NGS theory: An introduction to next generation sequencing

Richard H. Scheuermann, PhD
Director, J. Craig Venter Institute
La Jolla, CA, USA

J. Craig Venter™
INSTITUTE

J. Craig Venter Institute (JCVI)

- Non-profit research institute
- 200 staff total
  - 120 in La Jolla, CA
  - 80 in Rockville, MD
- By function:
  - 100 wet lab scientists
  - 70 bioinformatics scientists and software engineers
  - 30 administrative support
- Operating budget of about $35 million/year
- 40 faculty manage >100 active sponsored projects

- Human Health: Genomes to Clinic
  - Genomic Medicine
  - Human Microbiome
  - Infectious Disease
- Environmental Sustainability and Discovery
  - Microbial & Environmental Genomics
  - Synthetic Biology
  - Microbial Fuel Cells and Bioenergy
- Platforms
  - Sequencing and Bioinformatics
  - Policy Center
  - Education

Craig Venter™
STITUTE

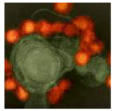
Viral Genomics

Humans are constantly exposed to a plethora of viruses that cause disease. Highly infectious pathogens such as influenza virus, rotavirus, enteroviruses, and respiratory syncytial virus are a persistent public health concern. Recent outbreaks caused by West Nile, Zika, yellow fever, and Middle-East Respiratory Syndrome (MERS) coronavirus remind us of the ongoing threat posed by emerging and re-remerging viruses. The virology group at the JCVI focuses on defining the genomes of viruses and characterizing virus-host interactions to better understand viral evolution, elucidate the underlying molecular mechanisms that cause disease, and develop novel vaccines and therapeutics.

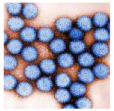
Key Staff >

Research Areas >

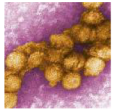
Research Projects



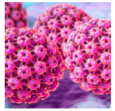
Protective Determinants of ZIKV NS1-specific Antibodies



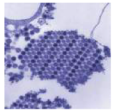
Rotavirus Genome Sequencing



Whole Genome Sequencing of West Nile Virus



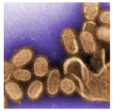
Uncovering Role of Vaginal Microbiome and Connection to HPV Infection and Cervical Cancer



Chikungunya Virus

Development of a Multivalent Rhinovirus Vaccine

Sequencing of Respiratory Syncytial Virus



Sequencing of Influenza A and Influenza B Viruses

Composite


Alerta en Latinoamérica por
EL VIRUS
ZIKA
Conoce las características de este virus de origen asiático y africano que está causando estragos en países de América Latina y el Caribe.
¿QUÉ ES?
Es un arbovirus transmitido por la picadura del mosquito Aedes Aegypti. Es similar al dengue, el chikungunya o la fiebre amarilla. También puede transmitirse vía sexual (se ha detectado virus en el semen) y vía sanguínea, pero son poco frecuentes.
SÍNTOMAS
• Sólo 1 de cada 5 los desarrolla
• Aparecen del día 2 al 12 después del contagio
• Duran entre 4 y 7 días
Los más frecuentes
• Fiebre de menos de 39° C
• Inflamación en manos y pies
• Conjuntivitis no purulenta
DIAGNÓSTICO
Ya que la mayoría de los síntomas se presentan en casos de dengue y chikungunya, se recomienda

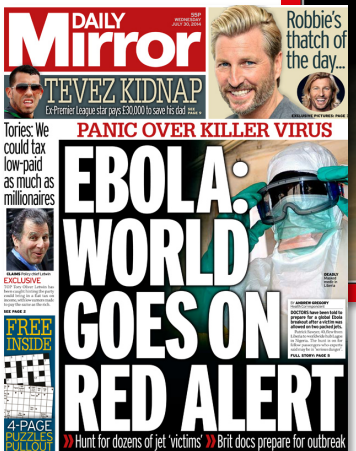

The spread of MERS
Examining the emergence and transfer of the deadly coronavirus


SPORTS ★ FINAL
Monday, April 27, 2009
DAILY NEWS
2.5 MILLION READERS EVERY DAY
NYDailyNews.com
BEANTOWN BEATDOWN
LOST WEEKEND AT FENWAY AS SOX SWEEP YANKS
SEE SPORTS


TIME
AVIAN FLU
DEATH THREAT
SPECIAL REPORT: Inside the global race to avert a pandemic


$1.25 · NYDailyNews.com
SPORTS FINAL
Partly sunny, 62/49. Friday, October 24, 2014
DAILY NEWS
NEW YORK'S HOMETOWN NEWSPAPER
First confirmed city case of deadly virus
NY DOC HAS EBOLA
Craig Spencer in full Ebola gear before his trip to Africa.
A DOCTOR from Harlem was confirmed Thursday night as the city's first victim of the Ebola virus.
PAGES 2,3,4,5,6,7
Harlem
He rode
Bowled
Cuomo,


SWINE FLU SPREADS!
U.S.
nfirmed
4–6


DAILY Mirror
Robbie's thatch of the day...
TEVEZ KIDNAP
Ex-Premier League star pays £30,000 to save his dad
Tories: We could tax low-paid as much as millionaires
PANIC OVER KILLER VIRUS
EBOLA: WORLD GOES ON RED ALERT
Hunt for dozens of jet 'victims'
Brit docs prepare for outbreak
FREE INSIDE
4-PAGE PUZZLES PULLOUT


nature
AVIAN FLU
Ready for a pandemic?


ENTEROVIRUS D68
SPECIAL ONGOING COVERAGE
Abled ALERT
USA
Photo: From pool video
OUTBREAK: Spike In Cases Overwhelms Hospitals | What Parents Need To Know


EXTREME HIGH
COVID-19
PANDEMIC


World Health Organization
Monkeypox

Sequencing and VEME

- Application of NGS to the study of Virus Evolution and Molecular Epidemiology
  - Track the evolution of viruses over time
  - Better understand the selective pressures that drive virus evolution
  - Identify the origins (reservoirs) of outbreak strains
  - Investigate transmission dynamics
  - Identify molecular determinants of host range
  - Identify molecular determinants of virulence
  - Identify evolutionarily conserved regions for targeted vaccines
  - Identify evolutionarily diverse regions for diagnostics

Outline

- A Brief History of Sequencing
- Next Generation Sequencing (NGS) Technologies
- Applications and Challenges of NGS
- Bioinformatics Methods and Resources

Some Trivia

- What year was the first whole genome sequence reported?

  a) 1969 b) 1977 c) 1981 d) 1985

- For which organism?

Some Trivia

- What year was the first whole genome sequence reported?

  a) 1969 *b) 1977* *c) 1981* *d) 1985*

- For which organism?

  Bacteriophage ΦX174 (5,375 bp)

- What method was used?

article

Nature **265**, 687 - 695 (24 February 1977); doi:10.1038/265687a0

Nucleotide sequence of bacteriophage ϕX174 DNA

F. SANGER, G. M. AIR*, B. G. BARRELL, N. L. BROWN†, A. R. COULSON, J. C. FIDDES, C. A. HUTCHISON III‡, P. M. SLOCOMBE§ & M. SMITH¶

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK
Present addresses : *John Curtin School of Medical Research, Microbiology Department, Canberra City ACT 2601, Australia, †Department of Biochemistry, University of Bristol, Bristol BS8 1TD, UK, ‡Department of Bacteriology and Immunology, University of North Carolina, Chapel Hill, North Carolina 27514, §Max-Planck-Institut für Molekulare Genetik, 1 Berlin 33, FRG, ¶Department of Biochemistry, University of British Columbia, Vancouver BC, Canada V6T 1W5,

A DNA sequence for the genome of bacteriophage ϕX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

J. Craig Venter™ INSTITUTE

Some Trivia

- What year was the first whole genome sequence reported?

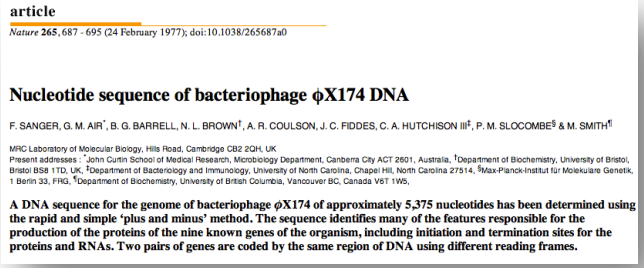
  a) 1969 *b) 1977* *c) 1981* *d) 1985*

- For which organism?

  Bacteriophage ΦX174 (5,375 bp)

- What method was used?

  dideoxy chain termination with ^{32}P (aka Sanger sequencing)



article

Nature **265**, 687 - 695 (24 February 1977); doi:10.1038/265687a0

Nucleotide sequence of bacteriophage ϕX174 DNA

F. SANGER, G. M. AIR*, B. G. BARRELL, N. L. BROWN†, A. R. COULSON, J. C. FIDDES, C. A. HUTCHISON III‡, P. M. SLOCOMBE§ & M. SMITH¶

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK
Present addresses : *John Curtin School of Medical Research, Microbiology Department, Canberra City ACT 2601, Australia, †Department of Biochemistry, University of Bristol, Bristol BS8 1TD, UK, ‡Department of Bacteriology and Immunology, University of North Carolina, Chapel Hill, North Carolina 27514, §Max-Planck-Institut für Molekulare Genetik, 1 Berlin 33, FRG, ¶Department of Biochemistry, University of British Columbia, Vancouver BC, Canada V6T 1W5,

A DNA sequence for the genome of bacteriophage ϕX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

J. Craig Venter™ INSTITUTE

Some Trivia

- What year was the first whole genome sequence reported?

  a) 1969 b) 1977 c) 1981 d) 1985

- For which organism?

