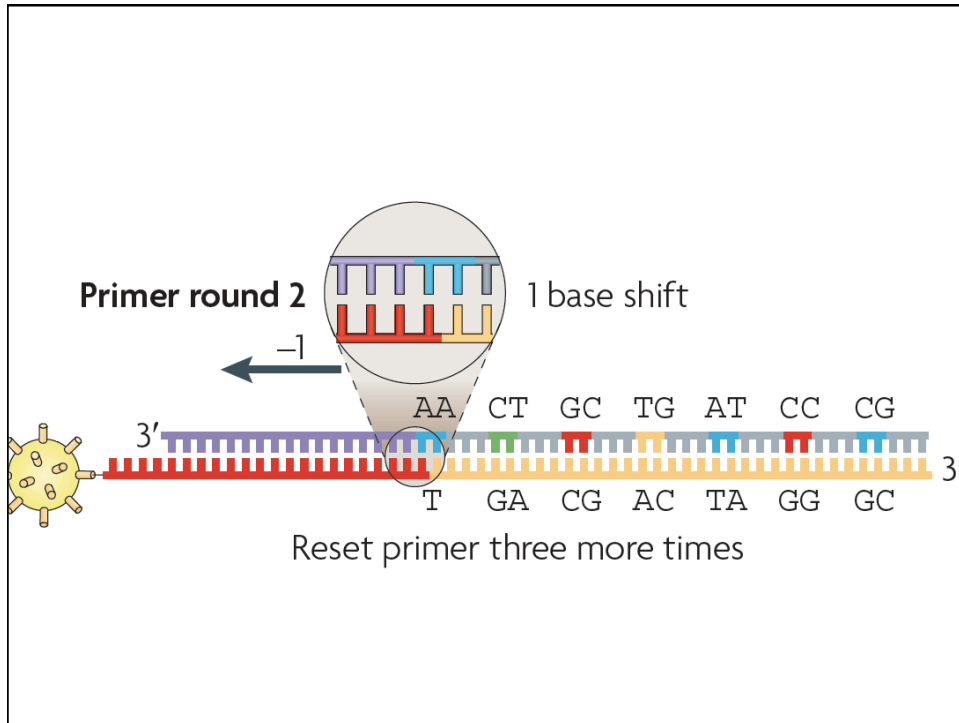


Two-base encoding: each target nucleotide is interrogated twice



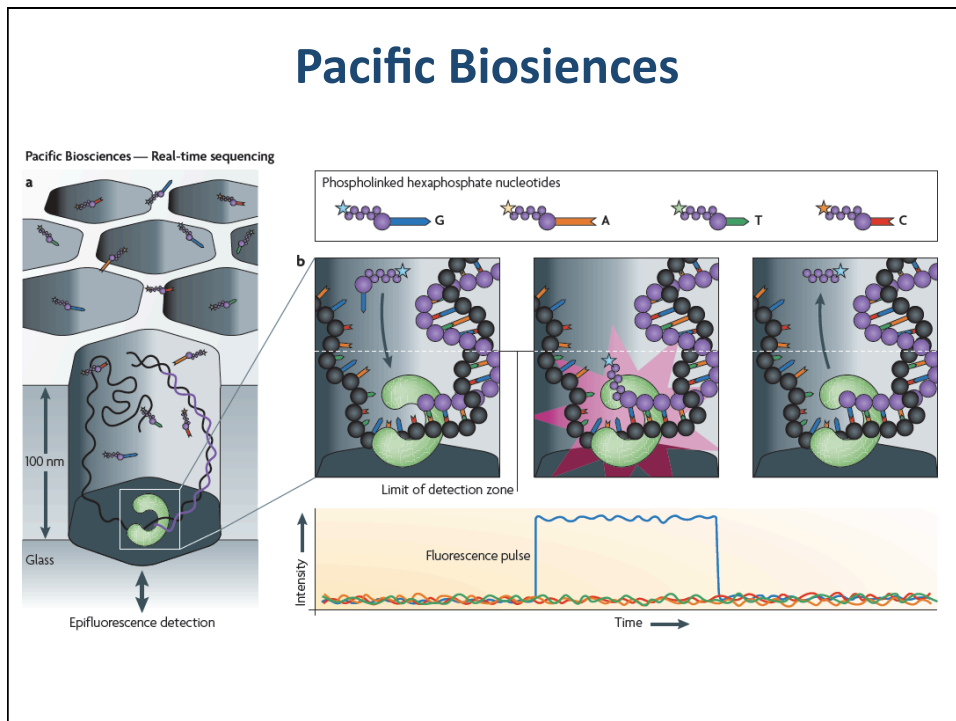
Must know identity of first base to decode color space





