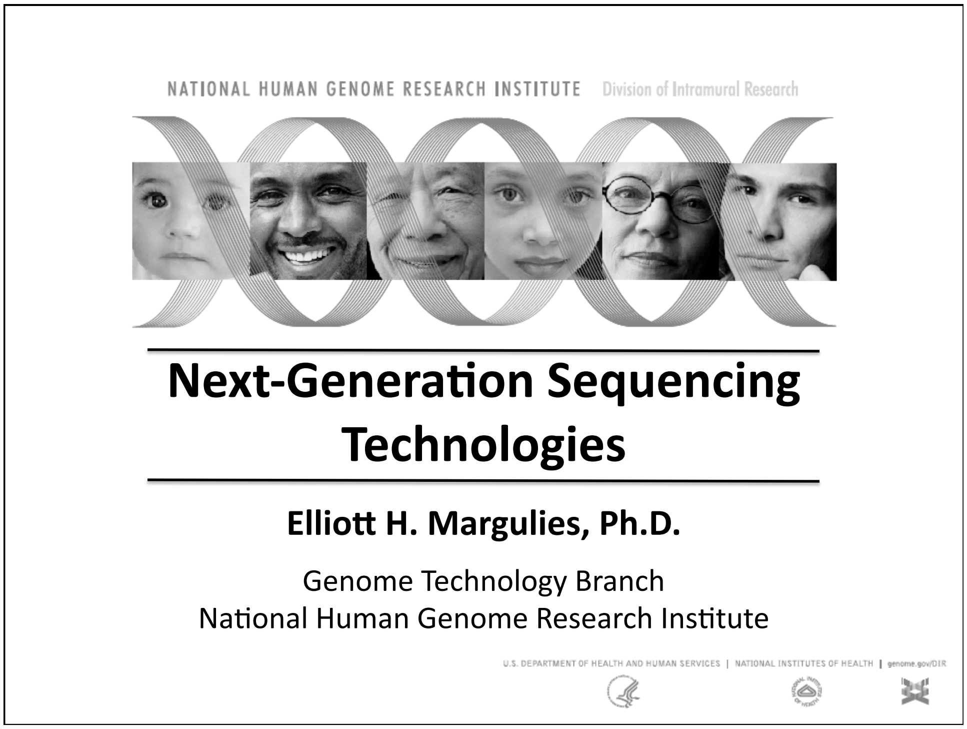


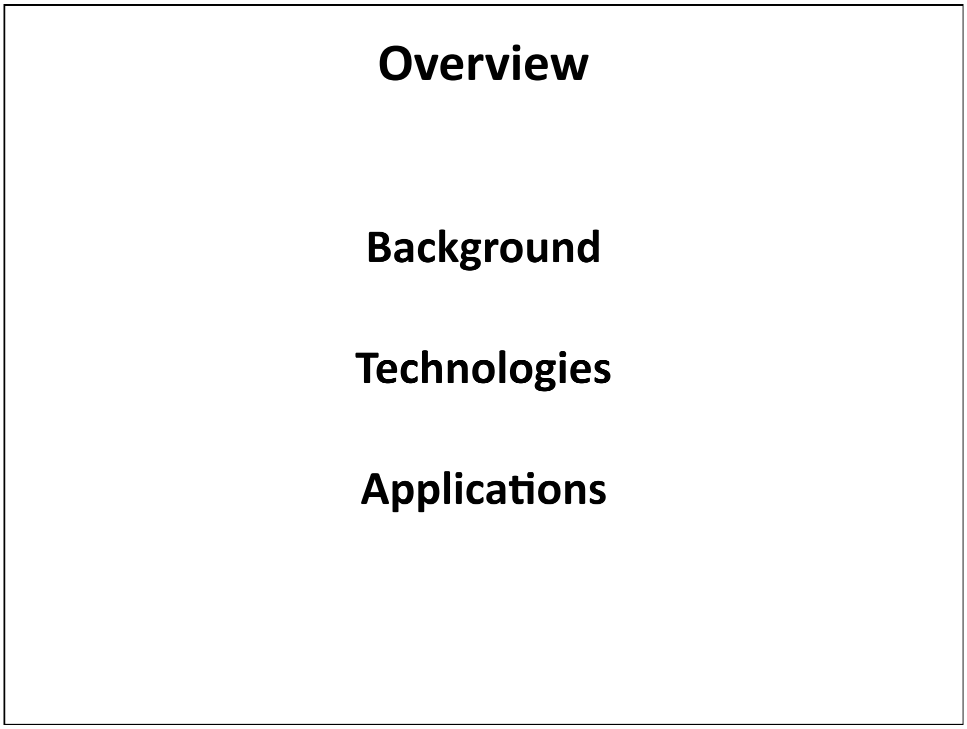
NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research

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

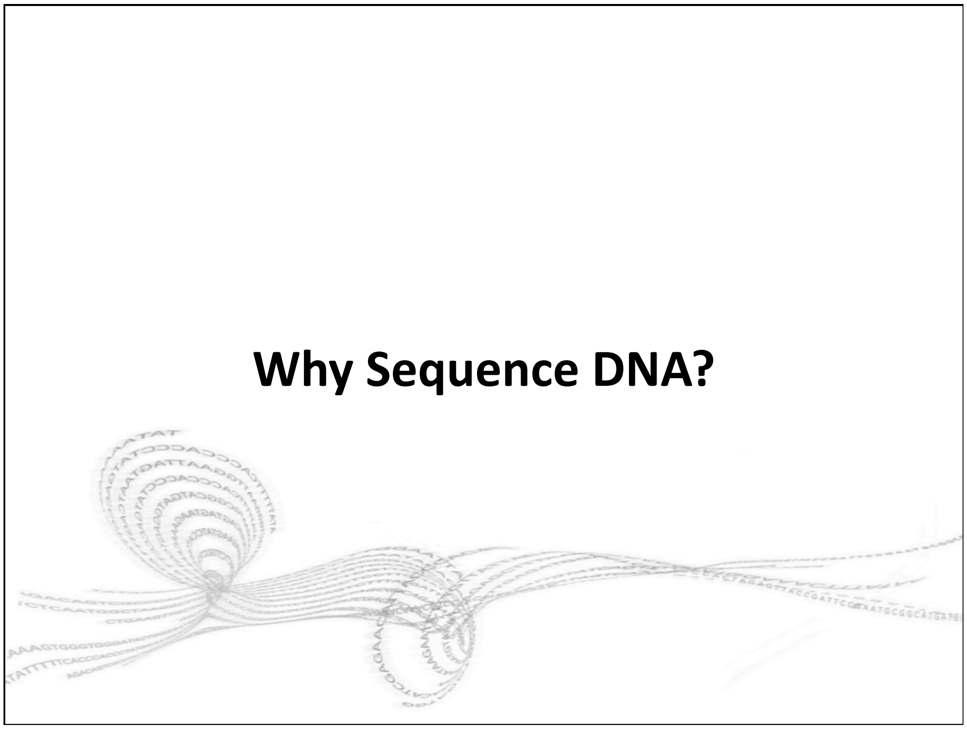


Overview

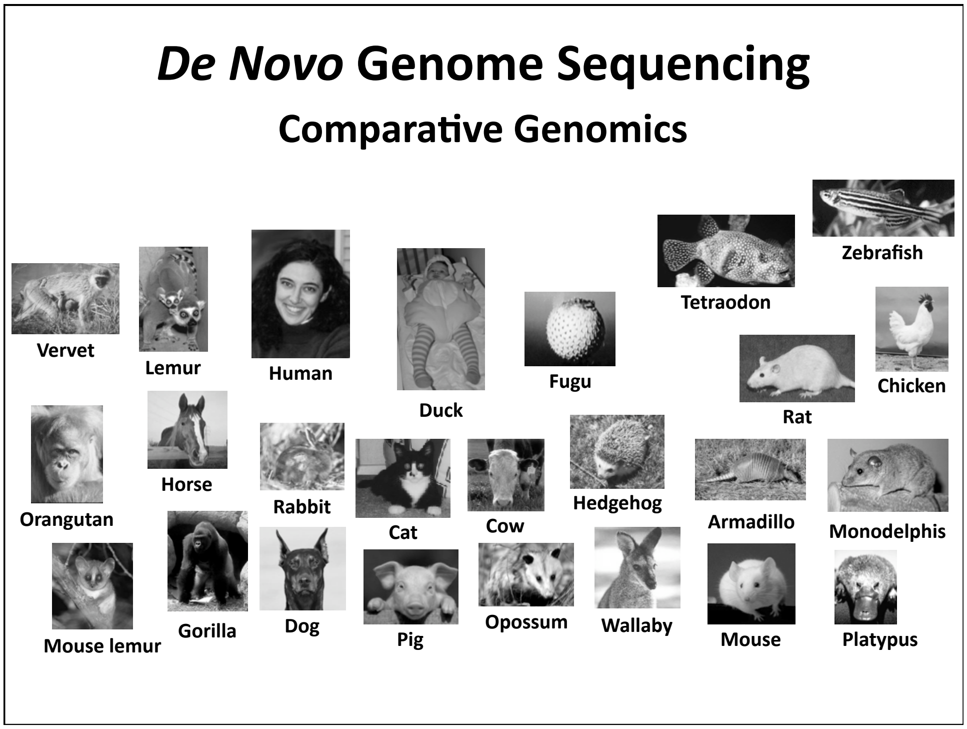
Background

Technologies

Applications



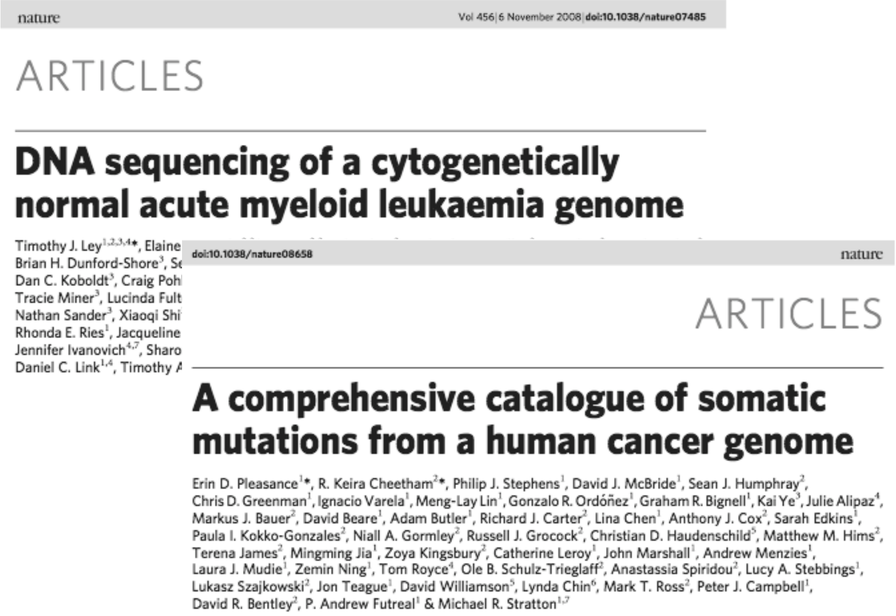
Why Sequence DNA?



De Novo Genome Sequencing

Comparative Genomics

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-Shore[3], Se
Dan C. Koboldt[3], Craig Pohl
Tracie Miner[3], Lucinda Fult
Nathan Sander[3], Xiaoqi Shi
Rhonda E. Ries[1], Jacqueline
Jennifer Ivanovich[4,7], 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

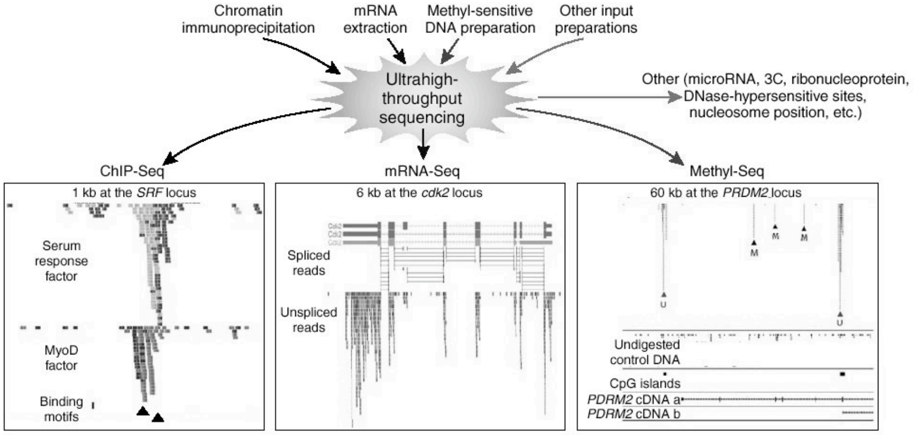
Erin D. Pleasance[1]*, R. Keira Cheetham[2]*, Philip J. Stephens[1], David J. McBride[1], Sean J. Humphray[2], Chris D. Greenman[1], Ignacio Varela[1], Meng-Lay Lin[1], Gonzalo R. Ordóñez[1], Graham R. Bignell[1], Kai Ye[3], Julie Alipaz[4], Markus J. Bauer[2], David Beare[1], Adam Butler[1], Richard J. Carter[2], Lina Chen[1], Anthony J. Cox[2], Sarah Edkins[1], Paula I. Kokko-Gonzales[2], Niall A. Gormley[2], Russell J. Grocock[2], Christian D. Haudenschild[5], Matthew M. Hims[2], Terena James[2], Mingming Jia[1], Zoya Kingsbury[2], Catherine Leroy[1], John Marshall[1], Andrew Menzies[1], Laura J. Mudie[1], Zemin Ning[1], Tom Royce[4], Ole B. Schulz-Trieglaff[2], Anastassia Spiridou[2], Lucy A. Stebbings[1], Lukasz Szajkowski[2], Jon Teague[1], David Williamson[5], Lynda Chin[6], Mark T. Ross[2], Peter J. Campbell[1], David R. Bentley[2], P. Andrew Futreal[1] & Michael R. Stratton[1,7]

"Counting Experiments"

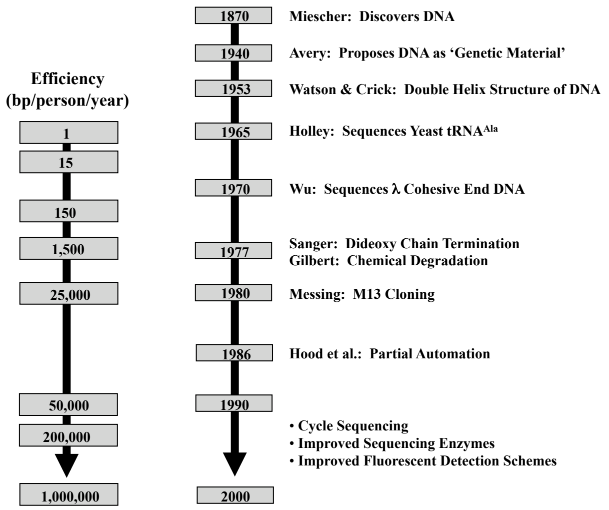
NATURE METHODS | VOL.5 NO.1 | JANUARY 2008 | **19**