  Bacteriophage ΦX174 (5,375 bp)

- What method was used?

  dideoxy chain termination with ^{32}P (aka Sanger sequencing)

- What year was the first whole genome sequence for a free-living organism reported?

  a) 1979 b) 1984 c) 1989 d) 1995

- For which organism?

Some Trivia

- What year was the first whole genome sequence reported?

  a) 1969 b) 1977 c) 1981 d) 1985

- For which organism?

  Bacteriophage ΦX174 (5,375 bp)

- What method was used?

  dideoxy chain termination with ^{32}P (aka Sanger sequencing)

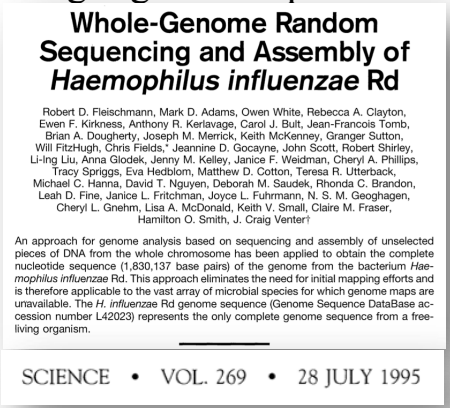
- What year was the first whole genome sequence for a free-living organism reported?

  a) 1979 b) 1984 c) 1989 d) 1995

- For which organism?

  Haemophilus influenza (1.8 x 10^6 bp)

- What method was used?



Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd

Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields,* Jeannine D. Gocayne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Hanna, David T. Nguyen, Deborah M. Saudek, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, N. S. M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, J. Craig Venter†

An approach for genome analysis based on sequencing and assembly of unselected pieces of DNA from the whole chromosome has been applied to obtain the complete nucleotide sequence (1,830,137 base pairs) of the genome from the bacterium *Haemophilus influenzae* Rd. This approach eliminates the need for initial mapping efforts and is therefore applicable to the vast array of microbial species for which genome maps are unavailable. The *H. influenzae* Rd genome sequence (Genome Sequence DataBase accession number L42023) represents the only complete genome sequence from a free-living organism.

SCIENCE • VOL. 269 • 28 JULY 1995

J. Craig Venter™ INSTITUTE

Some Trivia

- What year was the first whole genome sequence reported?

  a) 1969 *b) 1977* *c) 1981* *d) 1985*

- For which organism?

  Bacteriophage ΦX174 (5,375 bp)

- What method was used?

  dideoxy chain termination with ^{32}P (aka Sanger sequencing)

- What year was the first whole genome sequence for a free-living organism reported?

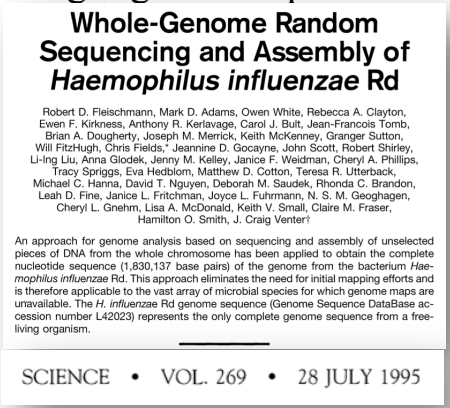
  a) 1979 *b) 1984* *c) 1989* *d) 1995*

- For which organism?

  Haemophilus influenza (1.8 x 10^6 bp)

- What method was used?

  Sanger sequencing with fluorescence



Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd

Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields,* Jeannine D. Gocayne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Hanna, David T. Nguyen, Deborah M. Saudek, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, N. S. M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, J. Craig Venter†

An approach for genome analysis based on sequencing and assembly of unselected pieces of DNA from the whole chromosome has been applied to obtain the complete nucleotide sequence (1,830,137 base pairs) of the genome from the bacterium *Haemophilus influenzae* Rd. This approach eliminates the need for initial mapping efforts and is therefore applicable to the vast array of microbial species for which genome maps are unavailable. The *H. influenzae* Rd genome sequence (Genome Sequence DataBase accession number L42023) represents the only complete genome sequence from a free-living organism.

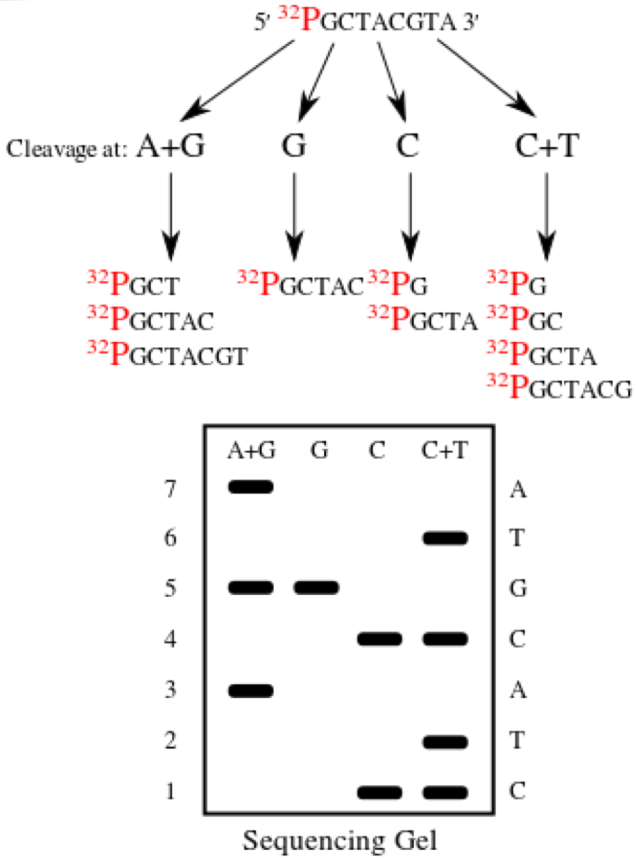
SCIENCE • VOL. 269 • 28 JULY 1995

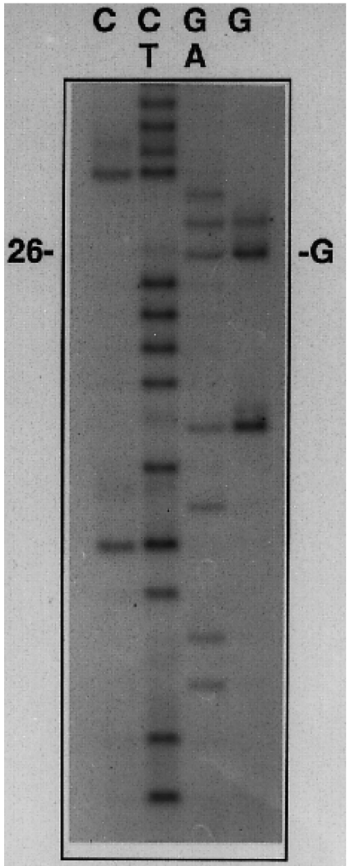
J. Craig Venter™ INSTITUTE

1st Generation Nucleic Acid Sequencing

- Maxim-Gilbert chemical method
  - Maxam AM, Gilbert W (1977). "A new method for sequencing DNA". Proc. Natl. Acad. Sci. U.S.A. 74 (2): 560–4.
- Sanger chain termination method
  - Sanger F, Coulson AR (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". J. Mol. Biol. 94 (3): 441–8.
  - Sanger F, Nicklen S, Coulson AR (1977). "DNA sequencing with chain-terminating inhibitors". Proc. Natl. Acad. Sci. U.S.A. 74 (12): 5463–7.