Summary of Three Platforms

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US\$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo- polymer repeats	Bacterial and insect genome <i>de novo</i> assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA _{II}	Frag, MP/ solid-phase	RTs	75 or 100	4 [†] , 9 [‡]	18 [†] , 35 [‡]	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7 [†] , 14 [‡]	30 [†] , 50 [‡]	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.

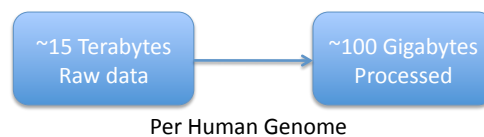




Systems engineering challenges

- **We are still learning what the important bits are that need to be stored.**

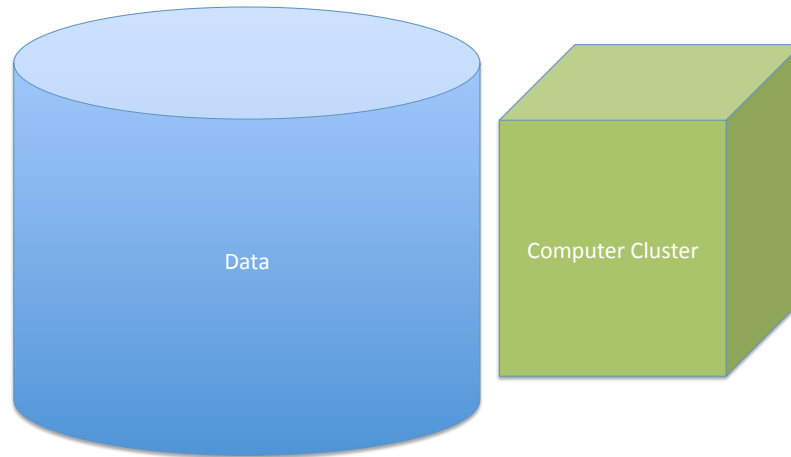
Data volumes are MASSIVE, especially short-term



- **Challenges with cloud computing**

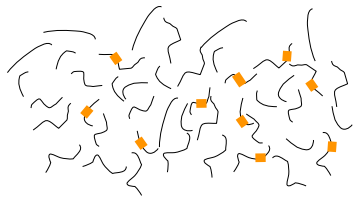
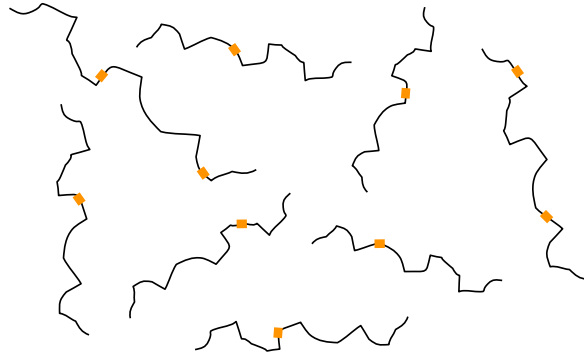