Sequence census methods for functional genomics

Barbara Wold & Richard M Myers



History of Nucleic Acid Sequencing

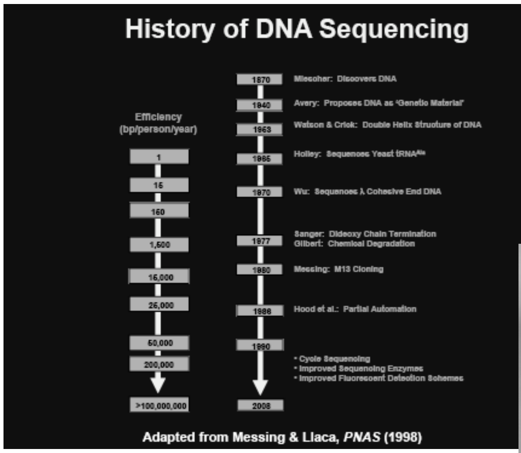

Efficiency (bp/person/year)

- 1
- 15
- 150
- 1,500
- 25,000
- 50,000
- 200,000
- 1,000,000

- 1870 Miescher: Discovers DNA
- 1940 Avery: Proposes DNA as 'Genetic Material'
- 1953 Watson & Crick: Double Helix Structure of DNA
- 1965 Holley: Sequences Yeast tRNA^{Ala}
- 1970 Wu: Sequences λ Cohesive End DNA
- 1977 Sanger: Dideoxy Chain Termination
Gilbert: Chemical Degradation
- 1980 Messing: M13 Cloning
- 1986 Hood et al.: Partial Automation
- 1990
  - Cycle Sequencing
  - Improved Sequencing Enzymes
  - Improved Fluorescent Detection Schemes
- 2000

Slide kindly provided by Eric Green

Plateau in Sequencing Technology


History of DNA Sequencing

Efficiency (bp/person/year)

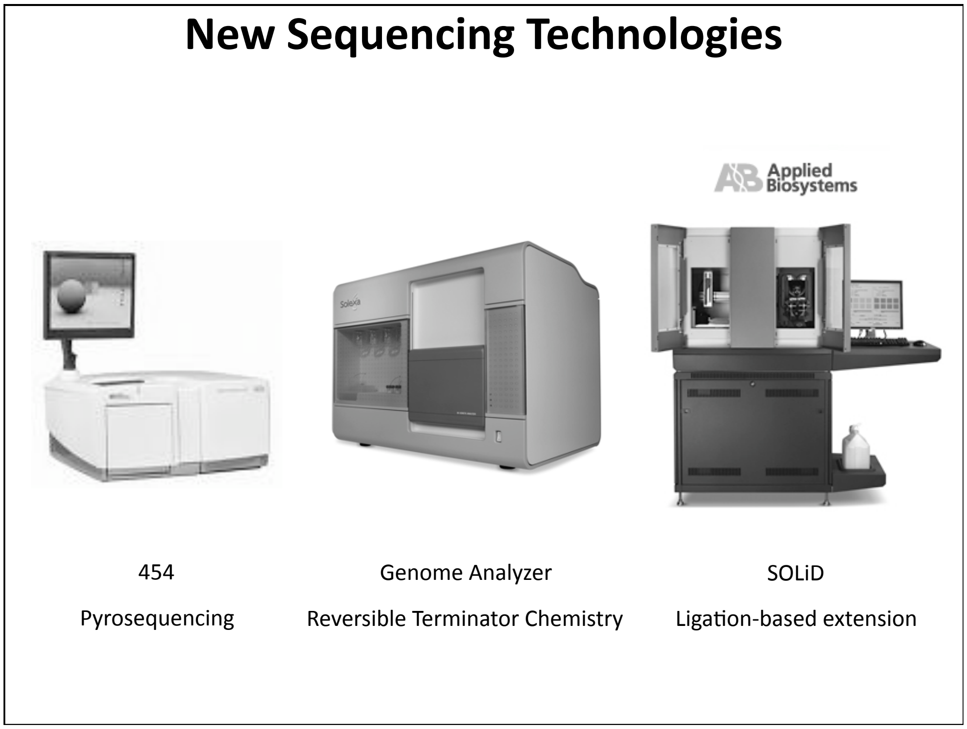
- 1
- 15
- 150
- 1,500
- 16,000
- 25,000
- 50,000
- 200,000
- >100,000,000

- 1870 Miescher: Discovers DNA
- 1940 Avery: Proposes DNA as 'Genetic Material'
- 1953 Watson & Crick: Double Helix Structure of DNA
- 1965 Holley: Sequences Yeast tRNA^{Ala}
- 1970 Wu: Sequences λ Cohesive End DNA
- 1977 Sanger: Dideoxy Chain Termination
Gilbert: Chemical Degradation
- 1980 Messing: M13 Cloning
- 1986 Hood et al.: Partial Automation
- 1990
  - Cycle Sequencing
  - Improved Sequencing Enzymes
  - Improved Fluorescent Detection Schemes
- 2008

Adapted from Messing & Llaca, *PNAS* (1998)

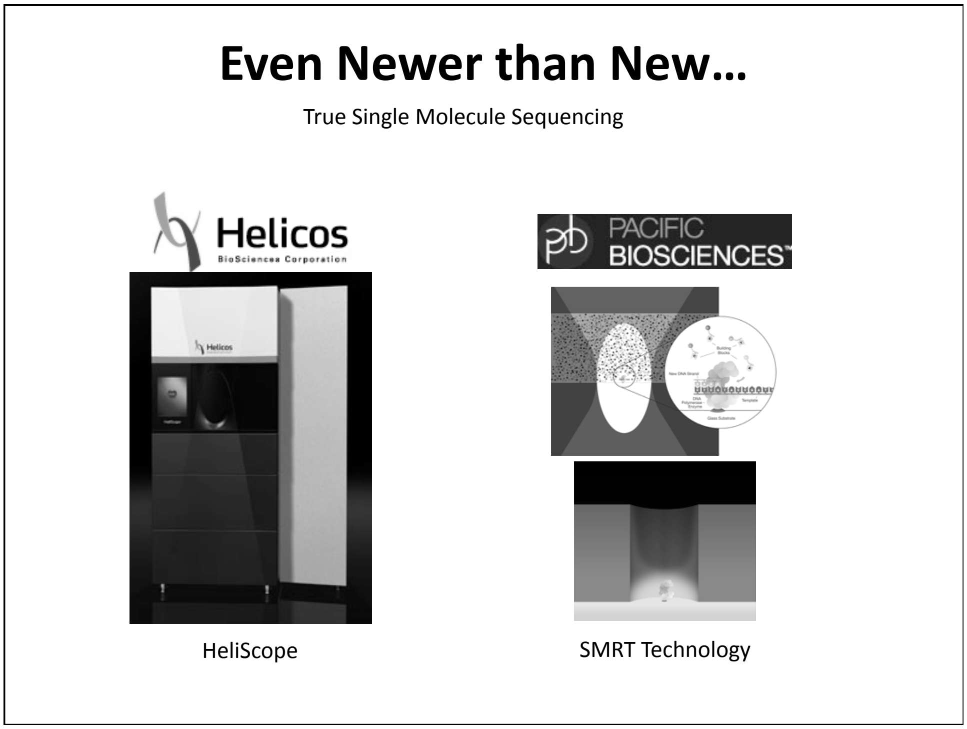
Current Topics in Genome Analysis, E. Green, Lecture 1

AB 3730 xl



New Sequencing Technologies

454	Genome Analyzer	SOLiD
Pyrosequencing	Reversible Terminator Chemistry	Ligation-based extension



Even Newer than New…

True Single Molecule Sequencing

HeliScope

SMRT Technology

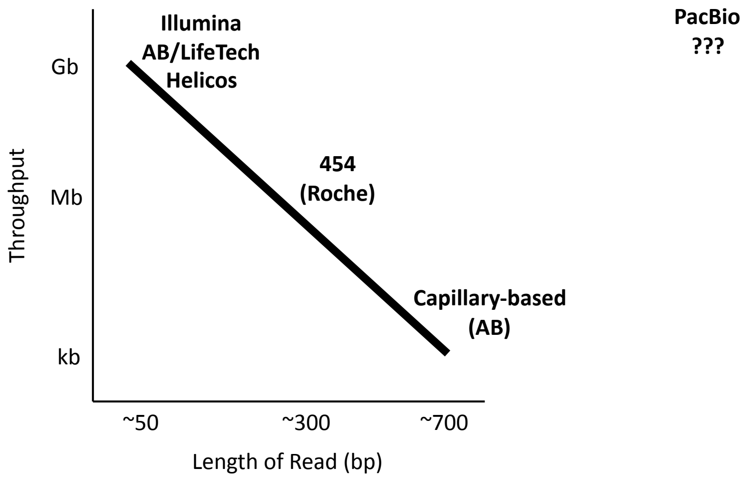
NATURE REVIEWS | GENETICS VOLUME 11 | JANUARY 2010 | 31

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing technologies — the next generation

*Michael L. Metzker**‡

Trade-offs with Newer Sequencing Technologies

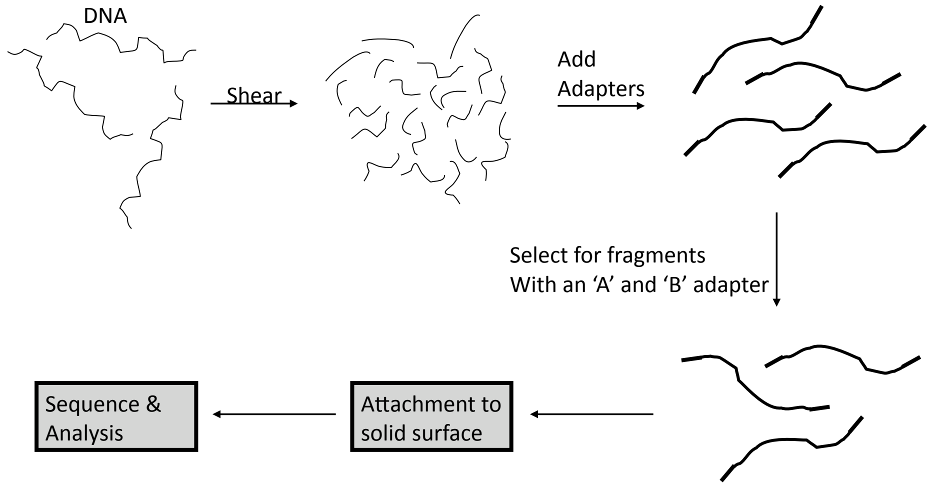


Throughput = $\frac{\text{Amount of Sequence Generated}}{\text{Unit of Time or Cost}}$