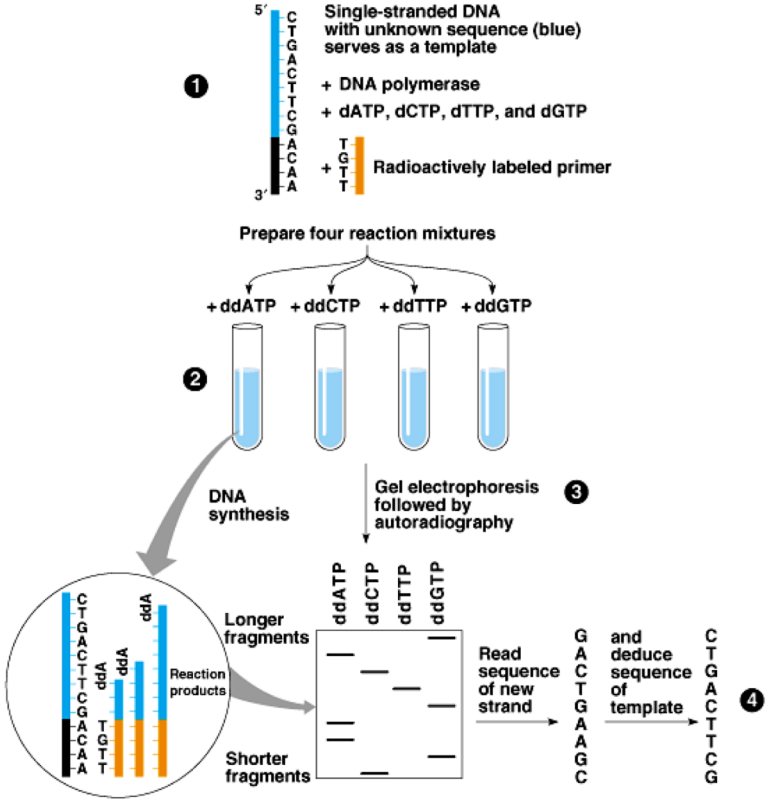
Maxim-Gilbert – chemical cleavage

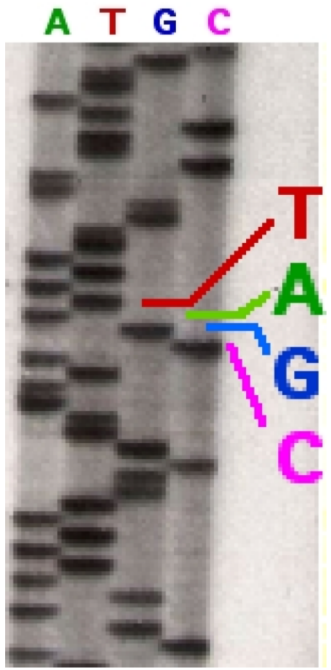




J. Craig Venter™ INSTITUTE

Sanger - chain termination

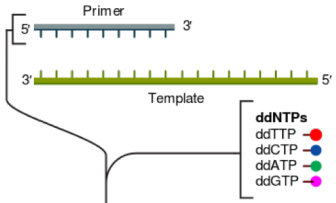




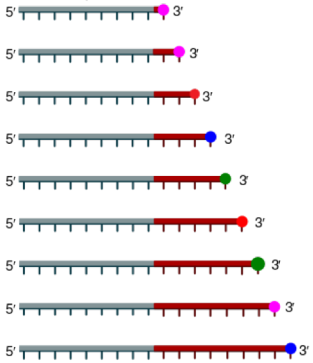
J. Craig Venter™ INSTITUTE

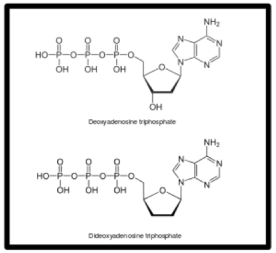
Sanger 1st => Sanger 2nd

① Reaction mixture
- Primer and DNA template
- DNA polymerase
- ddNTPs with flourochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)

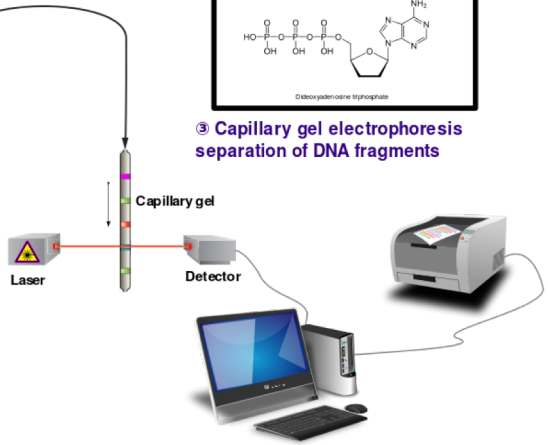


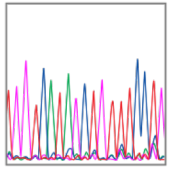
② Primer elongation and chain termination





③ Capillary gel electrophoresis separation of DNA fragments



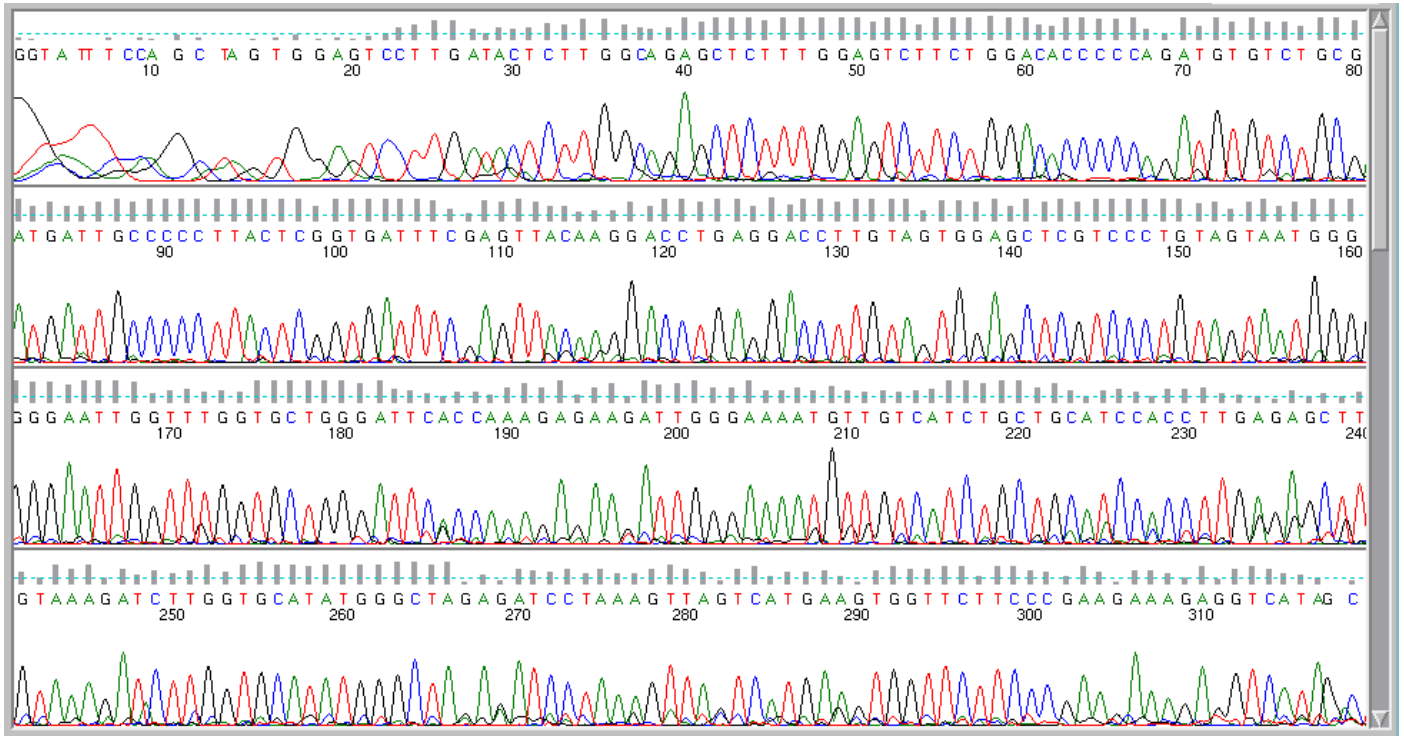


Chromatograph

④ Laser detection of flourochromes and computational sequence analysis

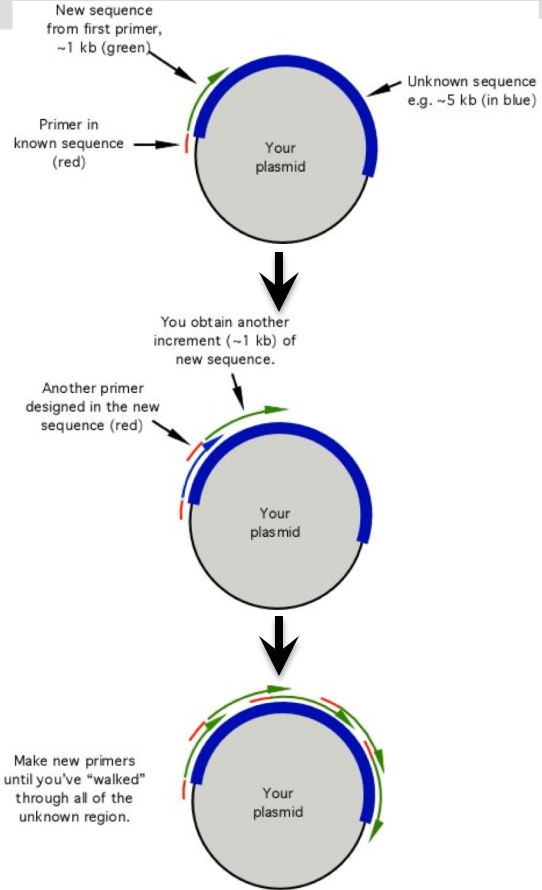
J. Craig Venter INSTITUTE

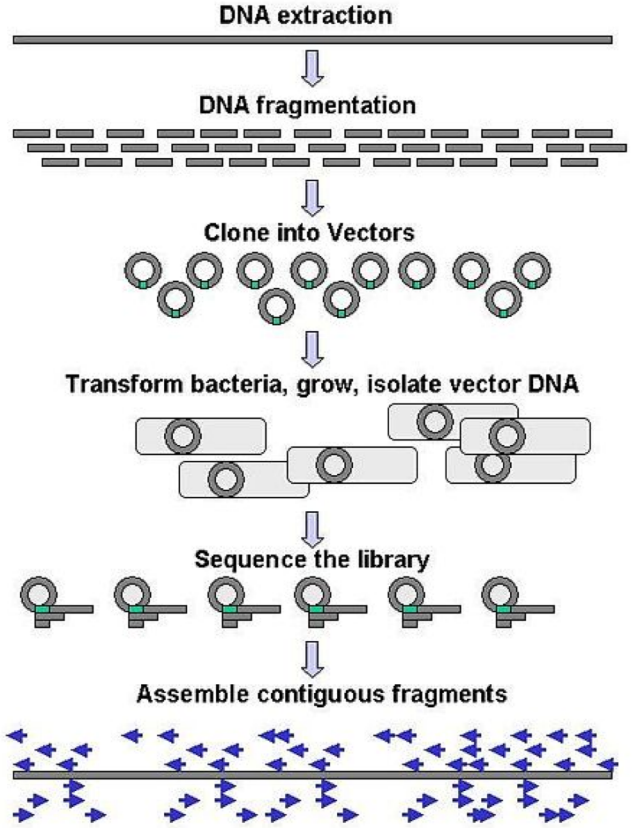
A chromatogram



J. Craig Venter™ INSTITUTE

Primer Walking vs Shotgun





J. Craig Venter™ INSTITUTE

Next Generation Sequencing (NGS)

- Massively parallel; sequencing by synthesis or ligation
- Common attributes
  - Random fragmentation of starting DNA
  - Ligation with custom linkers/adapters
  - Library amplification on a solid surface (either bead or glass slide)
  - Direct step-by-step detection of each nucleotide base incorporated during the sequencing reaction
  - Hundreds of thousands to hundreds of millions of reactions imaged per instrument run = "massively parallel sequencing"
  - Shorter read lengths than capillary sequencers