Data and Compute – a shift in complexity

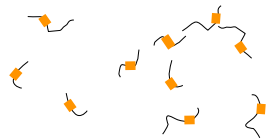


Applications

**Isolate parts of the genome that are
“interesting”**

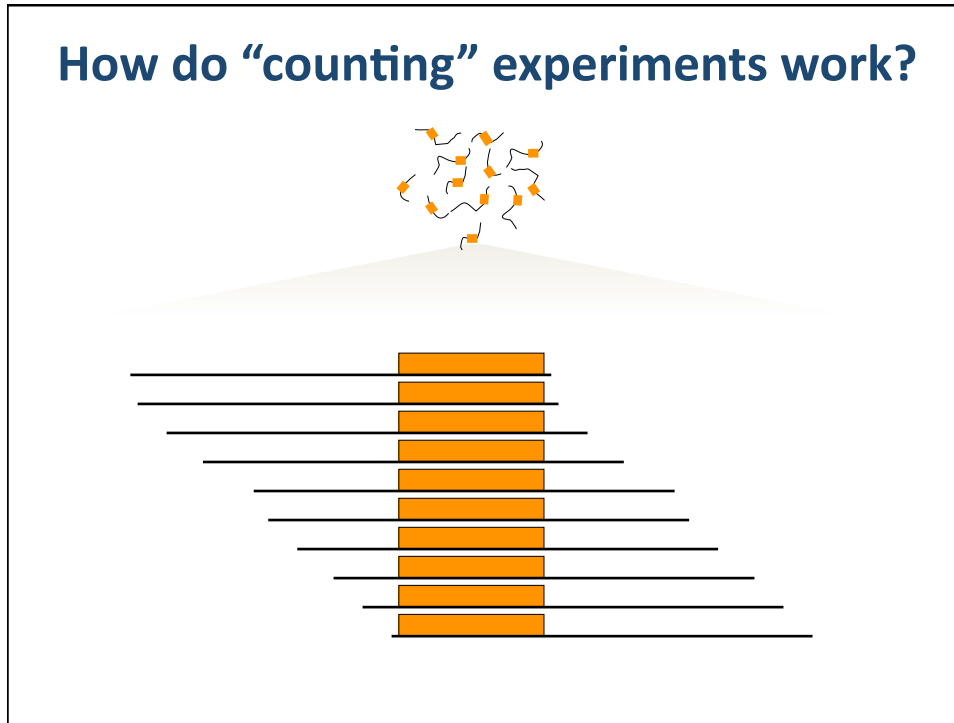


Shear DNA into “small” fragments



Purify DNA fragments of interest

SEQUENCE



Cell (2007) May 18;129(4):823-37.

High-Resolution Profiling of Histone Methylation in the Human Genome

Artem Barski,^{1,2} Suresh Cuddapah,^{1,2} Kairong Cui,^{1,2} Tae-Young Roh,^{1,2} Dustin E. Schones,^{1,2} Zhibin Wang,^{1,2} Gang Wei,^{1,2} Iouri Chepelev,² and Keji Zhao^{1*}

¹Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA
²Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA

*Correspondence: zhaok@nlb.nih.gov
 DOI 10.1016/j.cell.2007.05.009

- One of the first publications using Solexa data
- Reproducible data production
- Correlates with other sequence-based counting experiments
- Identify biologically-relevant patterns of histone methylation
 - Transcription
 - Enhancers
 - Insulators
- Stay tuned for Laura Elnitski’s lecture!

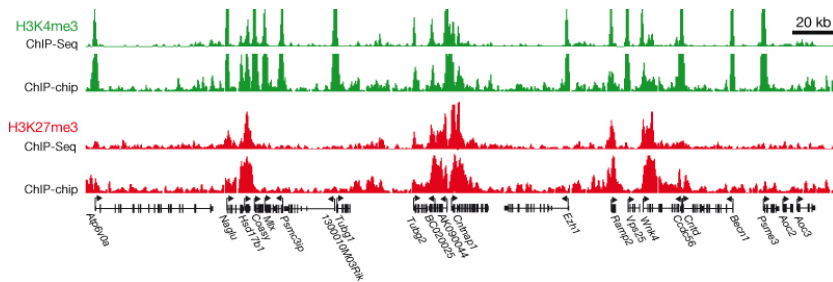
Sequencing-based methods equivalent to Microarray-based methods