Template/Library Preparation Methods

"Single Molecule" Sequencing

Really a clonal amplification of a single DNA molecule



DNA

Shear

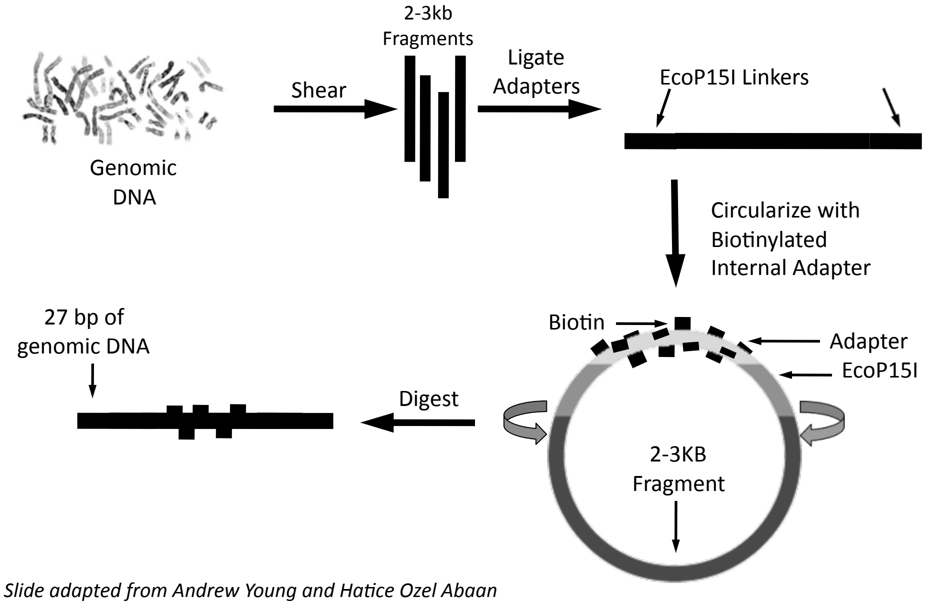
Add
Adapters

Select for fragments
With an 'A' and 'B' adapter

Attachment to
solid surface

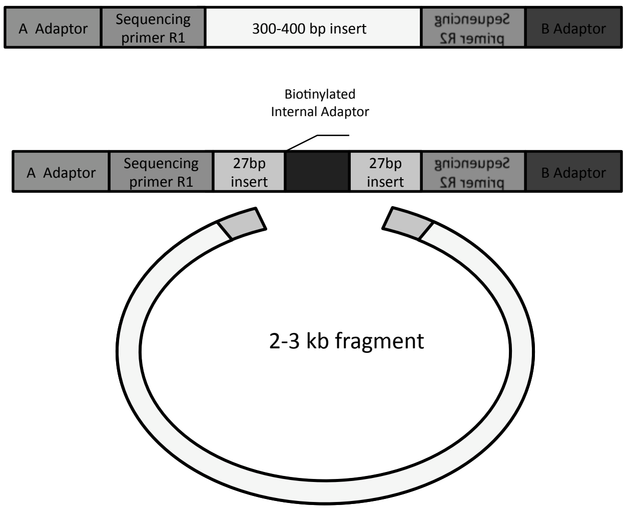
Sequence &
Analysis

Creating Large Insert Paired-End Libraries



Slide adapted from Andrew Young and Hatice Ozel Abaan

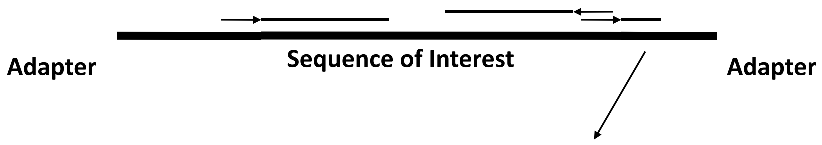
Paired-End Sample Preparation



Slide adapted from Andrew Young and Hatice Ozel Abaan

Multiplexing Samples

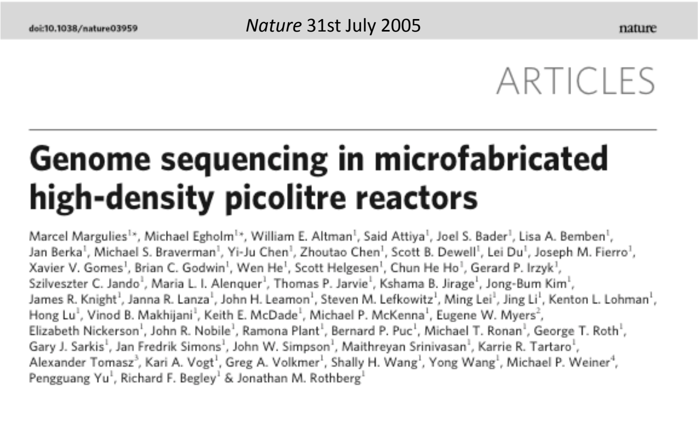
Paired End Indexing Strategy



Index

```
ACAGTG
ACTTGA
ATCACG
CAGATC
CGATGT
GATCAG
GCCAAT
GGCTAC
TAGCTT
TGACCA
TTAGGC
```

454 Sequencing Technology



doi:10.1038/nature03959 *Nature* 31st July 2005 nature

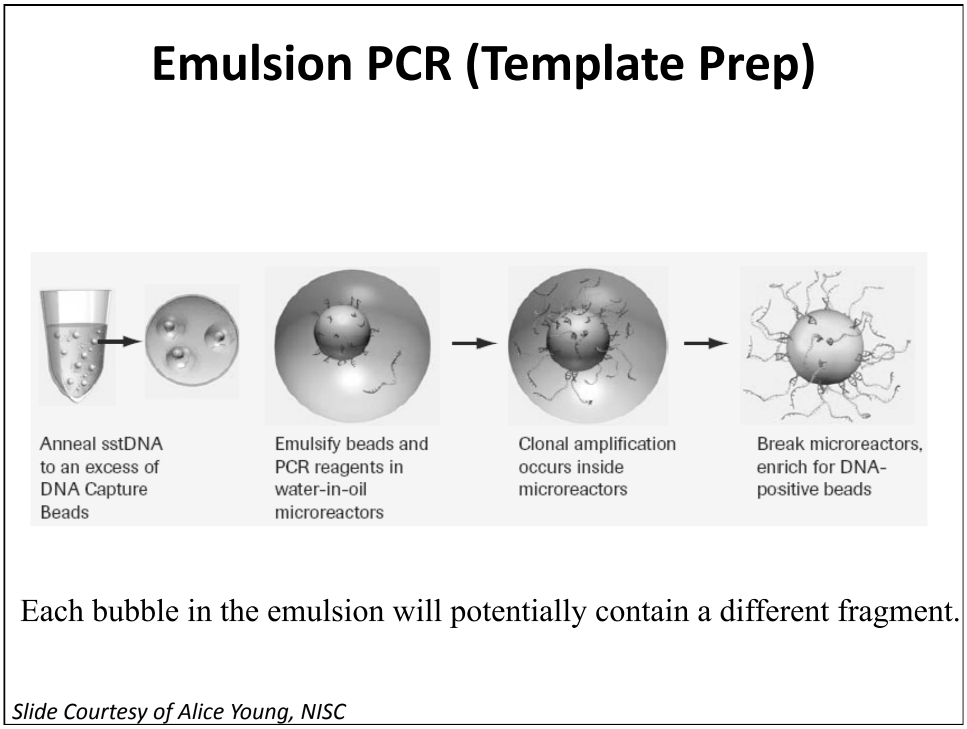
ARTICLES

Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies[1*], Michael Egholm[1*], William E. Altman[1], Said Attiya[1], Joel S. Bader[1], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[1], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gomes[1], Brian C. Godwin[1], Wen He[1], Scott Helgesen[1], Chun He Ho[1], Gerard P. Irzyk[1], Szilveszter C. Jando[1], Maria L. I. Alenquer[1], Thomas P. Jarvie[1], Kshama B. Jirage[1], Jong-Bum Kim[1], James R. Knight[1], Janna R. Lanza[1], John H. Leamon[1], Steven M. Lefkowitz[1], Ming Lei[1], Jing Li[1], Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[3], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

454 LIFE SCIENCES

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



Emulsion PCR (Template Prep)

Anneal sstDNA to an excess of DNA Capture Beads

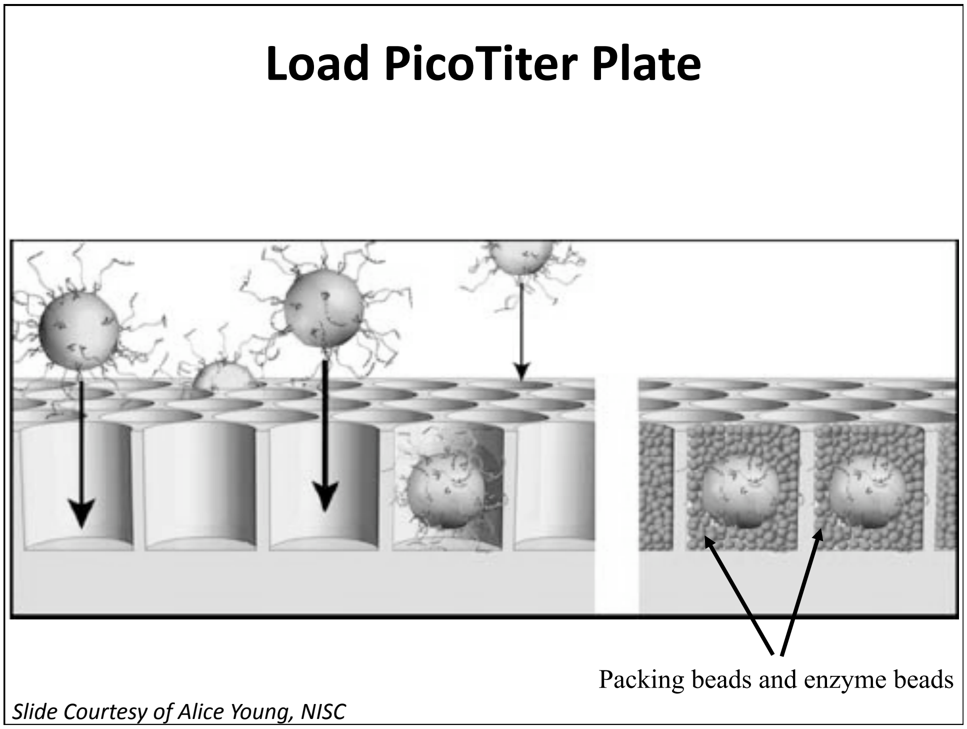
Emulsify beads and PCR reagents in water-in-oil microreactors

Clonal amplification occurs inside microreactors

Break microreactors, enrich for DNA-positive beads

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

Slide Courtesy of Alice Young, NISC

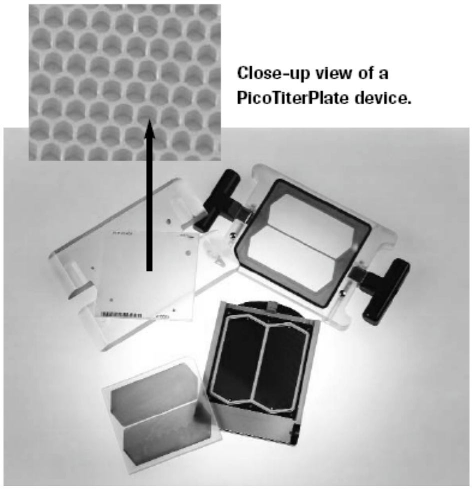


Load PicoTiter Plate

Packing beads and enzyme beads

Slide Courtesy of Alice Young, NISC

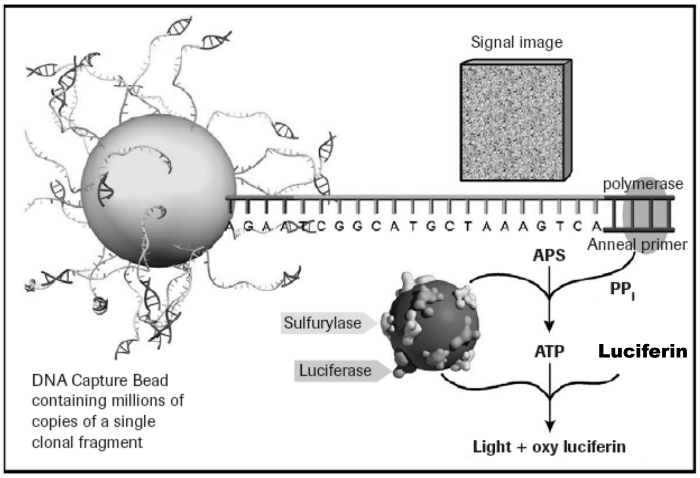
PicoTiter Plate Apparatus



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

Slide Courtesy of Alice Young, NISC

PyroSequencing


Signal image
polymerase
A G A A T C G G C A T G C T A A A G T C A
Anneal primer
APS
PP$_i$
Sulfurylase
Luciferase
ATP
Luciferin
Light + oxy luciferin
DNA Capture Bead containing millions of copies of a single clonal fragment

Slide Courtesy of Alice Young, NISC

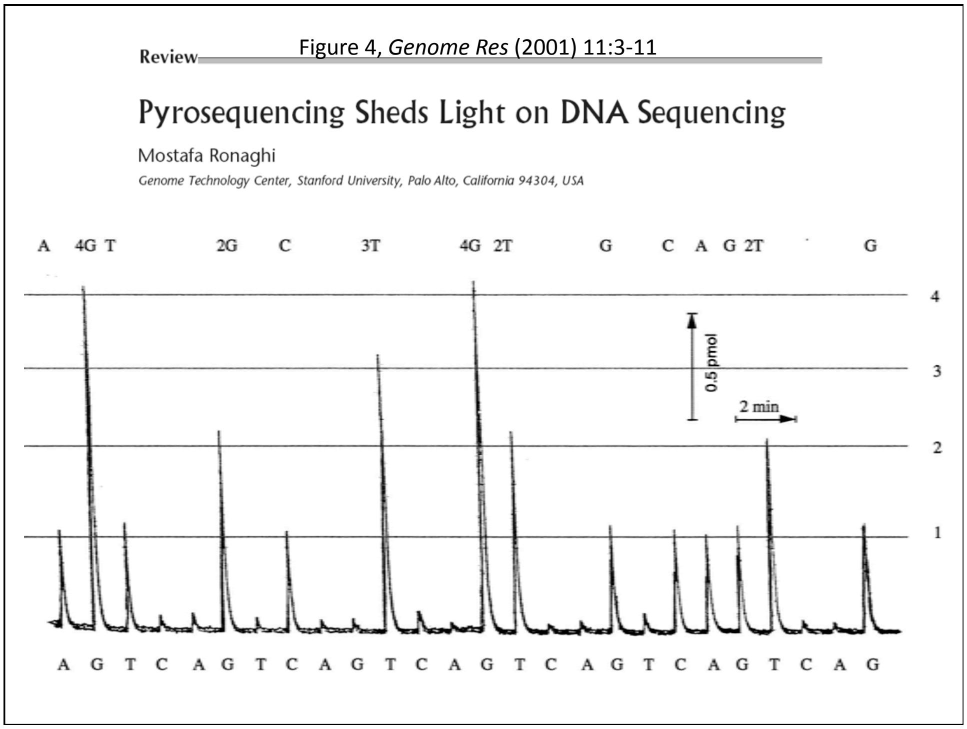


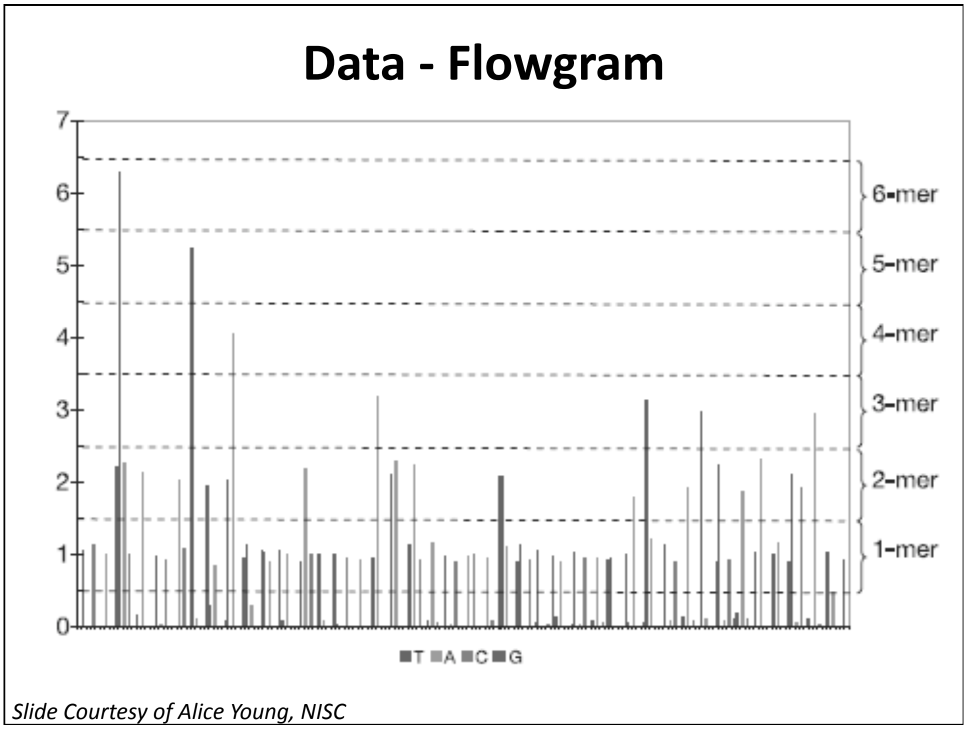
Figure 4, *Genome Res* (2001) 11:3-11

Review

Pyrosequencing Sheds Light on DNA Sequencing

Mostafa Ronaghi

Genome Technology Center, Stanford University, Palo Alto, California 94304, USA



Data - Flowgram

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. Green[1], Johannes Krause[1], Susan E. Ptak[1], Adrian W. Briggs[1], Michael T. Ronan[2], Jan F. Simons[2], Lei Du[2], Michael Egholm[2], Jonathan M. Rothberg[2], Maja Paunovic[3]‡ & Svante Pääbo[1]

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,[4] Joe Alessi,[1] Feng Chen,[1] Darren Platt,[1] Svante Pääbo,[4] Jonathan K. Pritchard,[3] Edward M. Rubin[1,2]*