454 GS FLX

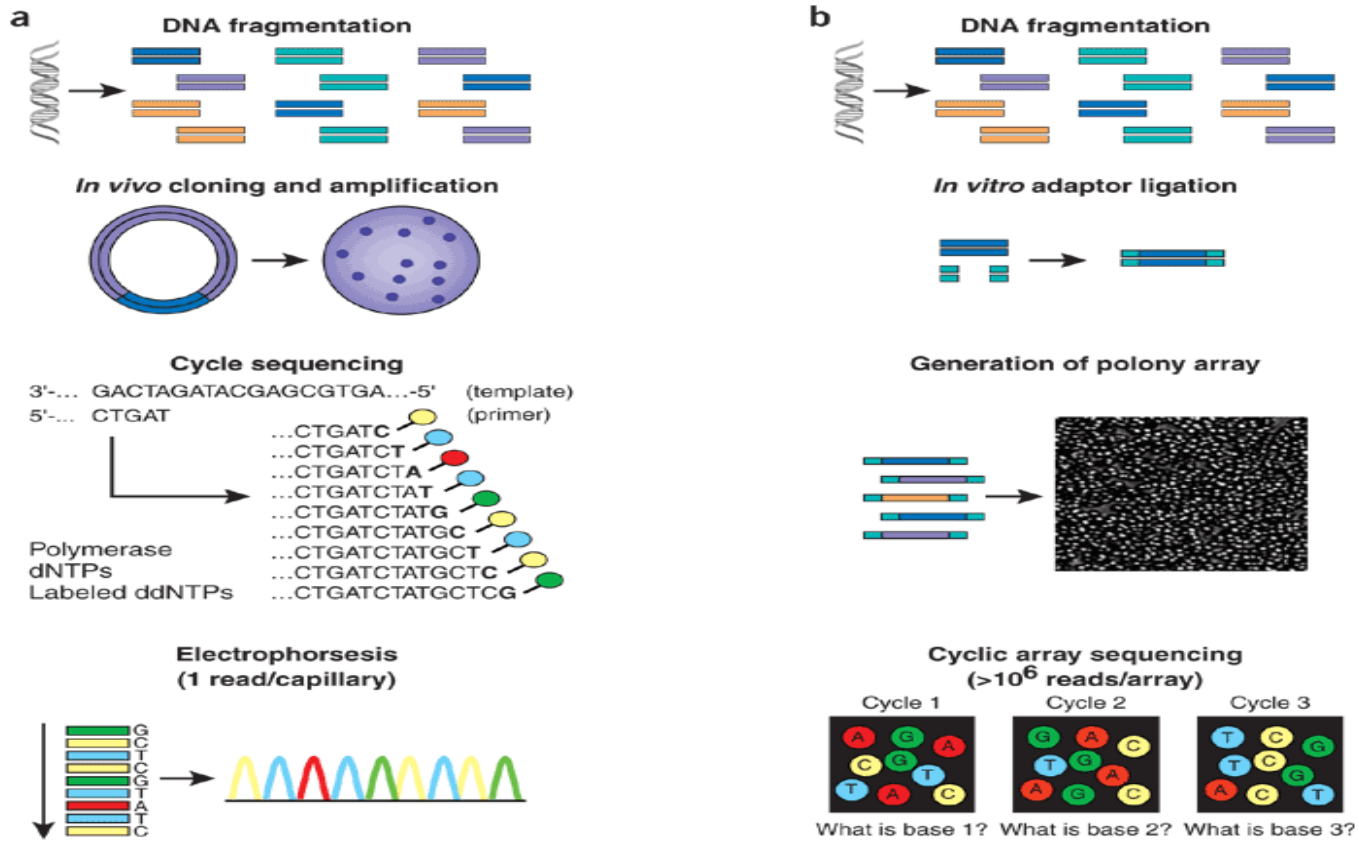
Illumina HiSeq 2000

Illumina MiSeq

Ion PGM

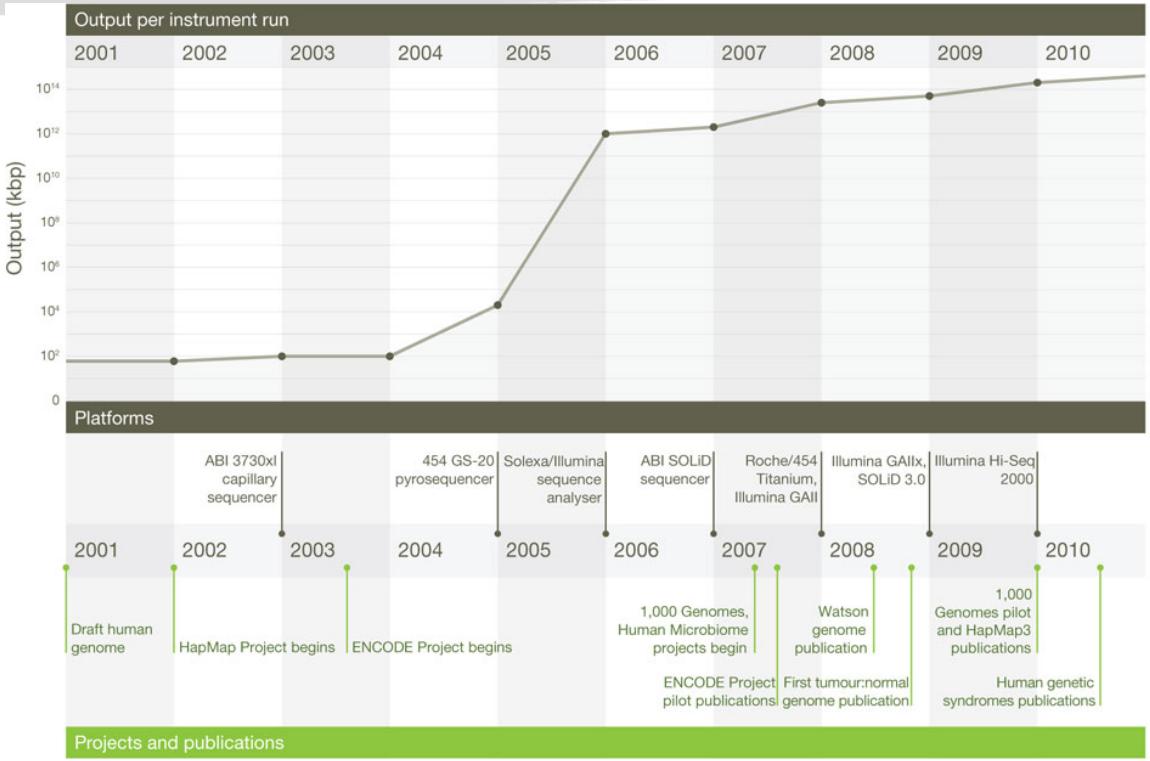
J. Craig Venter™ INSTITUTE

1st Generation vs Next Generation



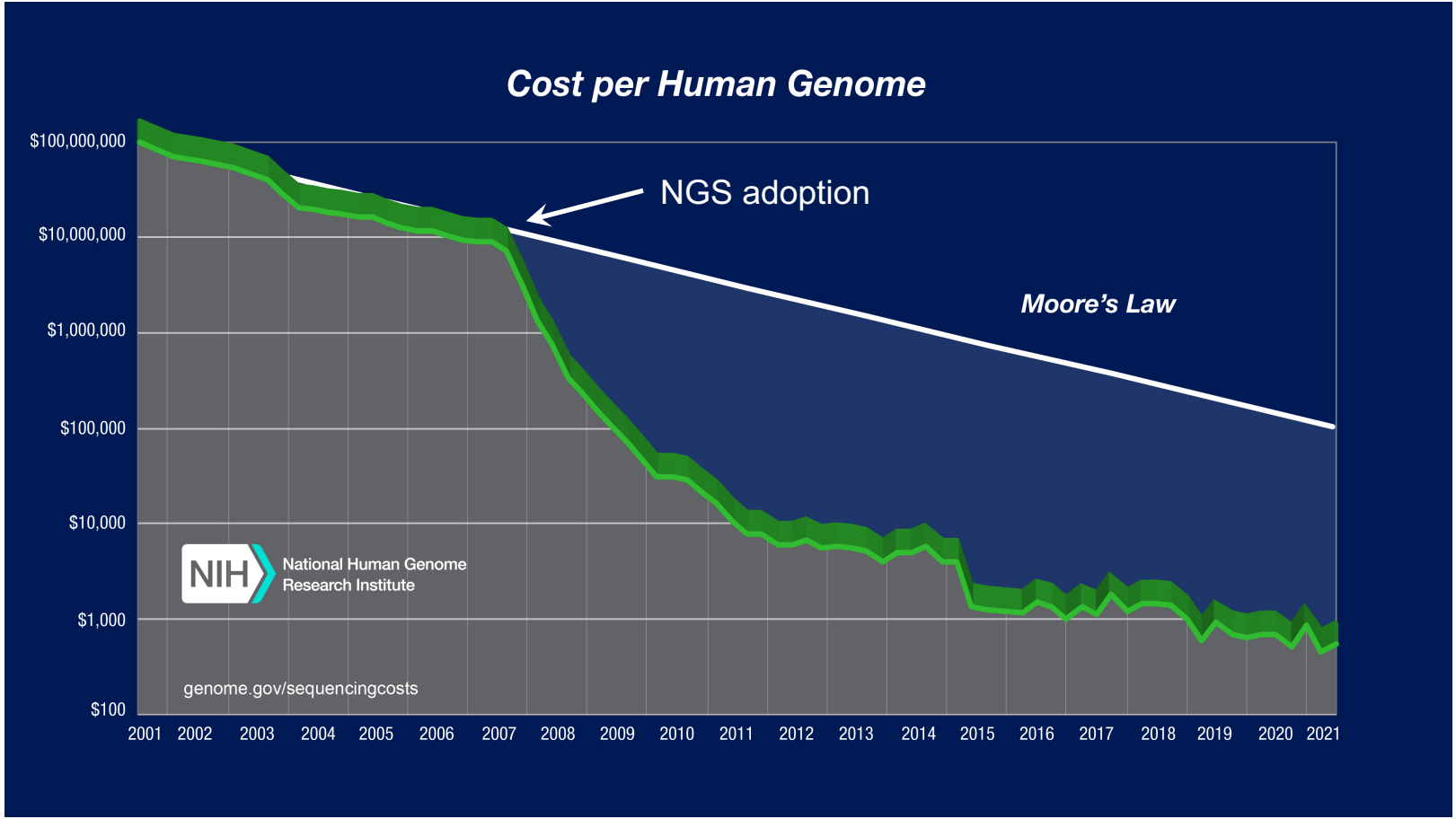
J. Craig Venter™ INSTITUTE

Change in Output



ER Mardis. *Nature* **470**, 198-203 (2011) doi:10.1038/nature09796

J. Craig Venter™ INSTITUTE



Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: http://www.genome.gov/sequencingcosts/. Accessed 15JUL2019.