ARTICLES

Nature. 2007 Aug 2;448(7153):553-60

Genome-wide maps of chromatin state in pluripotent and lineage-committed cells

Tarjei S. Mikkelsen^{1,2}, Manching Ku^{1,4}, David B. Jaffe¹, Biju Issac^{1,4}, Erez Lieberman^{1,2}, Georgia Giannoukos¹, Pablo Alvarez², William Brockman², Tae-Kyung Kim², Richard P. Koche^{1,2,4}, William Lee¹, Eric Mendenhall^{1,4}, Aisling O'Donovan⁴, Aviva Presser¹, Carsten Russ¹, Xiaohui Xie¹, Alexander Meissner², Marius Wernig¹, Rudolf Jaenisch³, Chad Nusbaum¹, Eric S. Lander^{1,2*} & Bradley E. Bernstein^{1,4*}



Whole Genome Sequencing



First two “personal” genomes sequenced



Watson



Venter

Table 2 | Sequencing statistics on personal genome projects

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

What does “whole-genome” mean?

Vol 456 | 6 November 2008 | doi:10.1038/nature07517

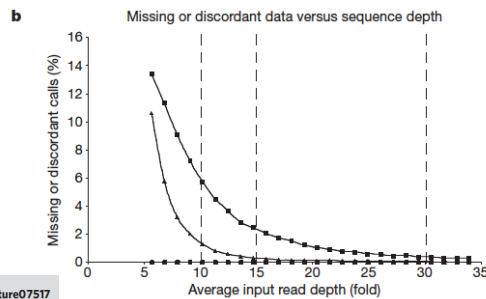
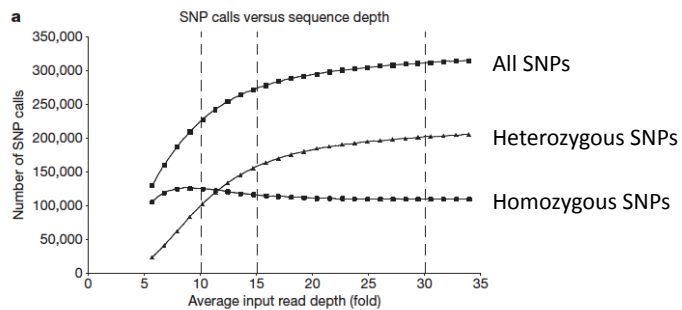
nature

ARTICLES

Accurate whole human genome sequencing using reversible terminator chemistry

- Average 30x base-wise alignment depth of coverage
- 90 Gigabases of aligned sequence
- 120 Gigabases of purity filtered data
- 600 Million paired-end 100bp reads
- Realignment back to reference sequence.

Why 30X?



Vol 456 | 6 November 2008 | doi:10.1038/nature07517

Tumor/Normal Whole-Genome Comparison

nature

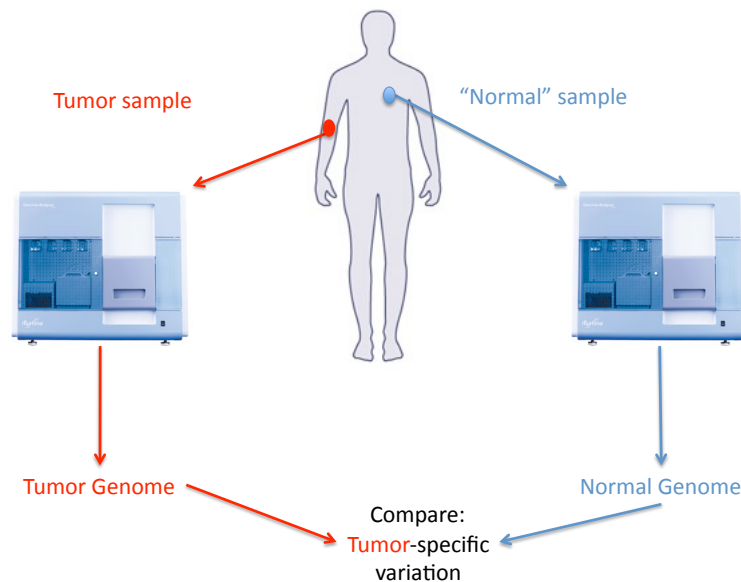
Vol 456 | 6 November 2008 | doi:10.1038/nature07485