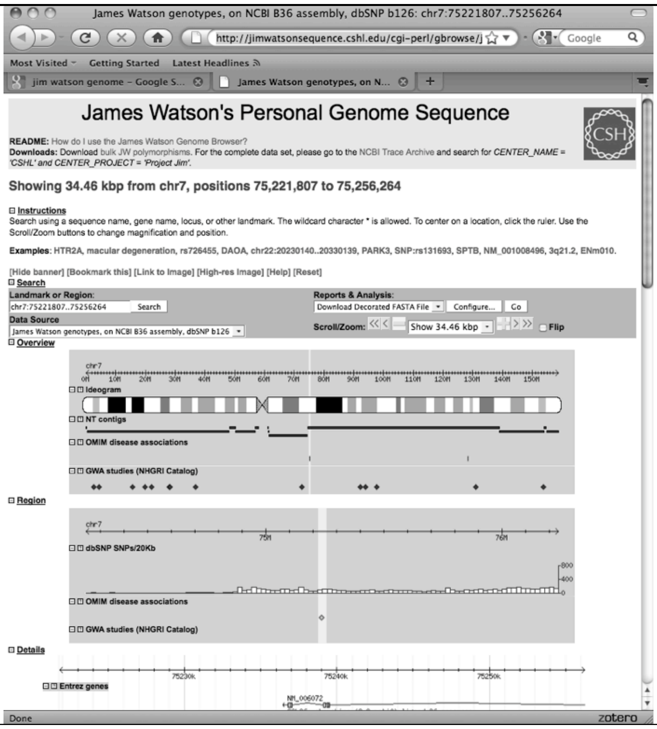
http://popsci.typepad.com/photos/uncategorized/2007/10/25/laluezafox1lr.jpg

nature Vol 452|17 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 Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†



James Watson genotypes, on NCBI B36 assembly, dbSNP b126: chr7:75221807..75256264

http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/j

James Watson's Personal Genome Sequence

README: How do I use the James Watson Genome Browser?
Downloads: Download bulk JW polymorphisms. For the complete data set, please go to the NCBI Trace Archive and search for *CENTER_NAME = 'CSHL' and CENTER_PROJECT = 'Project Jim'.*

Showing 34.46 kbp from chr7, positions 75,221,807 to 75,256,264

⊟ **Instructions**
Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: HTR2A, macular degeneration, rs726455, DAOA, chr22:20230140..20330139, PARK3, SNP:rs131693, SPTB, NM_001008496, 3q21.2, ENm010.

[Hide banner] [Bookmark this] [Link to Image] [High-res Image] [Help] [Reset]

⊟ **Search**

Landmark or Region: chr7:75221807..75256264

Data Source: James Watson genotypes, on NCBI B36 assembly, dbSNP b126

Reports & Analysis: Download Decorated FASTA File

Scroll/Zoom: Show 34.46 kbp

⊟ **Overview**

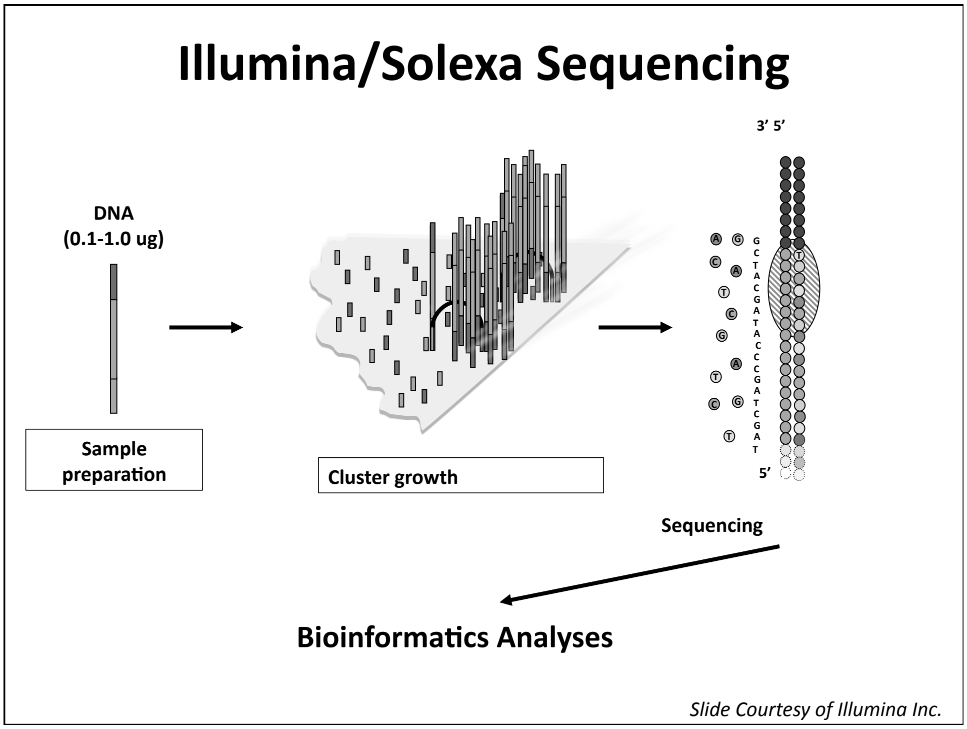
⊟ **Region**

⊟ **Details**

Done

Most Visited
Getting Started
Latest Headlines
jim watson genome – Google S...
James Watson genotypes, on N...
Google
CSH
Search
Configure...
Go
Flip
chr7
0M 10M 20M 30M 40M 50M 60M 70M 80M 90M 100M 110M 120M 130M 140M 150M
Ideogram
NT contigs
OMIM disease associations
GWA studies (NHGRI Catalog)
chr7
75M
76M
dbSNP SNPs/20Kb
800
400
0
OMIM disease associations
GWA studies (NHGRI Catalog)
75230k
75240k
75250k
Entrez genes
NM_006072
zotero

Illumina Genome Analyzer



Illumina/Solexa Sequencing

DNA
(0.1-1.0 ug)

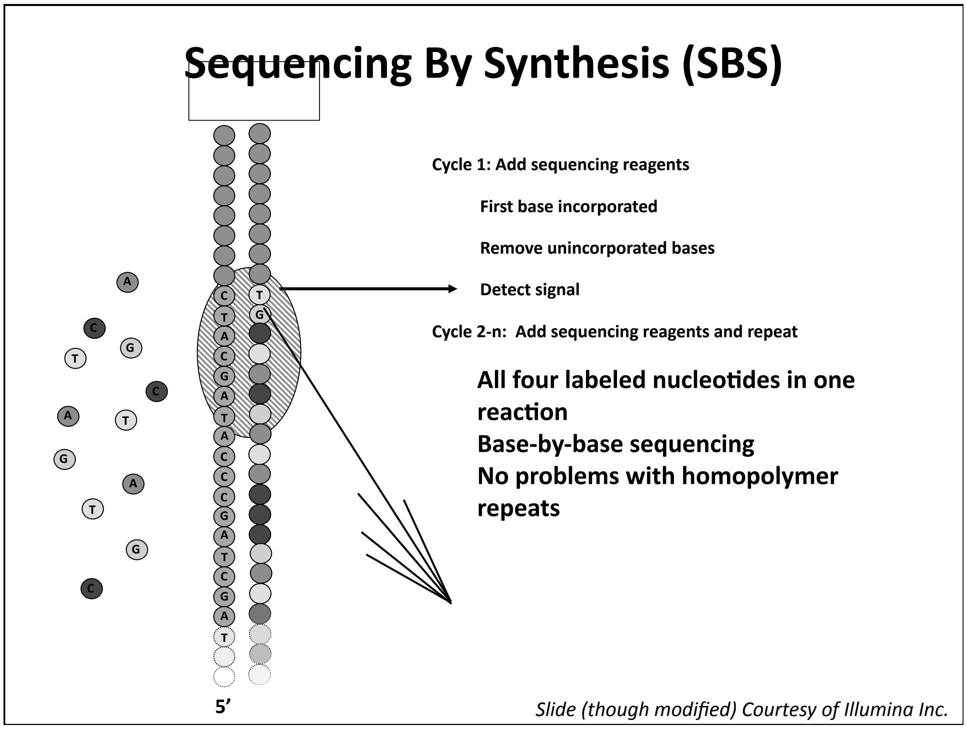
Sample preparation

Cluster growth

Sequencing

Bioinformatics Analyses

Slide Courtesy of Illumina Inc.



Sequencing By Synthesis (SBS)

Cycle 1: Add sequencing reagents

First base incorporated

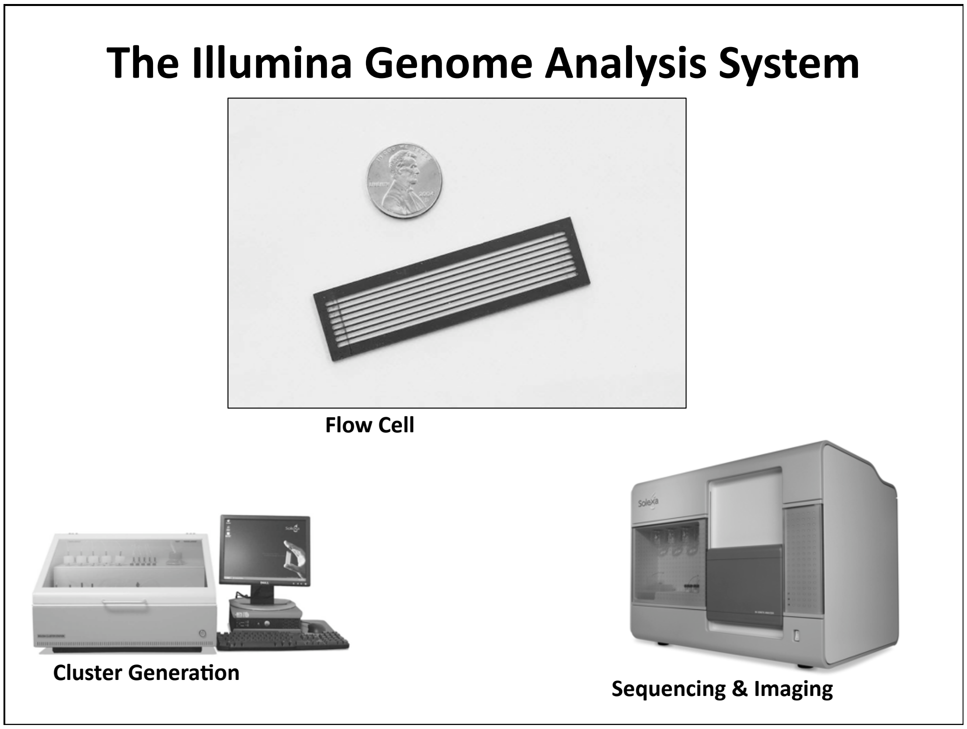
Remove unincorporated bases

Detect signal

Cycle 2-n: Add sequencing reagents and repeat

All four labeled nucleotides in one reaction
Base-by-base sequencing
No problems with homopolymer repeats

Slide (though modified) Courtesy of Illumina Inc.