JCVI Joint Technology Core

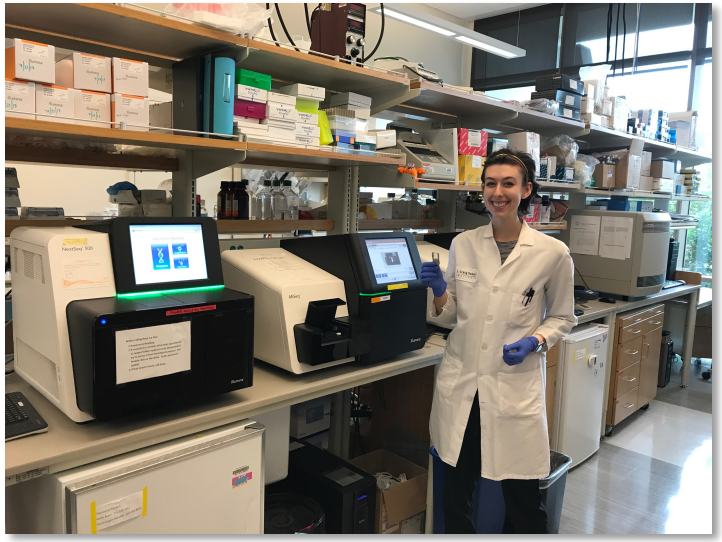
ABI 3730xl

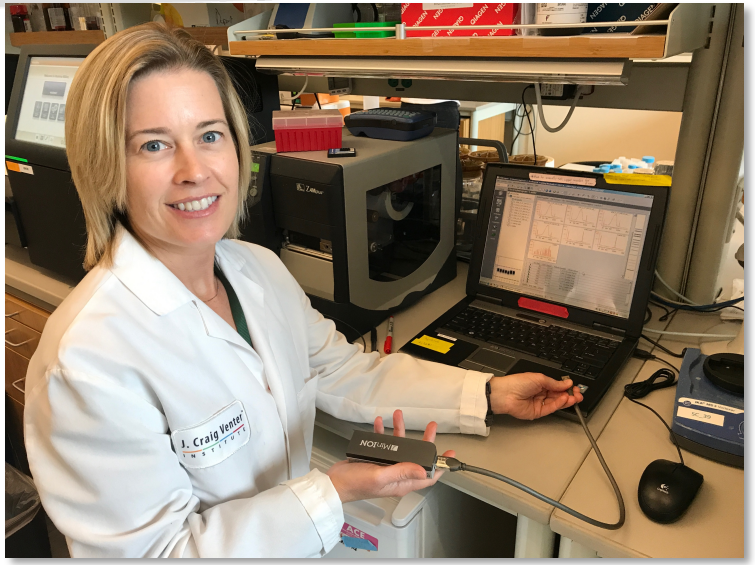
Capacity: 240,000 sequences/day or 80 million lanes/year at 24 runs per day

J. Craig Venter™ INSTITUTE

New JCVI Sequencing Core

Illumina NextSeq/MiSeq
800 million reads/runs

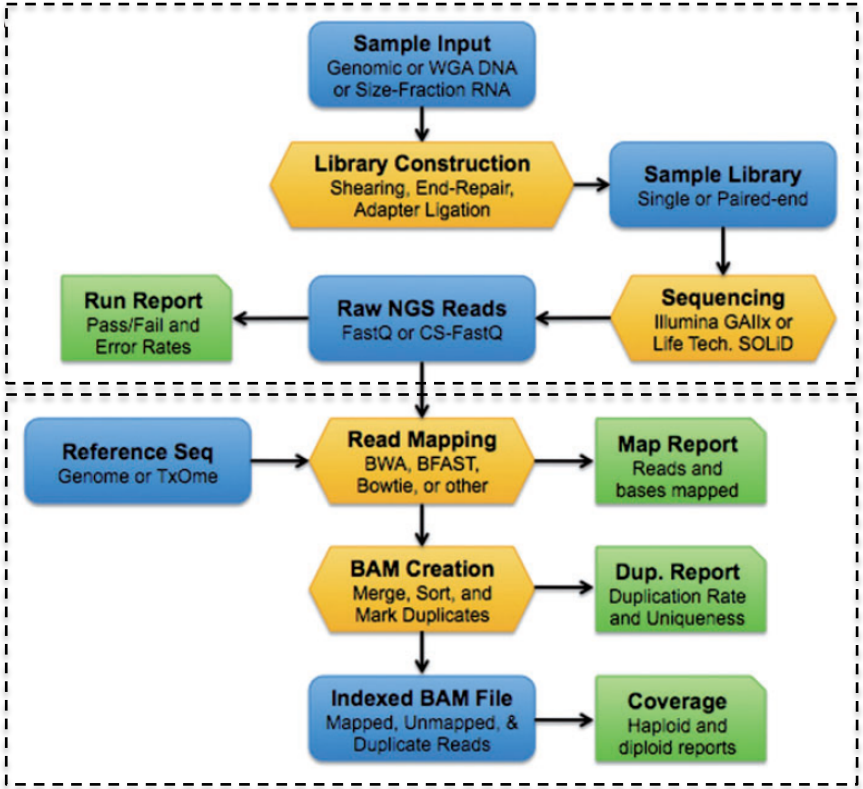




Oxford Nanopore
MinION

J. Craig Venter™
INSTITUTE

General Workflow



Wet lab

Dry lab

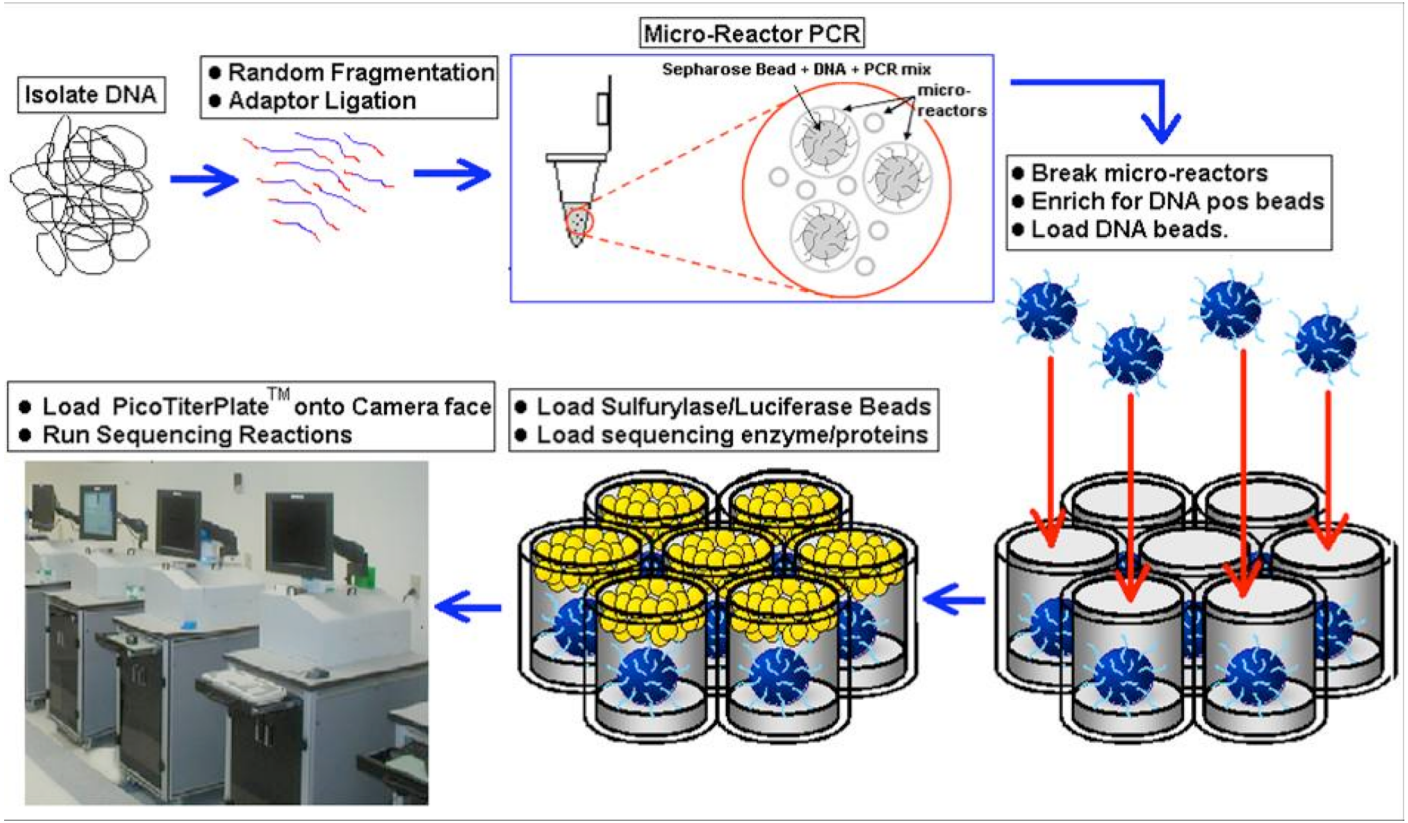
D Koboldt *et al. Briefings in Bioinformatics* (2010)