ARTICLES

DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome

Timothy J. Ley^{1,2,3,4*}, Elaine R. Mardis^{2,3*}, Li Ding^{2,3}, Bob Fulton³, Michael D. McLellan³, Ken Chen³, David Dooling³, Brian H. Dunford-Shore³, Sean McGrath³, Matthew Hickenbotham³, Lisa Cook³, Rachel Abbott³, David E. Larson³, Dan C. Koboldt³, Craig Pohl³, Scott Smith³, Amy Hawkins³, Scott Abbott³, Devin Locke³, LaDeana W. Hillier^{3,5}, Tracie Miner³, Lucinda Fulton³, Vincent Magrini^{2,3}, Todd Wylie³, Jarret Glasscock³, Joshua Conyers³, Nathan Sander³, Xiaoqi Shi³, John R. Osborne³, Patrick Minx³, David Gordon⁸, Asif Chinwalla³, Yu Zhao¹, Rhonda E. Ries¹, Jacqueline E. Payton⁵, Peter Westervelt^{1,4}, Michael H. Tomasson^{1,4}, Mark Watson^{3,4,5}, Jack Baty⁶, Jennifer Ivanovich^{4,7}, Sharon Heath^{1,4}, William D. Shannon^{1,4}, Rakesh Nagarajan^{4,5}, Matthew J. Walter^{1,4}, Daniel C. Link^{1,4}, Timothy A. Graubert^{1,4}, John F. DiPersio^{1,4} & Richard K. Wilson^{2,3,4}

Experimental Design



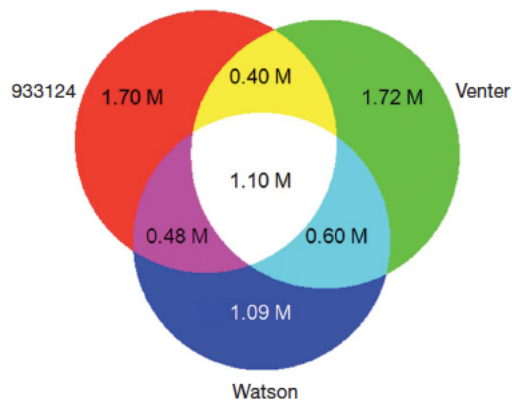
Summary of Data Generated

Table 1 | Tumour and skin genome coverage from patient 933124

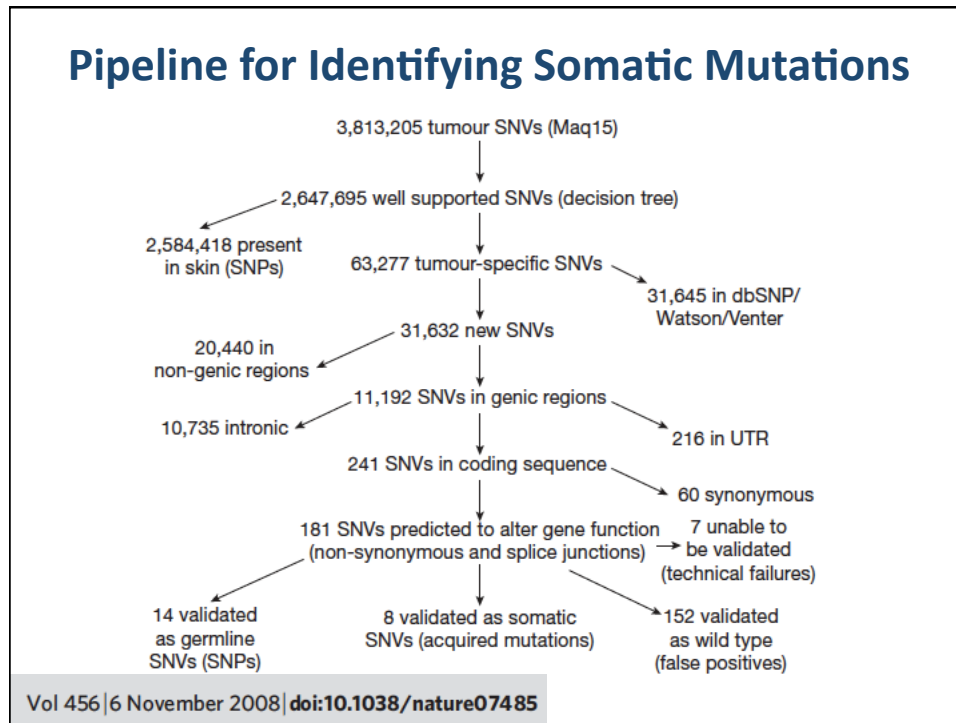
	Tumour	Skin
Libraries	4	3
Runs	98	34
Reads obtained	5,858,992,064	2,122,836,148
Reads passing quality filter	3,025,923,365	1,228,177,690
Bases passing quality filter	98,184,511,523	41,783,794,834
Reads aligned by Maq	2,729,957,053	1,080,576,680
Reads unaligned by Maq	295,966,312	138,276,594
SNVs detected with respect to hg18 (no Y)	3,811,115	2,918,446