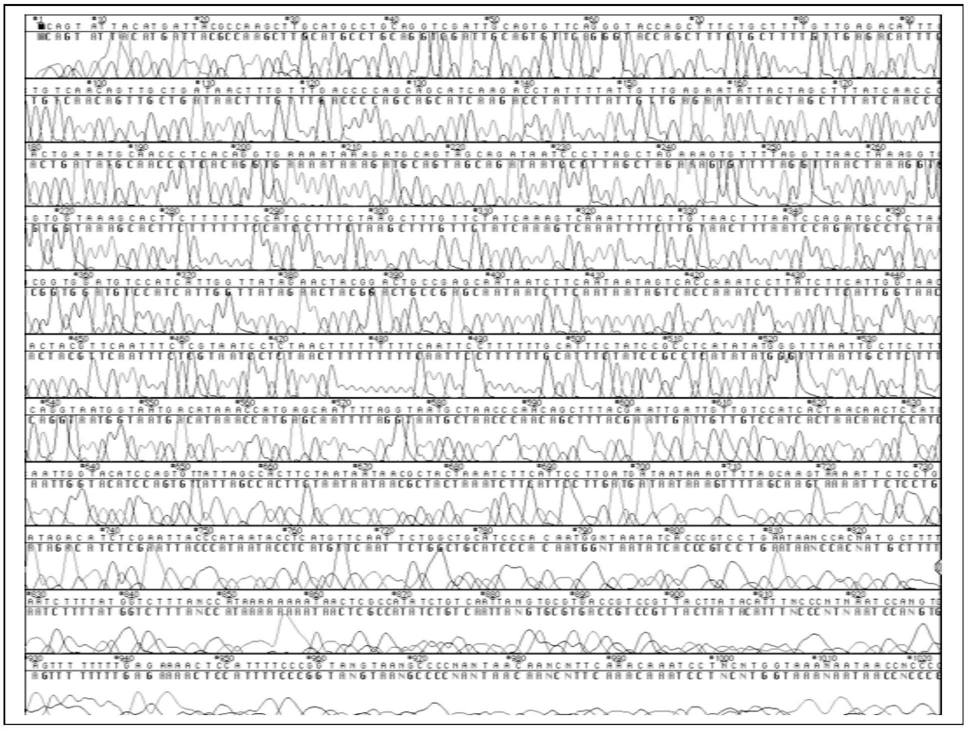


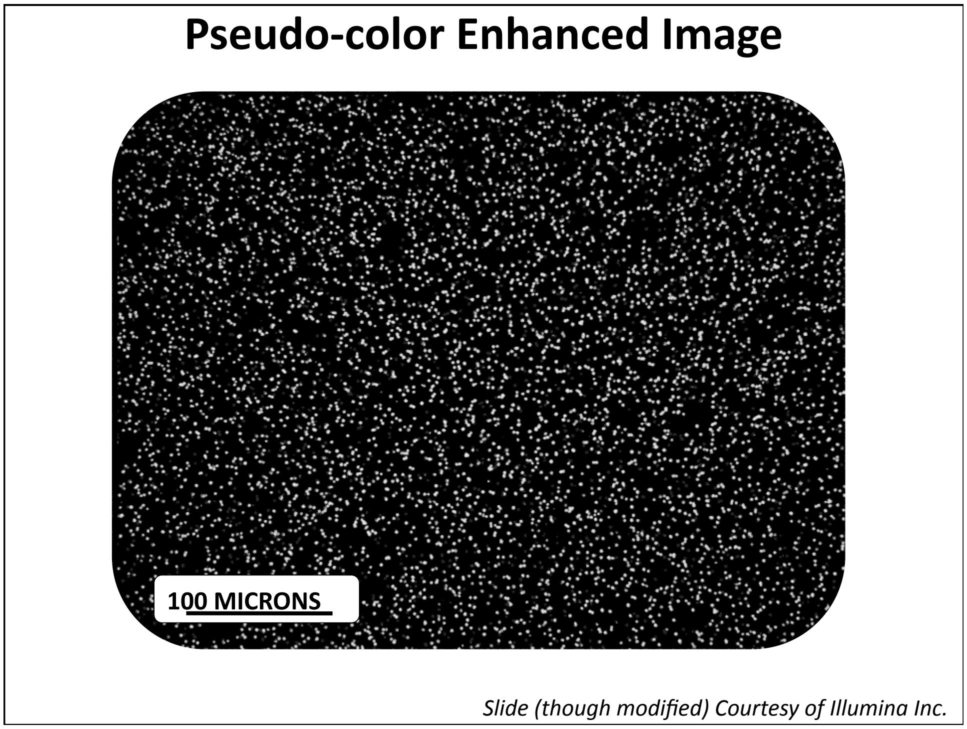
The Illumina Genome Analysis System

Flow Cell

Cluster Generation

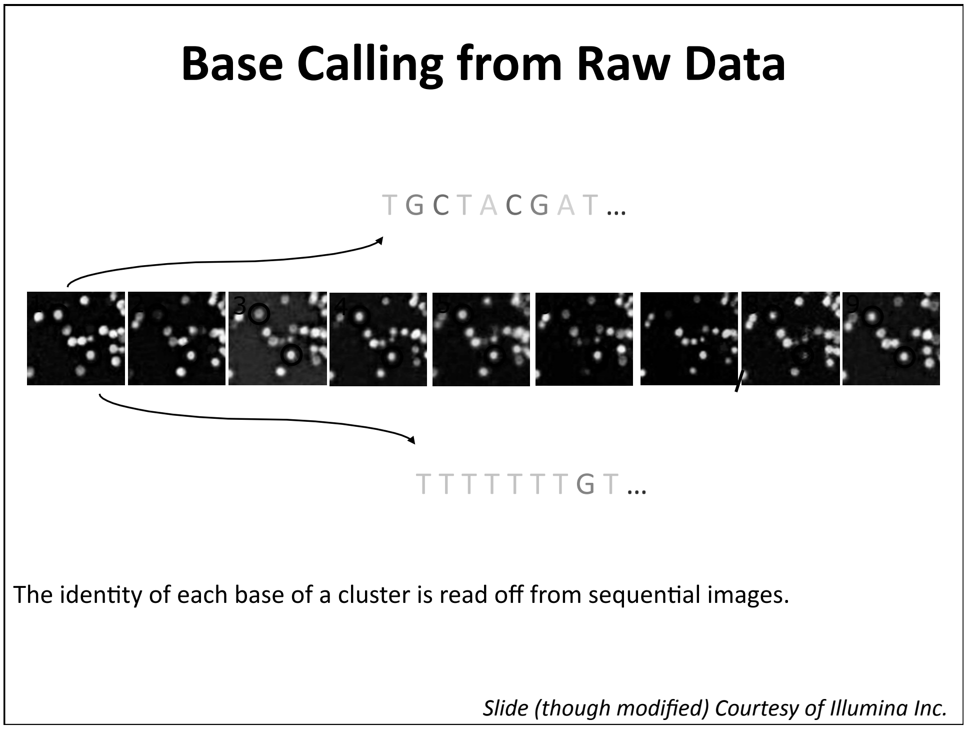
Sequencing & Imaging





Pseudo-color Enhanced Image

Slide (though modified) Courtesy of Illumina Inc.



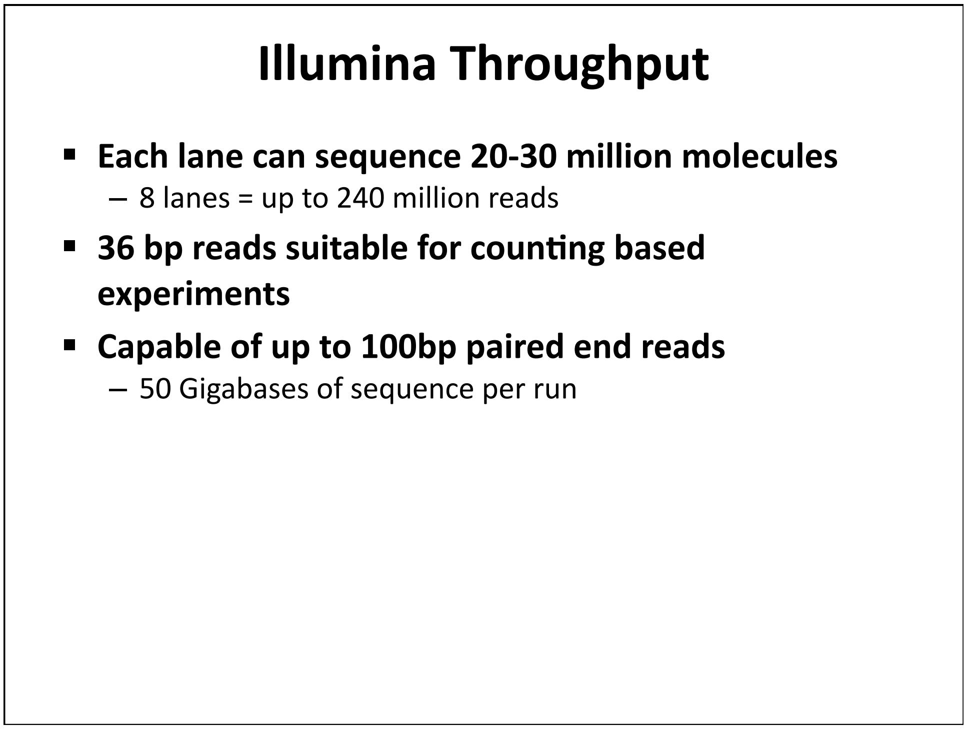
Base Calling from Raw Data

T G C T A C G A T …

T T T T T T T G T …

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

Slide (though modified) Courtesy of Illumina Inc.



Illumina Throughput

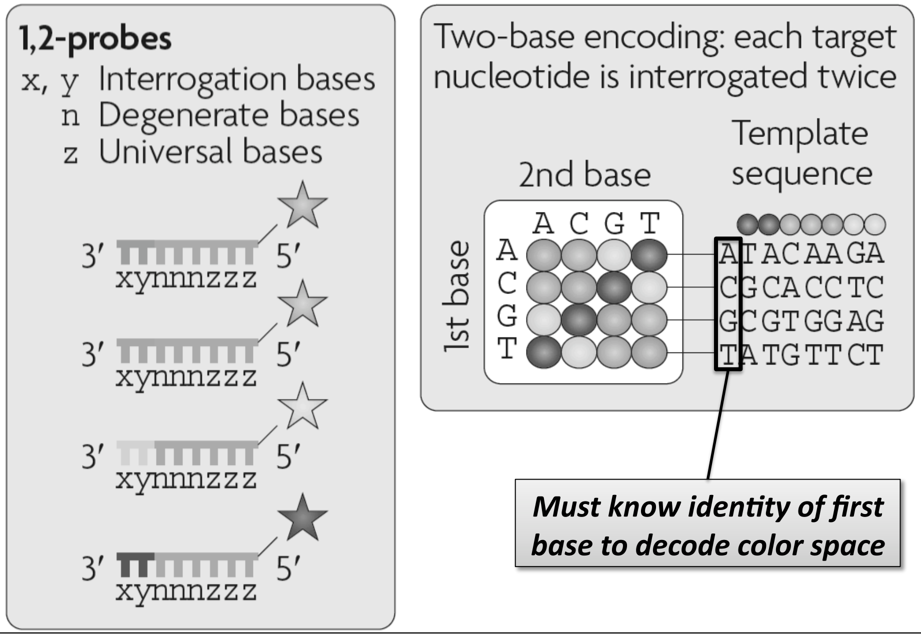
- **Each lane can sequence 20-30 million molecules**
  - 8 lanes = up to 240 million reads
- **36 bp reads suitable for counting based experiments**
- **Capable of up to 100bp paired end reads**
  - 50 Gigabases of sequence per run

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



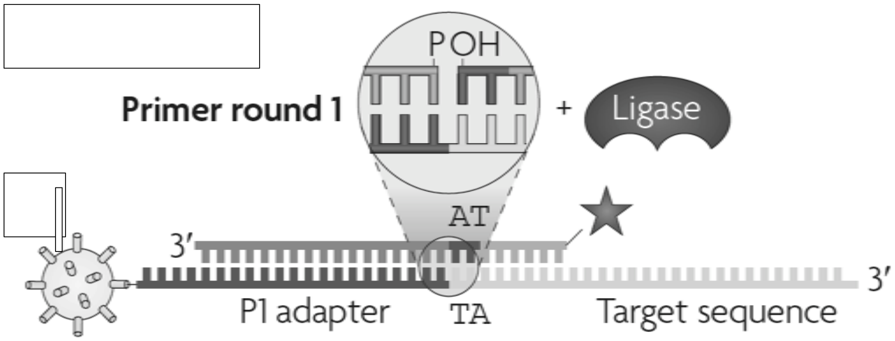
1,2-probes

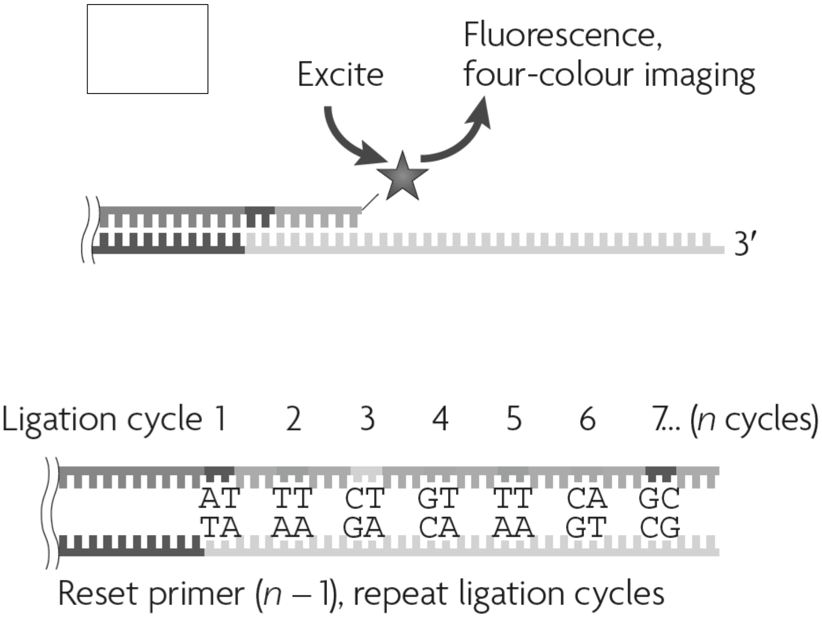
x, y Interrogation bases
n Degenerate bases
z Universal bases

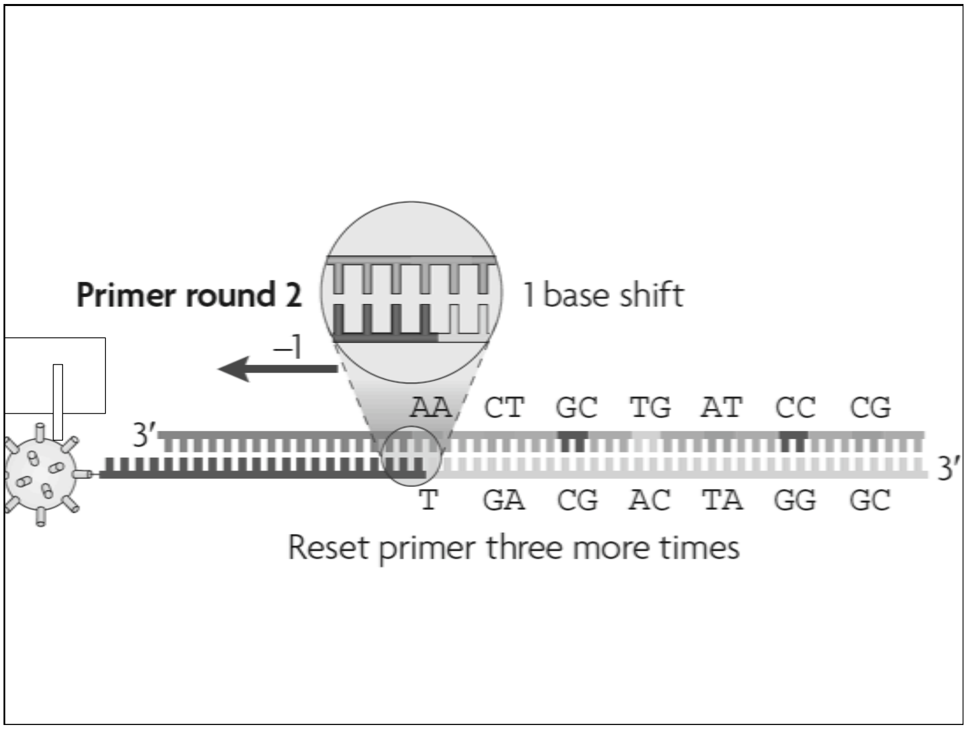
Two-base encoding: each target nucleotide is interrogated twice

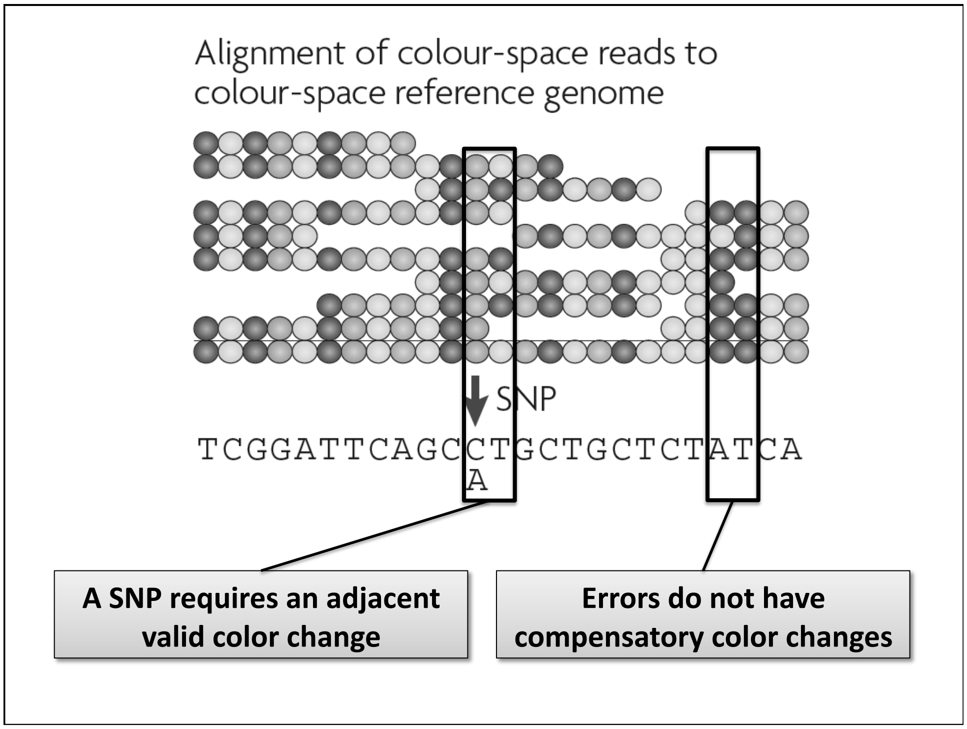
Must know identity of first base to decode color space