J. Craig Venter™ INSTITUTE

Major NGS Technologies (2nd Generation)

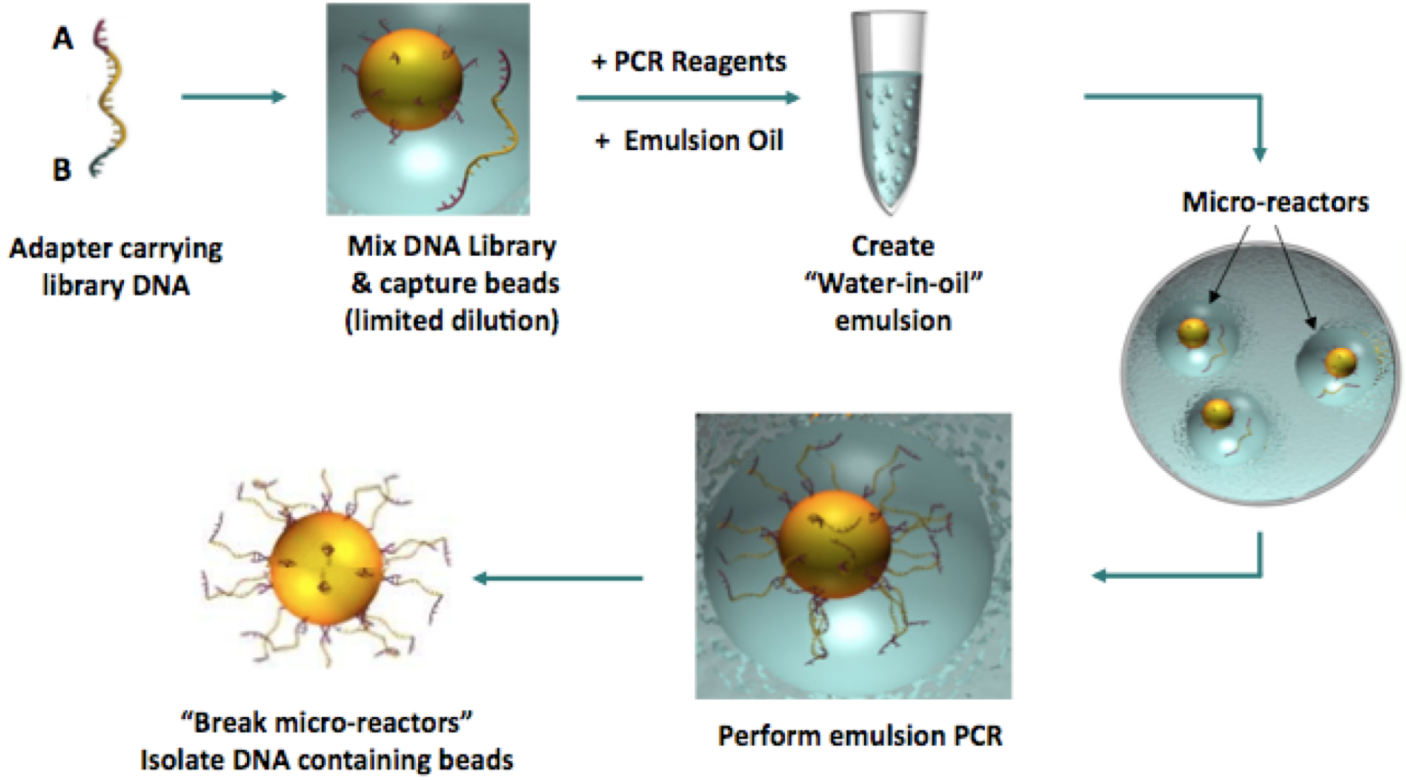
- 454 Life Sciences/Roche
- Ion Torrent/PGM/Life Technologies
- Illumina HiSeq/MiSeq/NovaSeq
- SOLiD/ABI