Vol 456 | 6 November 2008 | doi:10.1038/nature07485

Comparison to Other "Personal" Genomes



Vol 456 | 6 November 2008 | doi:10.1038/nature07485



Melanoma Tumor Cell Line

doi:10.1038/nature08658 nature

ARTICLES

A comprehensive catalogue of somatic mutations from a human cancer genome

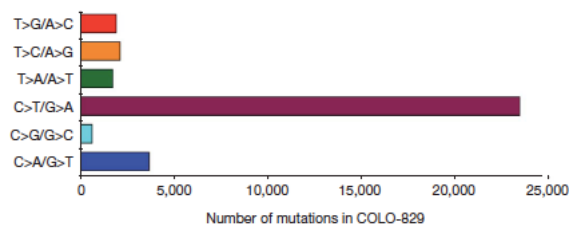
Erin D. Pleasance^{1*}, R. Keira Cheetham^{2*}, Philip J. Stephens¹, David J. McBride¹, Sean J. Humphray², Chris D. Greenman¹, Ignacio Varela¹, Meng-Lay Lin¹, Gonzalo R. Ordóñez¹, Graham R. Bignell¹, Kai Ye³, Julie Alipaz⁴, Markus J. Bauer², David Beare¹, Adam Butler¹, Richard J. Carter², Lina Chen¹, Anthony J. Cox², Sarah Edkins¹, Paula I. Kokko-Gonzales², Niall A. Gormley², Russell J. Grocock², Christian D. Haudenschild⁵, Matthew M. Hims², Terena James², Mingming Jia¹, Zoya Kingsbury², Catherine Leroy¹, John Marshall¹, Andrew Menzies¹, Laura J. Mudie¹, Zemin Ning¹, Tom Royce⁴, Ole B. Schulz-Trieglaff², Anastassia Spiridou², Lucy A. Stebbings¹, Lukasz Szajkowski², Jon Teague¹, David Williamson⁵, Lynda Chin⁶, Mark T. Ross², Peter J. Campbell¹, David R. Bentley², P. Andrew Futreal¹ & Michael R. Stratton^{1,7}

Defining a Somatic Variant

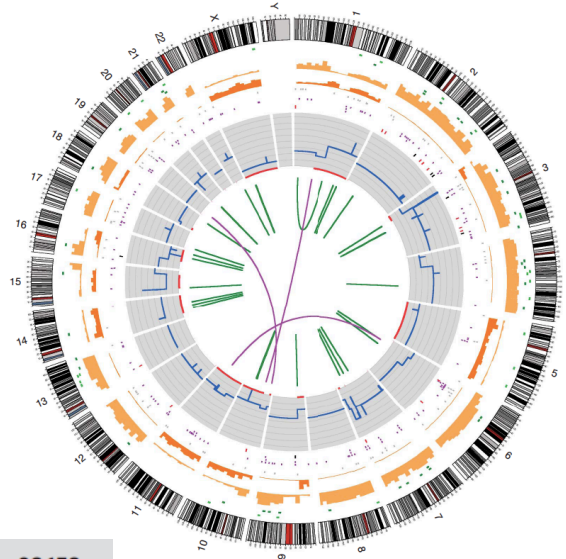
- **Minimum of 3 high-quality reads in the tumor with a variant**
- **Minimum of 10X coverage in the normal and no evidence of the variant**
- **Systematic biases/errors are eliminated**
 - Library preparation
 - Sequencing chemistry
 - Read alignment