Ligation-based extension




Excite
Fluorescence,
four-colour imaging
3′
Ligation cycle 1 2 3 4 5 6 7... (n cycles)
AT TT CT GT TT CA GC
TA AA GA CA AA GT CG
Reset primer (n – 1), repeat ligation cycles

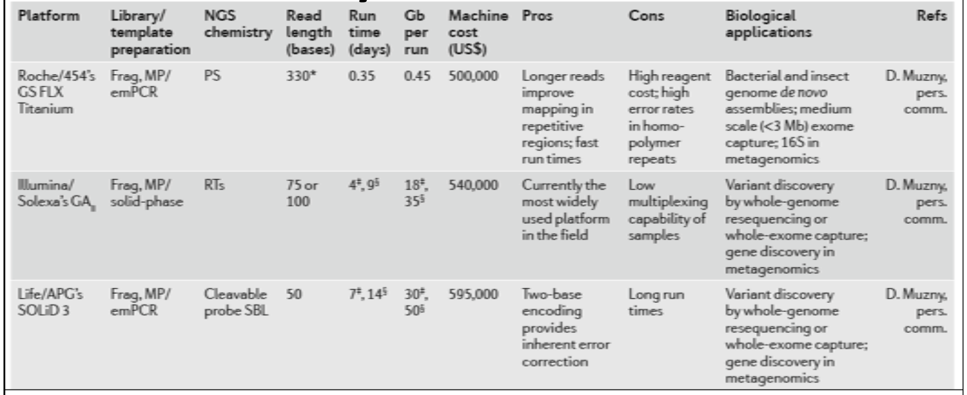



Alignment of colour-space reads to colour-space reference genome
SNP
TCGGATTCAGCCTGCTGCTCTATCA
A

A SNP requires an adjacent valid color change

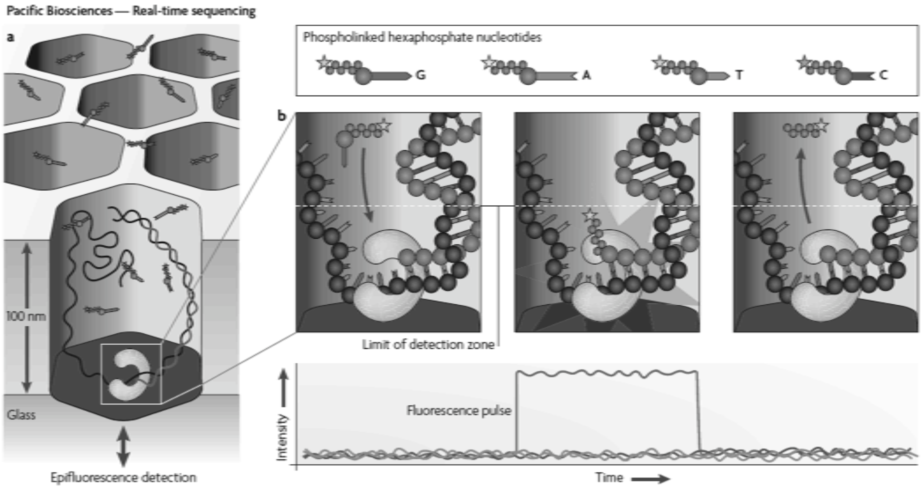
Errors do not have compensatory color changes

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 *de novo* 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	$4^{\ddagger}$, $9^{\S}$	$18^{\ddagger}$, $35^{\S}$	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	$7^{\ddagger}$, $14^{\S}$	$30^{\ddagger}$, $50^{\S}$	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.

Pacific Biosiences

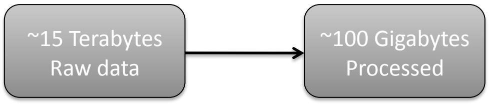

Pacific Biosciences — Real-time sequencing
a
100 nm
Glass
Epifluorescence detection
Phospholinked hexaphosphate nucleotides
G
A
T
C
b
Limit of detection zone
Fluorescence pulse
Intensity
Time

4 September 2008 | www.nature.com/nature | $10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA

Systems engineering challenges

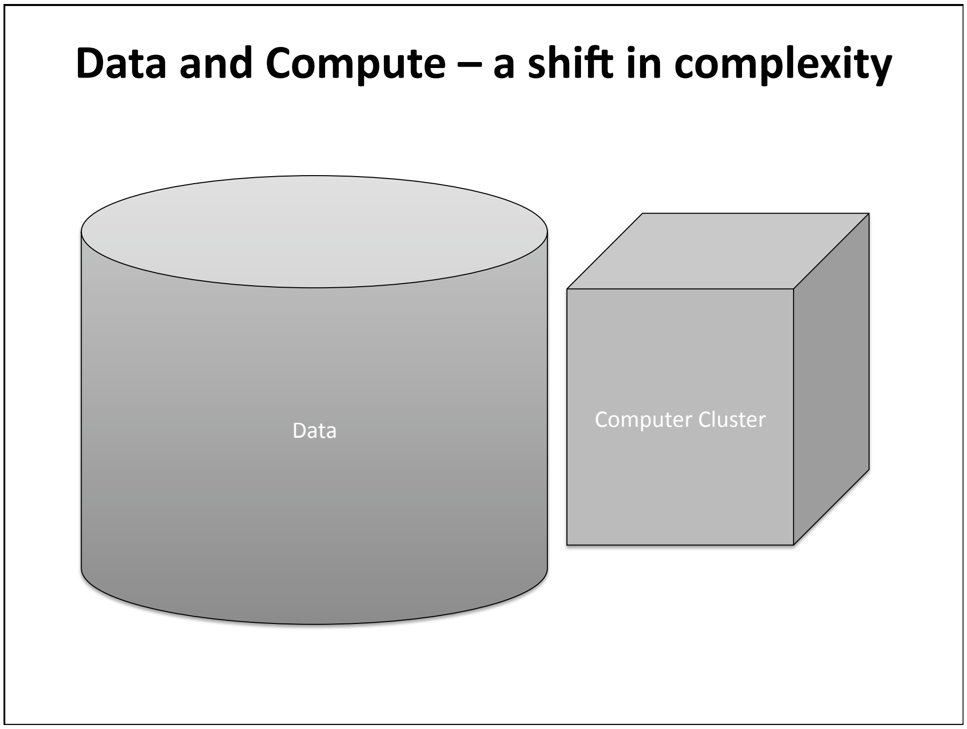
- **We are still learning what the important bits are that need to be stored.**

  Data volumes are MASSIVE, especially short-term

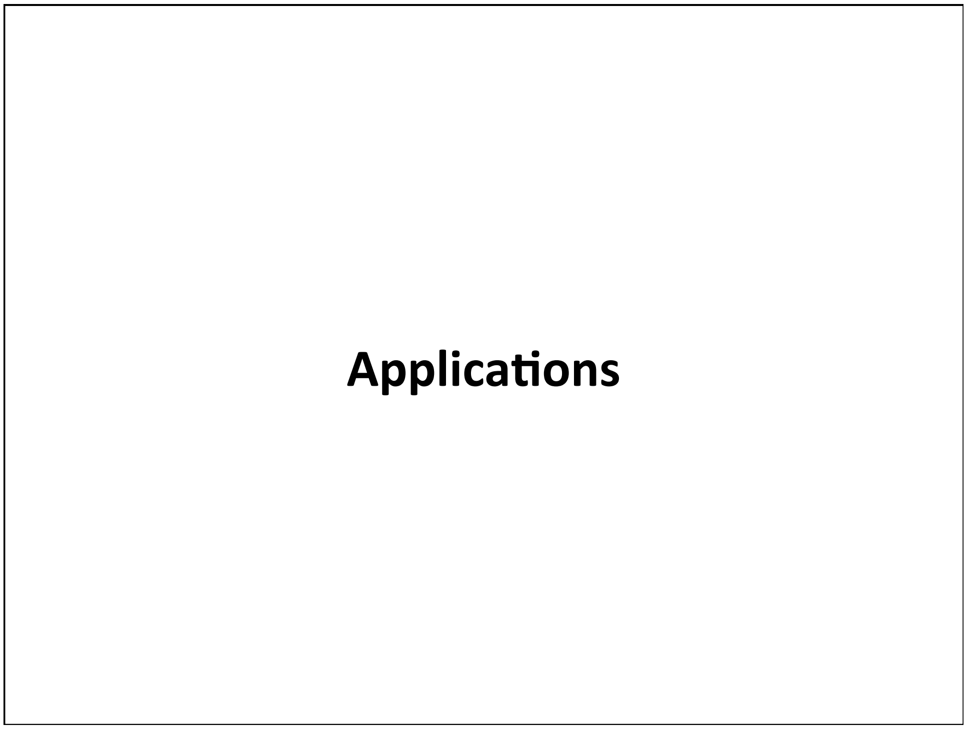


Per Human Genome

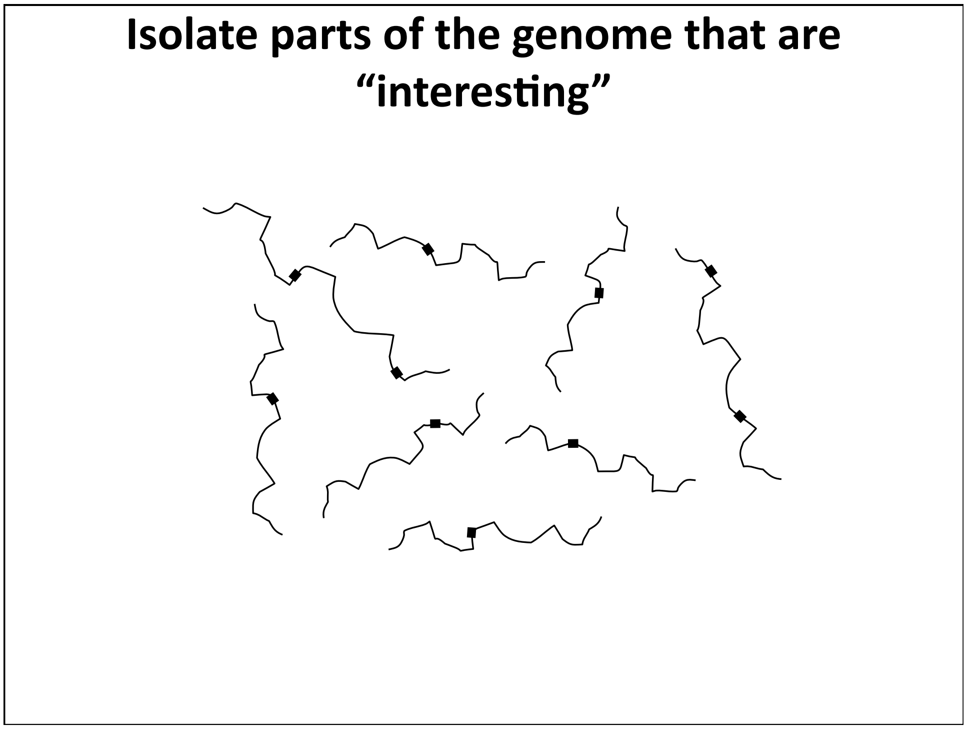
- **Challenges with cloud computing**



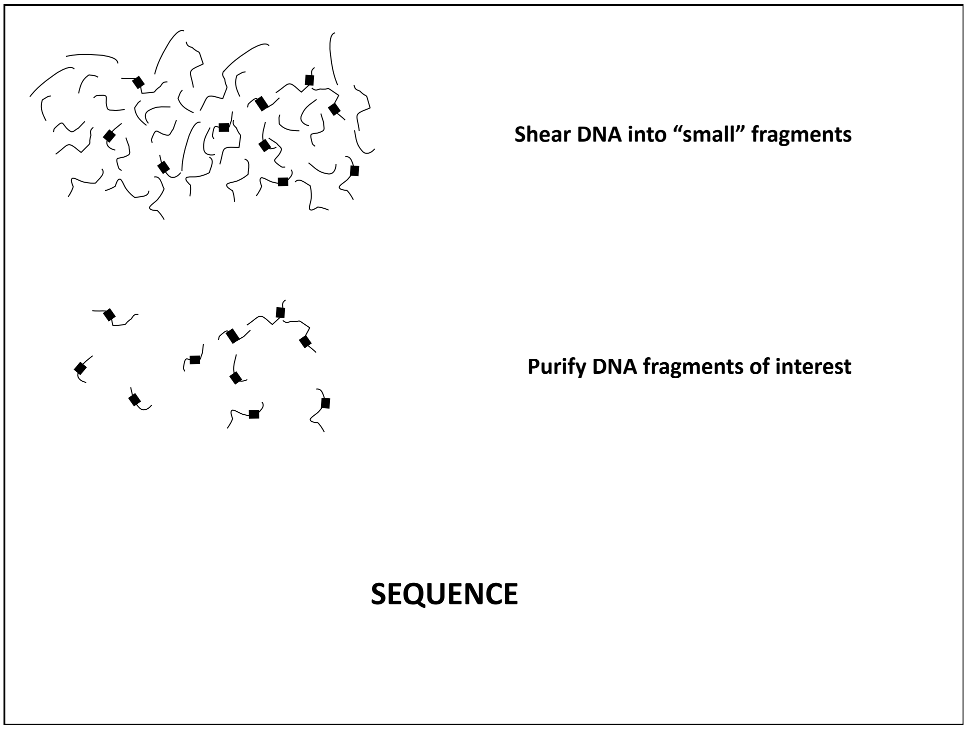
Data and Compute – a shift in complexity



Applications



Isolate parts of the genome that are "interesting"

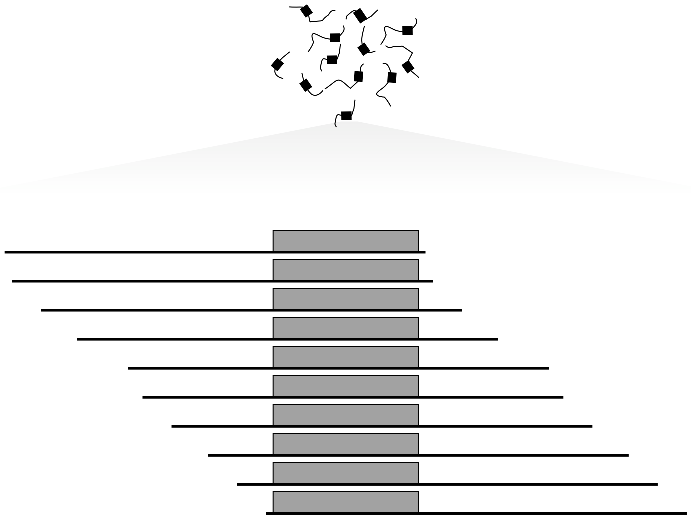


Shear DNA into "small" fragments

Purify DNA fragments of interest

SEQUENCE

How do "counting" experiments work?