Validation Efforts

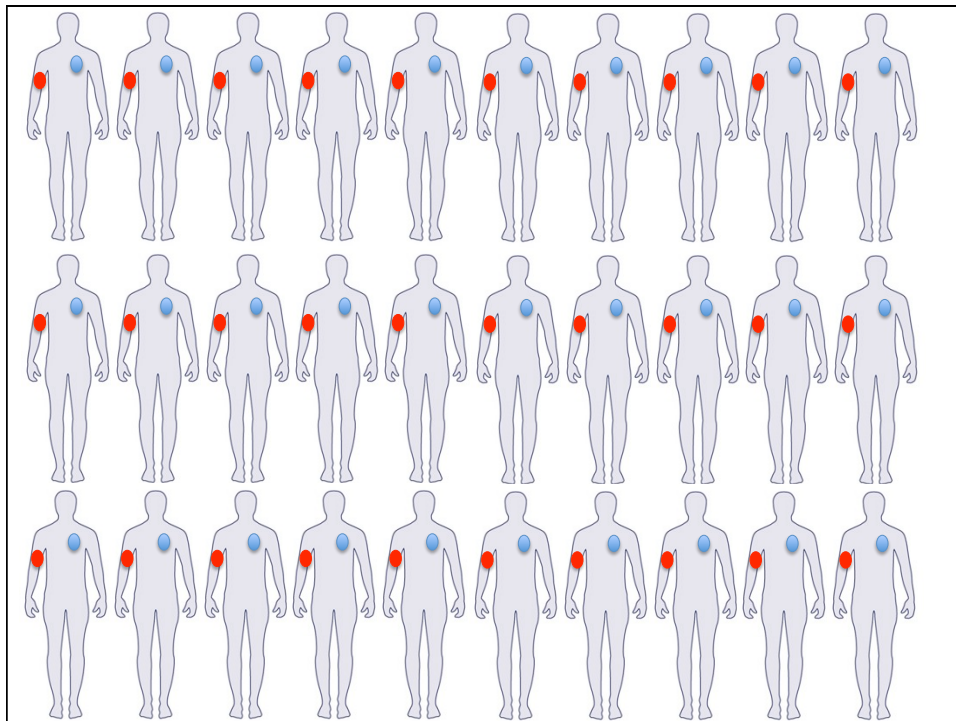
- **32,325 somatic variants detected**
- **Validation against Sanger sequencing data:**
 - 42 of 48 previously known somatic variants detected
88% sensitivity
 - 452 of 470 newly detected somatic variants confirmed
3% false positive rate
- **Mutational profile reflective of UV DNA damage:**

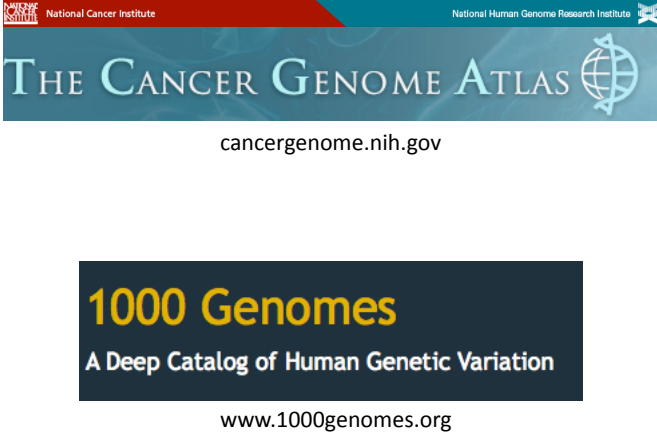


Translocations and Copy Number Variations



doi:10.1038/nature08658





The Cancer Genome Atlas logo features the text "THE CANCER GENOME ATLAS" in white serif font on a blue background, with a globe icon to the right. Above it are logos for the National Cancer Institute and the National Human Genome Research Institute.

1000 Genomes logo features the text "1000 Genomes" in yellow and "A Deep Catalog of Human Genetic Variation" in white on a dark blue background.

cancergenome.nih.gov

www.1000genomes.org