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

High-Resolution Profiling of Histone Methylations in the Human Genome

Artem Barski,[1,3] Suresh Cuddapah,[1,3] Kairong Cui,[1,3] Tae-Young Roh,[1,3] Dustin E. Schones,[1,3] Zhibin Wang,[1,3] Gang Wei,[1,3] Iouri Chepelev,[2] and Keji Zhao[1,*]

[1] Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA

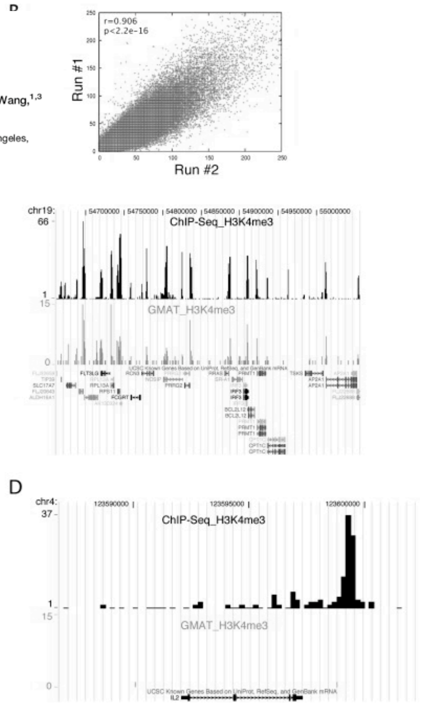
[2] Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA

[3] These authors contributed equally to this work and are listed alphabetically.

*Correspondence: zhaok@nhlbi.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!**



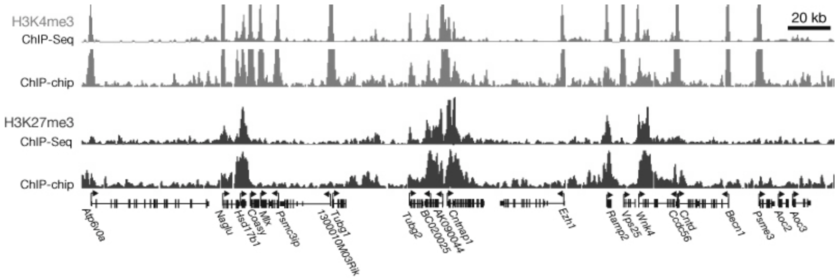
Sequencing-based methods equivalent to Microarray-based methods

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

ARTICLES

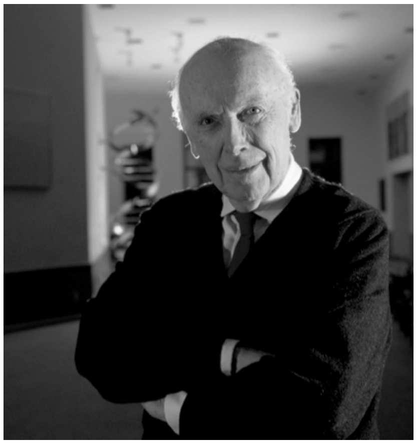
Genome-wide maps of chromatin state in pluripotent and lineage-committed cells

Tarjei S. Mikkelsen[1,2], Manching Ku[1,4], David B. Jaffe[1], Biju Issac[1,4], Erez Lieberman[1,2], Georgia Giannoukos[1], Pablo Alvarez[1], William Brockman[1], Tae-Kyung Kim[5], Richard P. Koche[1,2,4], William Lee[1], Eric Mendenhall[1,4], Aisling O'Donovan[4], Aviva Presser[1], Carsten Russ[1], Xiaohui Xie[1], Alexander Meissner[3], Marius Wernig[3], Rudolf Jaenisch[3], Chad Nusbaum[1], Eric S. Lander[1,3]* & Bradley E. Bernstein[1,4,6]*



Whole Genome Sequencing

First two "personal" genomes sequenced



Watson

Venter

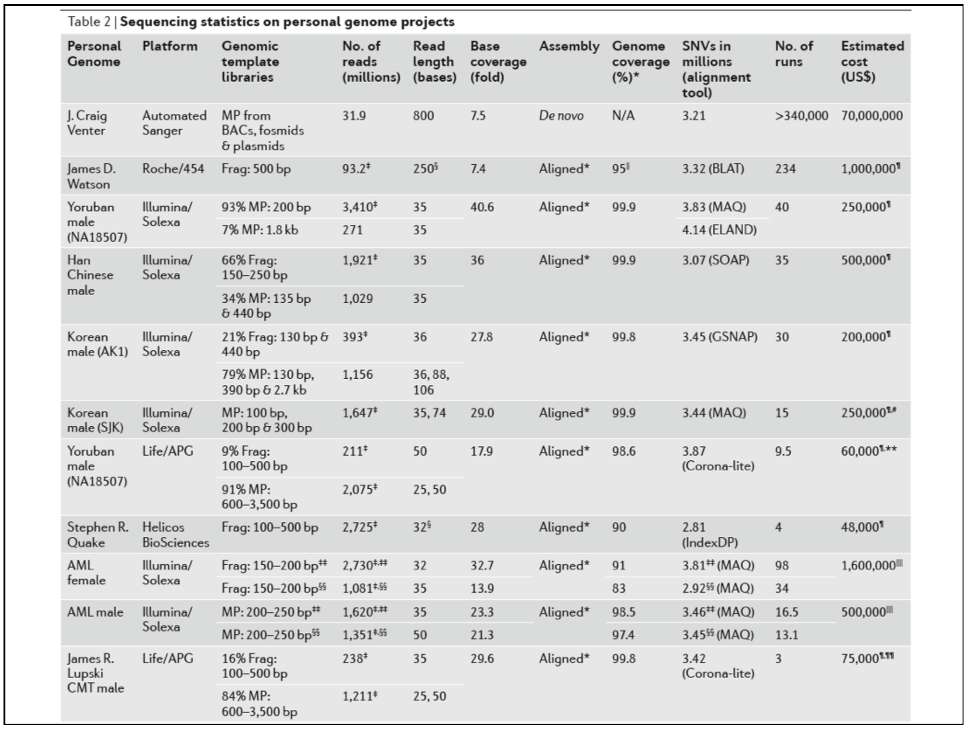
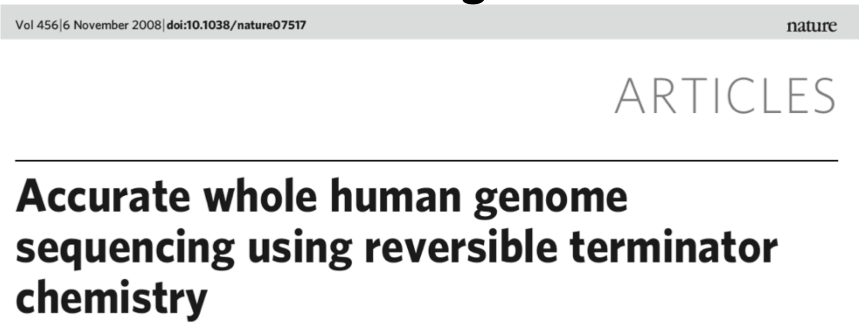


Table 2 | **Sequencing statistics on personal genome projects**

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, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2[‡]	250[§]	7.4	Aligned*	95[‖]	3.32 (BLAT)	234	1,000,000[¶]
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3,410[‡]	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000[¶]
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921[‡]	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000[¶]
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393[‡]	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000[¶]
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647[‡]	35, 74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000[¶,#]
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211[‡]	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000[¶,**]
		91% MP: 600–3,500 bp	2,075[‡]	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725[‡]	32[§]	28	Aligned*	90	2.81 (IndexDP)	4	48,000[¶]
AML female	Illumina/ Solexa	Frag: 150–200 bp[##]	2,730[‡,##]	32	32.7	Aligned*	91	3.81[##] (MAQ)	98	1,600,000[‖‖]
		Frag: 150–200 bp[§§]	1,081[‡,§§]	35	13.9		83	2.92[§§] (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp[##]	1,620[‡,##]	35	23.3	Aligned*	98.5	3.46[##] (MAQ)	16.5	500,000[‖‖]
		MP: 200–250 bp[§§]	1,351[‡,§§]	50	21.3		97.4	3.45[§§] (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238[‡]	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000[¶,¶¶]
		84% MP: 600–3,500 bp	1,211[‡]	25, 50						

What does "whole-genome" mean?



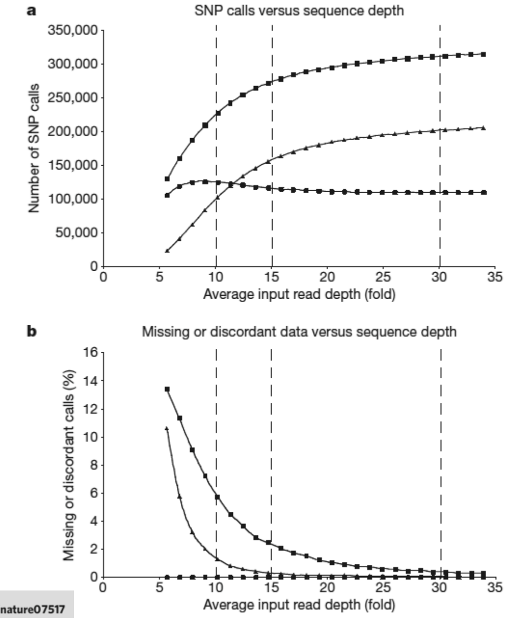
Vol 456|6 November 2008|doi:10.1038/nature07517 nature

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?



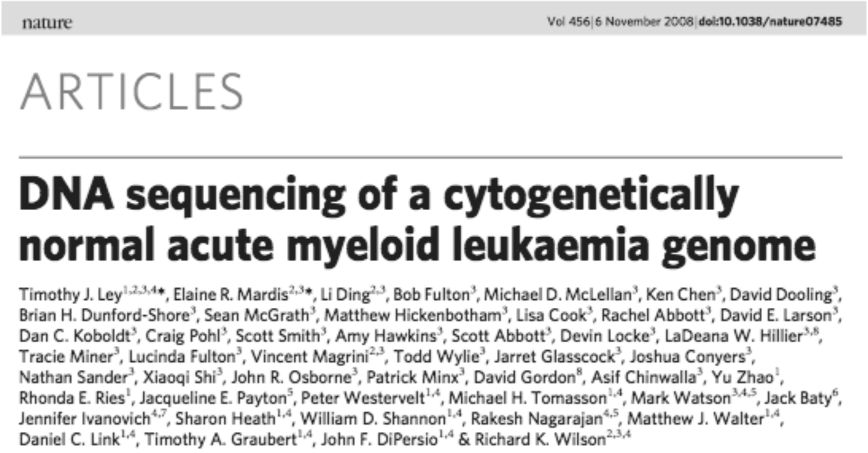
All SNPs

Heterozygous SNPs

Homozygous SNPs

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

Tumor/Normal Whole-Genome Comparison



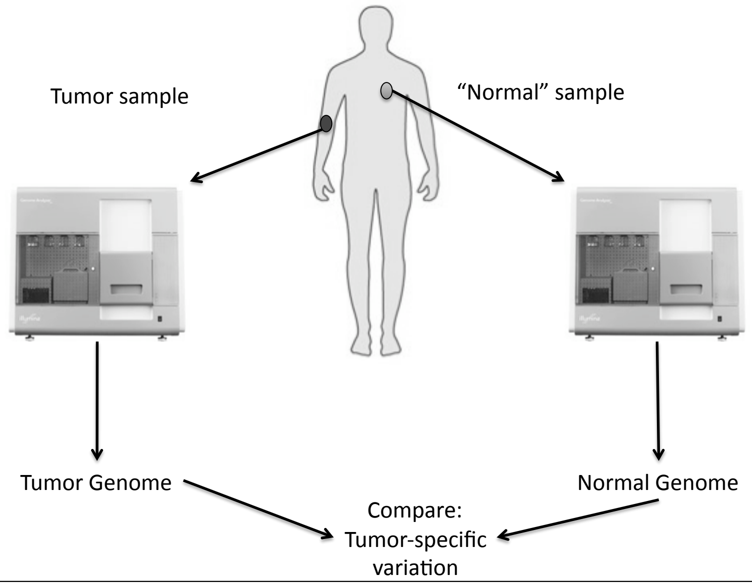
nature   Vol 456 | 6 November 2008 | **doi:10.1038/nature07485**

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 Fulton[3], Michael D. McLellan[3], Ken Chen[3], David Dooling[3], Brian H. Dunford-Shore[3], Sean McGrath[3], Matthew Hickenbotham[3], Lisa Cook[3], Rachel Abbott[3], David E. Larson[3], Dan C. Koboldt[3], Craig Pohl[3], Scott Smith[3], Amy Hawkins[3], Scott Abbott[3], Devin Locke[3], LaDeana W. Hillier[3,8], Tracie Miner[3], Lucinda Fulton[3], Vincent Magrini[2,3], Todd Wylie[3], Jarret Glasscock[3], Joshua Conyers[3], Nathan Sander[3], Xiaoqi Shi[3], John R. Osborne[3], Patrick Minx[3], David Gordon[8], Asif Chinwalla[3], Yu Zhao[1], Rhonda E. Ries[1], Jacqueline E. Payton[5], Peter Westervelt[1,4], Michael H. Tomasson[1,4], Mark Watson[3,4,5], Jack Baty[6], 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



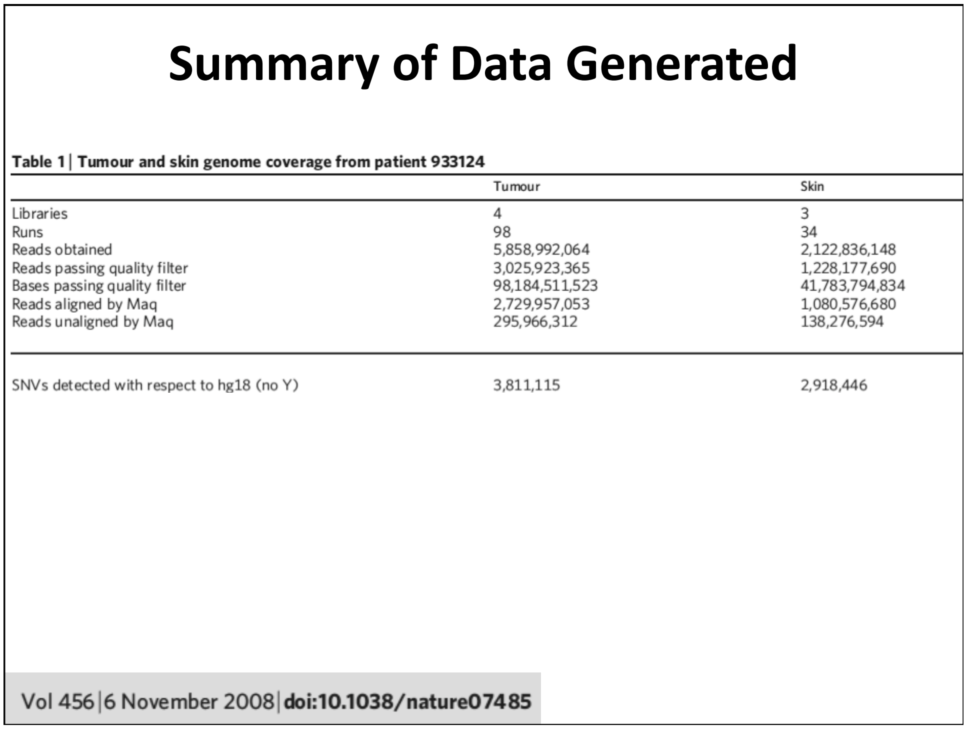
Tumor sample

“Normal” sample

Tumor Genome

Normal Genome

Compare:
Tumor-specific
variation

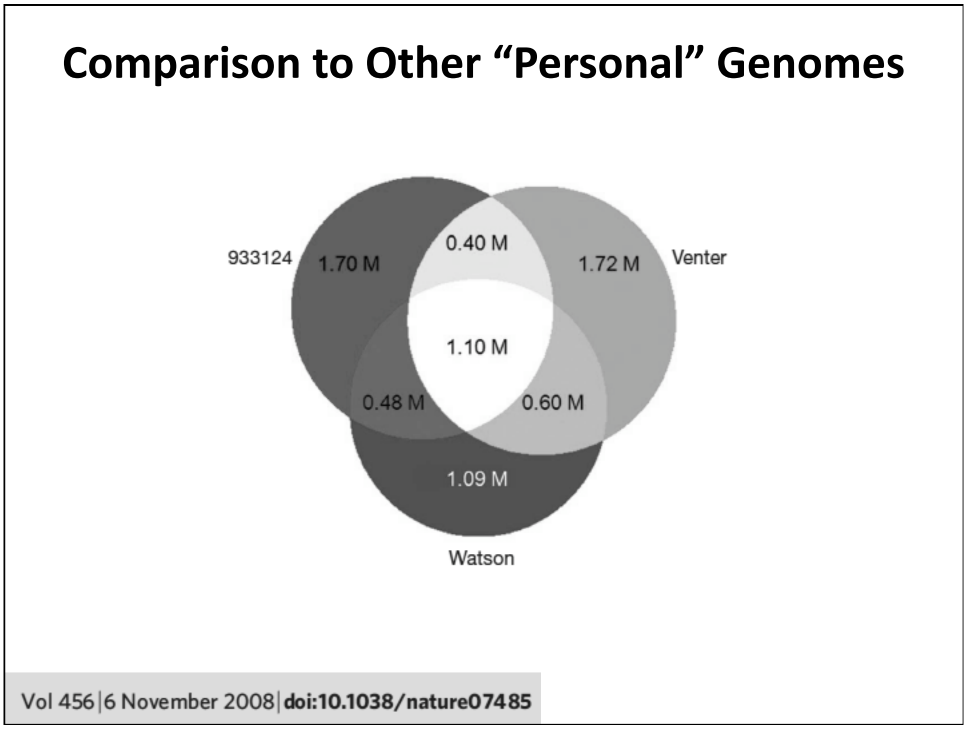


Summary of Data Generated

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

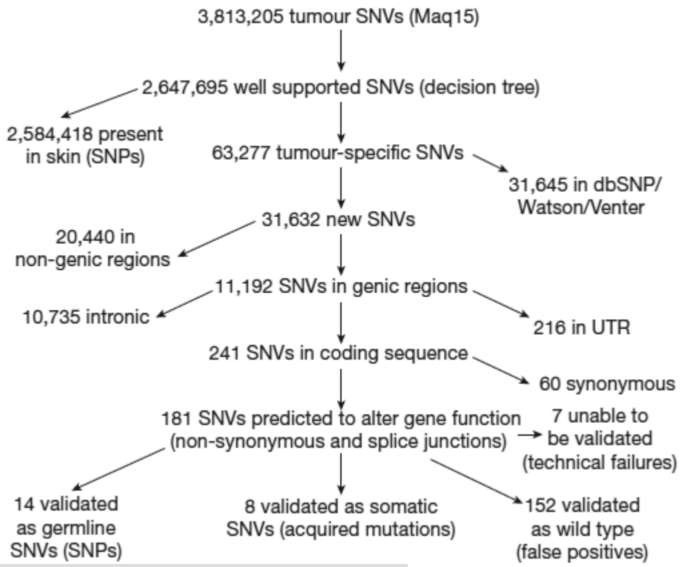
Vol 456 | 6 November 2008 | **doi:10.1038/nature07485**



Comparison to Other “Personal” Genomes

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

Pipeline for Identifying Somatic Mutations



3,813,205 tumour SNVs (Maq15)

2,647,695 well supported SNVs (decision tree)

2,584,418 present in skin (SNPs)

63,277 tumour-specific SNVs

31,645 in dbSNP/ Watson/Venter

31,632 new SNVs

20,440 in non-genic regions

11,192 SNVs in genic regions

10,735 intronic

216 in UTR

241 SNVs in coding sequence

60 synonymous

181 SNVs predicted to alter gene function (non-synonymous and splice junctions)

7 unable to be validated (technical failures)

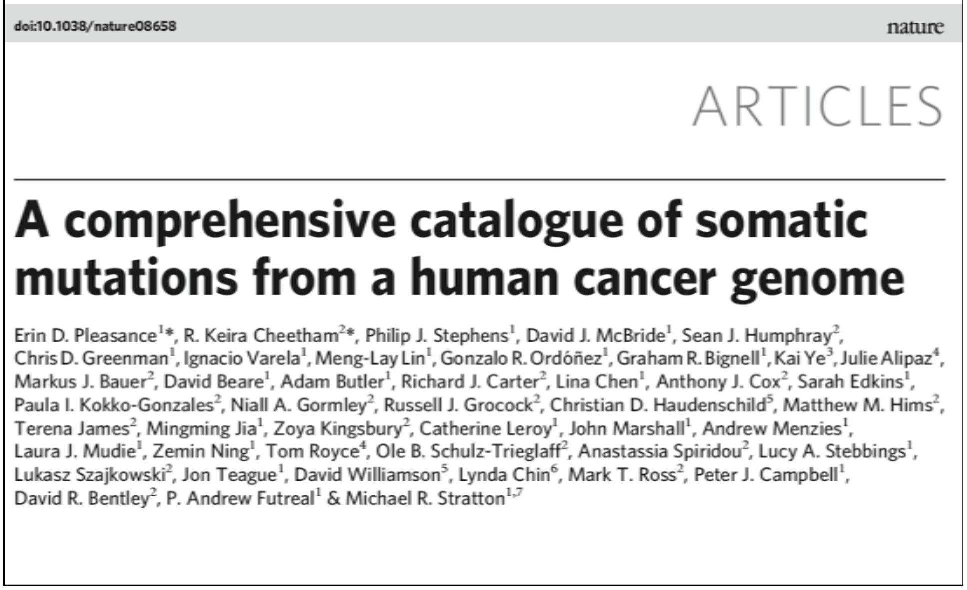
14 validated as germline SNVs (SNPs)

8 validated as somatic SNVs (acquired mutations)

152 validated as wild type (false positives)

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

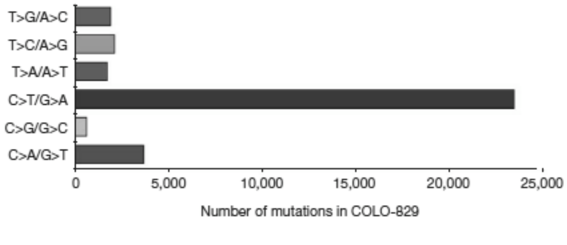
Erin D. Pleasance[1]*, R. Keira Cheetham[2]*, Philip J. Stephens[1], David J. McBride[1], Sean J. Humphray[2], Chris D. Greenman[1], Ignacio Varela[1], Meng-Lay Lin[1], Gonzalo R. Ordóñez[1], Graham R. Bignell[1], Kai Ye[3], Julie Alipaz[4], Markus J. Bauer[2], David Beare[1], Adam Butler[1], Richard J. Carter[2], Lina Chen[1], Anthony J. Cox[2], Sarah Edkins[1], Paula I. Kokko-Gonzales[2], Niall A. Gormley[2], Russell J. Grocock[2], Christian D. Haudenschild[5], Matthew M. Hims[2], Terena James[2], Mingming Jia[1], Zoya Kingsbury[2], Catherine Leroy[1], John Marshall[1], Andrew Menzies[1], Laura J. Mudie[1], Zemin Ning[1], Tom Royce[4], Ole B. Schulz-Trieglaff[2], Anastassia Spiridou[2], Lucy A. Stebbings[1], Lukasz Szajkowski[2], Jon Teague[1], David Williamson[5], Lynda Chin[6], Mark T. Ross[2], Peter J. Campbell[1], David R. Bentley[2], P. Andrew Futreal[1] & Michael R. Stratton[1,7]

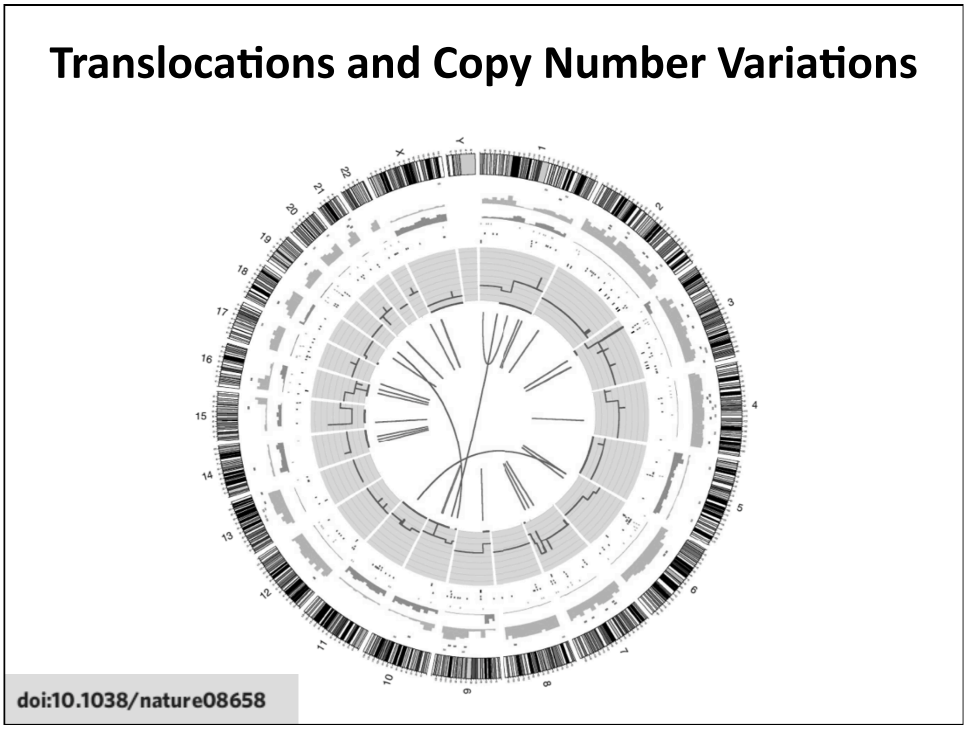
Defining a Somatic Variant

- **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

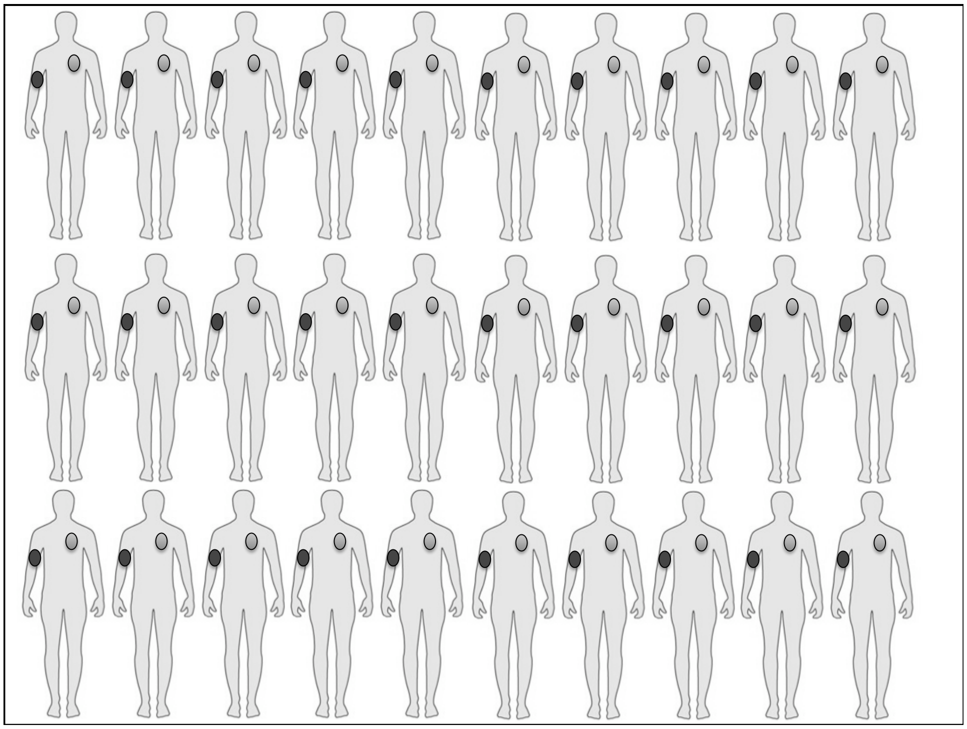
- **32,325 somatic variants detected**
- **Validation against Sanger sequencing data:**
  - 42 of 48 previously known somatic variants detected
    88% sensitivity
  - 452 of 470 newly detected somatic variants confirmed
    3% false positive rate
- **Mutational profile reflective of UV DNA damage:**





Translocations and Copy Number Variations

doi:10.1038/nature08658



National Cancer Institute National Human Genome Research Institute

THE CANCER GENOME ATLAS

cancergenome.nih.gov

1000 Genomes

A Deep Catalog of Human Genetic Variation

www.1000genomes.org