454 Sequencing



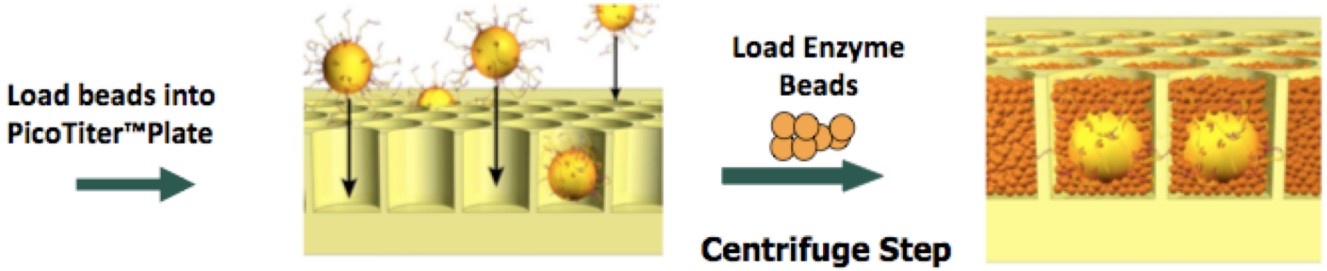
J. Craig Venter™ INSTITUTE

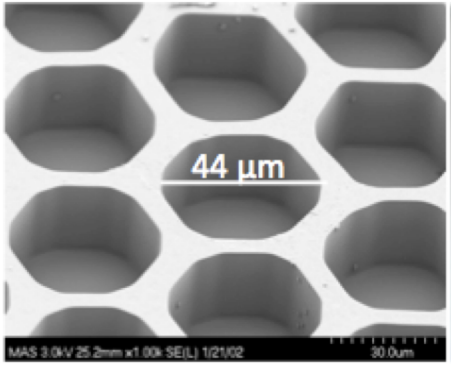
Emulsion PCR

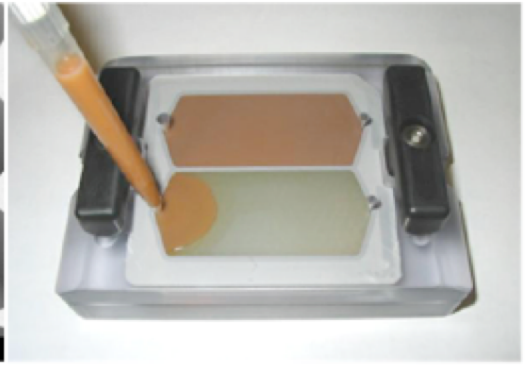


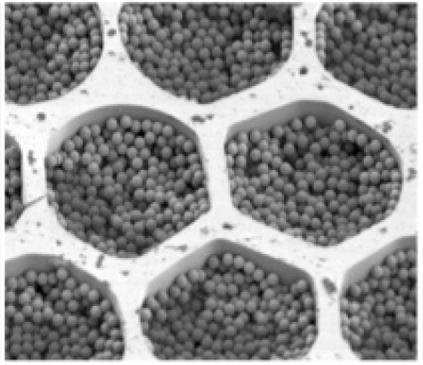
J. Craig Venter™ INSTITUTE

Picotiter Plates



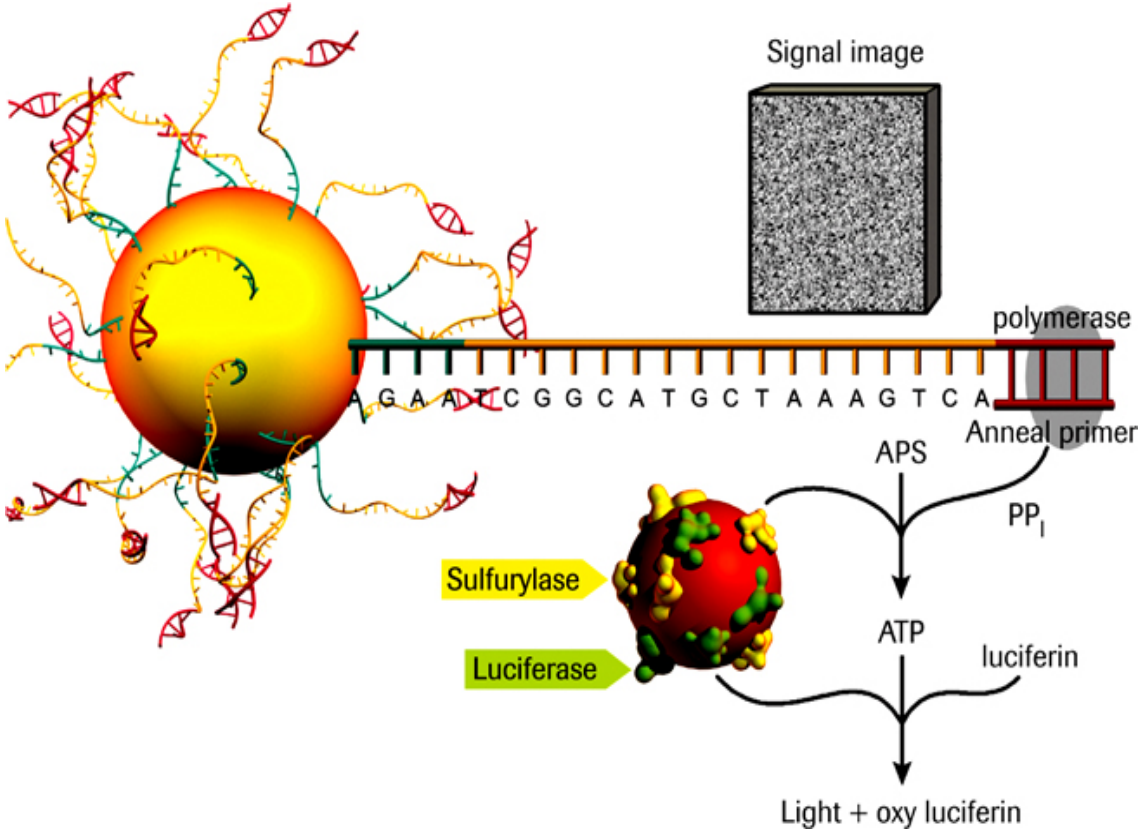






J. Craig Venter™
INSTITUTE

Pyrosequencing



J. Craig Venter™ INSTITUTE

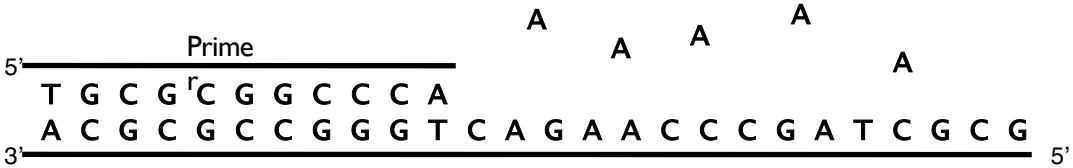
Shotgun Sequencing by 454

Only give polymerase one nucleotide at a time:

If that nucleotide is incorporated, enzymes turn PPi by-products into light:

J. Craig Venter™ INSTITUTE

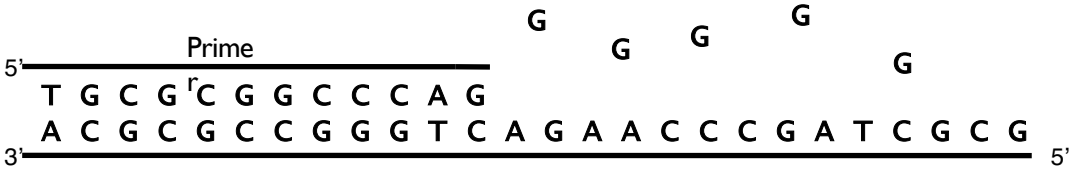
Only give polymerase one nucleotide at a time:



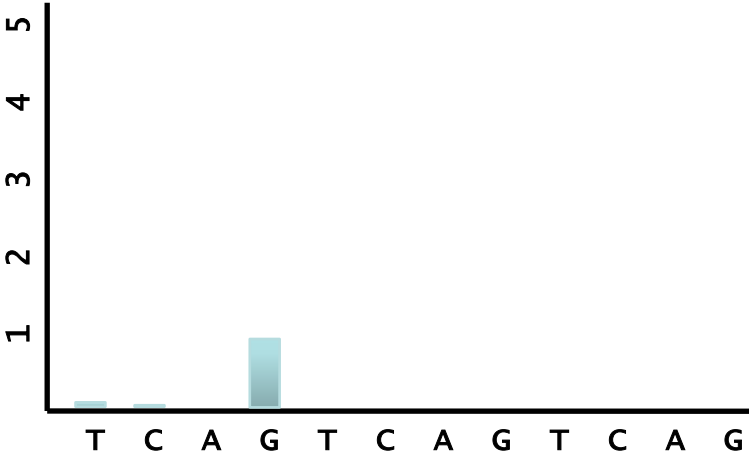
If that nucleotide is incorporated, enzymes turn PPi by-products into light:

J. Craig Venter™ INSTITUTE

Only give polymerase one nucleotide at a time:

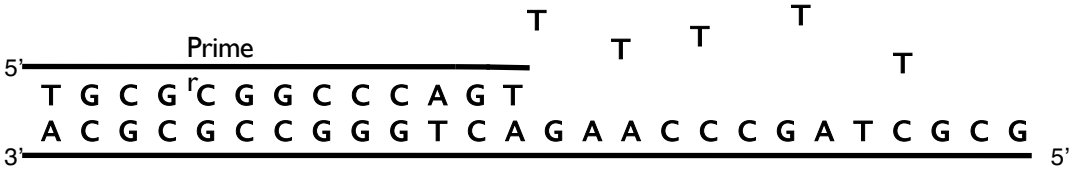


If that nucleotide is incorporated, enzymes turn PPi by-products into light:

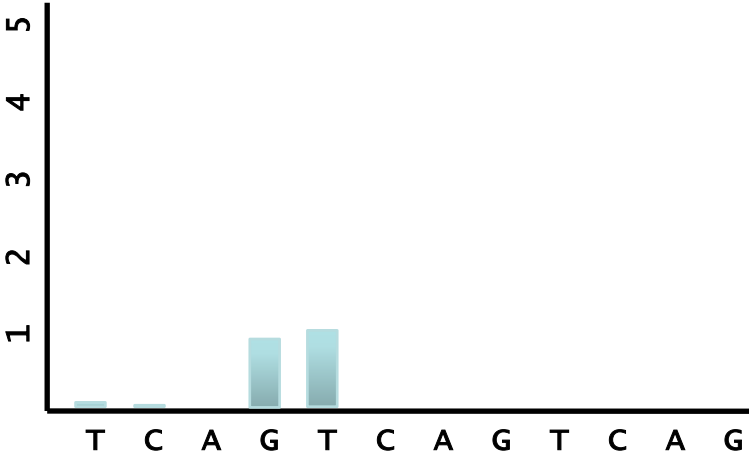


J. Craig Venter™ INSTITUTE

Only give polymerase one nucleotide at a time:

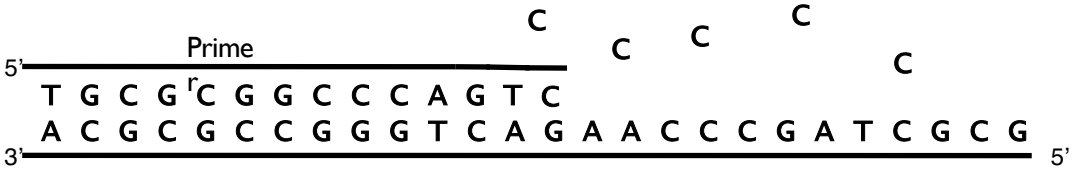


If that nucleotide is incorporated, enzymes turn PPi by-products into light:

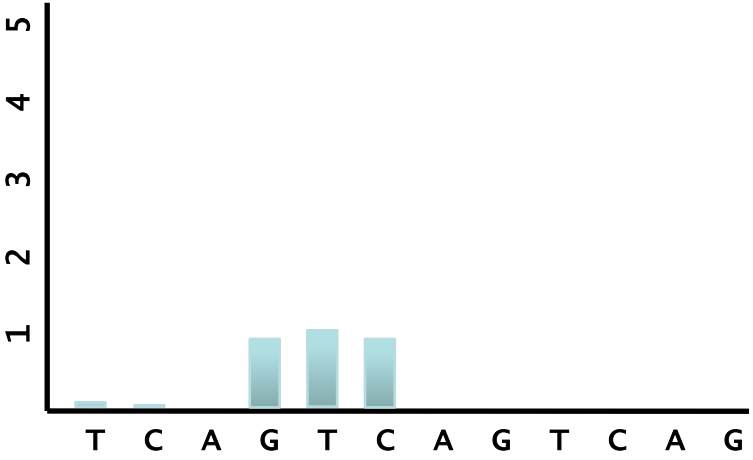


J. Craig Venter INSTITUTE

Only give polymerase one nucleotide at a time:

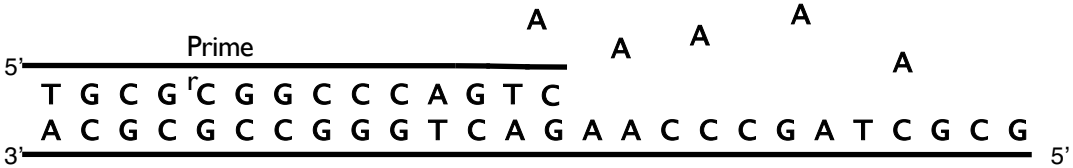


If that nucleotide is incorporated, enzymes turn PPi by-products into light:

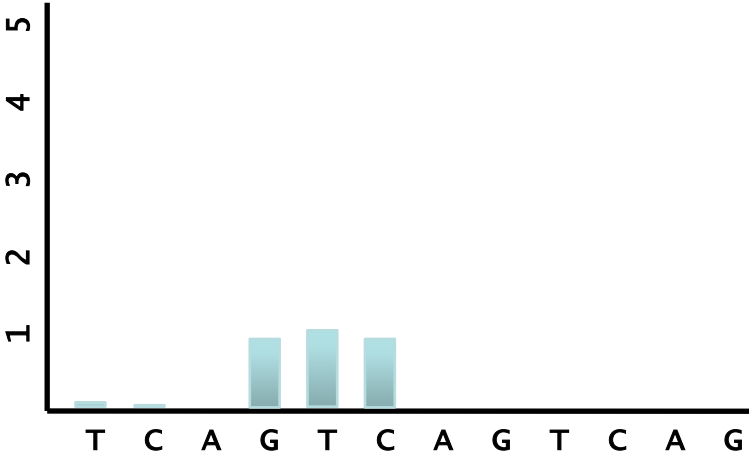


J. Craig Venter™ INSTITUTE

Only give polymerase one nucleotide at a time:

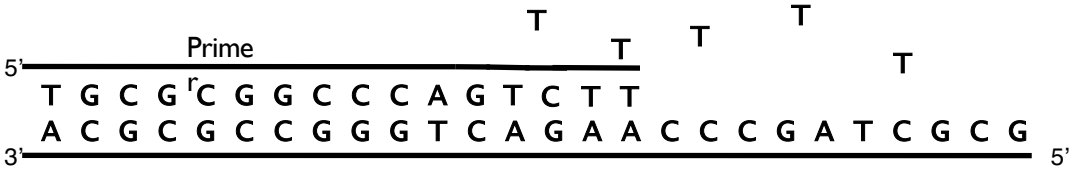


If that nucleotide is incorporated, enzymes turn PPi by-products into light:

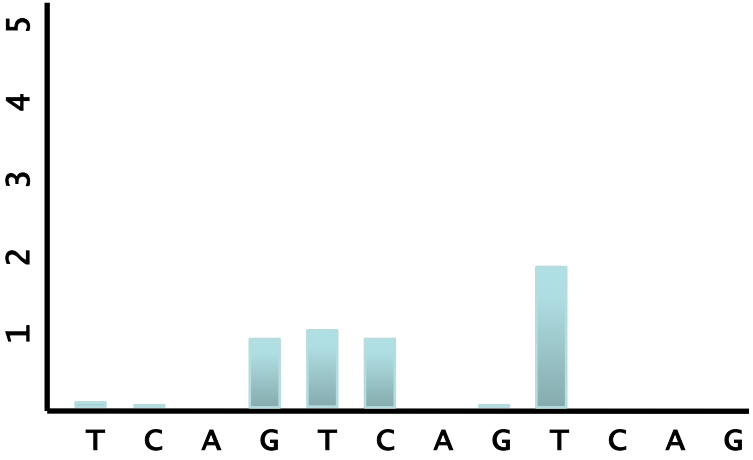


J. Craig Venter™ INSTITUTE

Only give polymerase one nucleotide at a time:

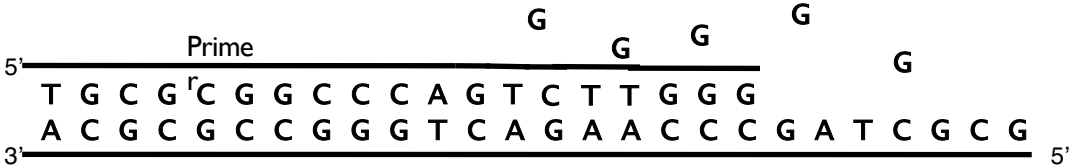


If that nucleotide is incorporated, enzymes turn PPi by-products into light:

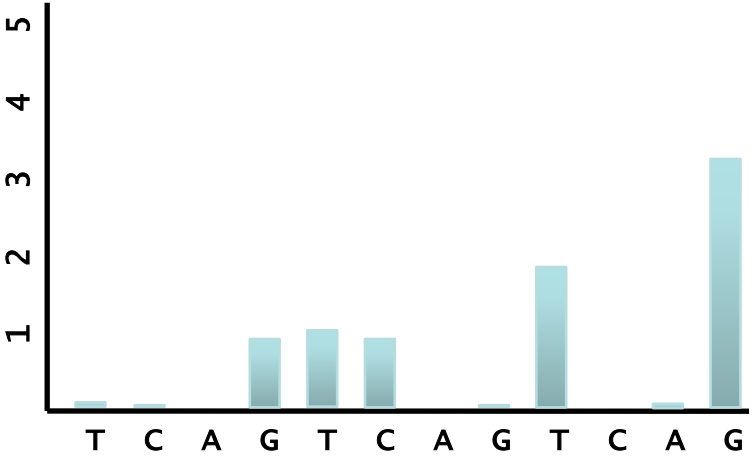


J. Craig Venter™ INSTITUTE

Only give polymerase one nucleotide at a time:



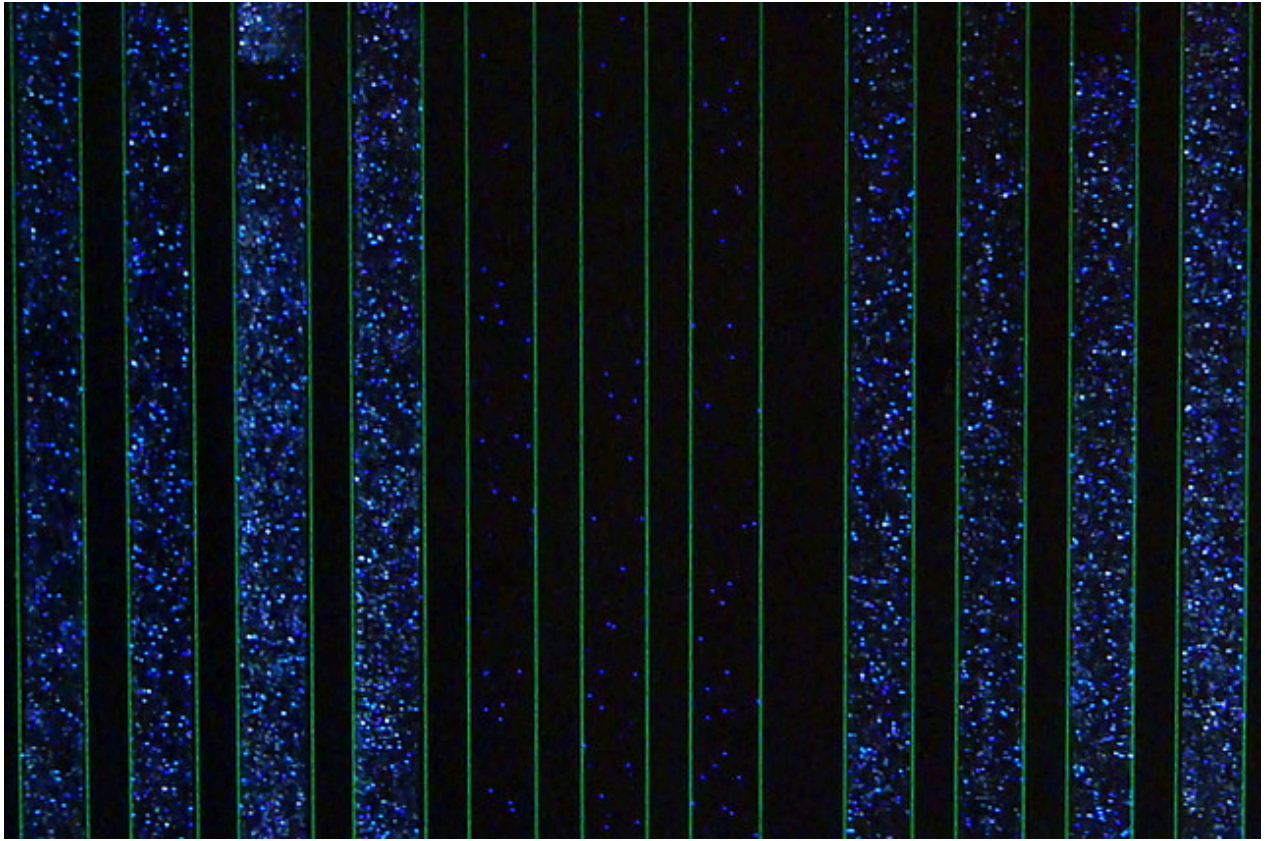
If that nucleotide is incorporated, enzymes turn PPi by-products into light:



The real power of this method is that it can take place in millions of tiny wells in a single plate at once.

J. Craig Venter™ INSTITUTE

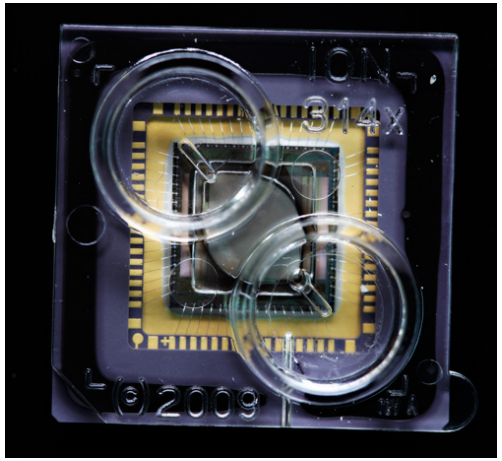
Raw 454 Data

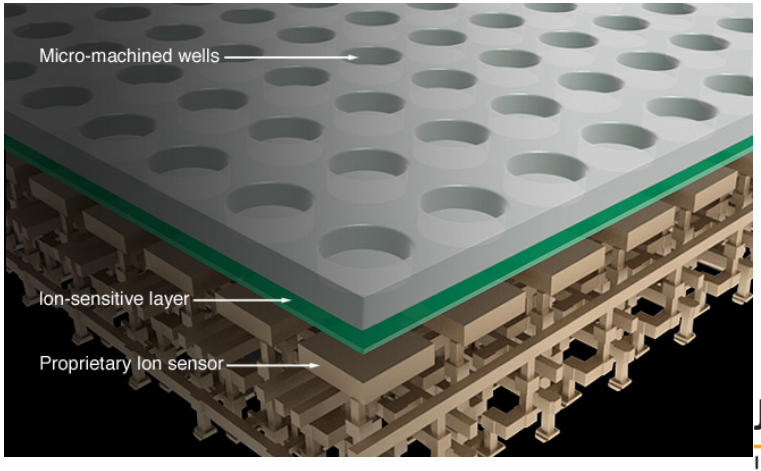


J. Craig Venter™
INSTITUTE

Ion Torrent/PGM/Life Technologies

- Ion semiconductor sequencing – detection of released hydrogen ions using ion-sensitive field-effect transistor technology
- Also uses emulsion PCR amplification step

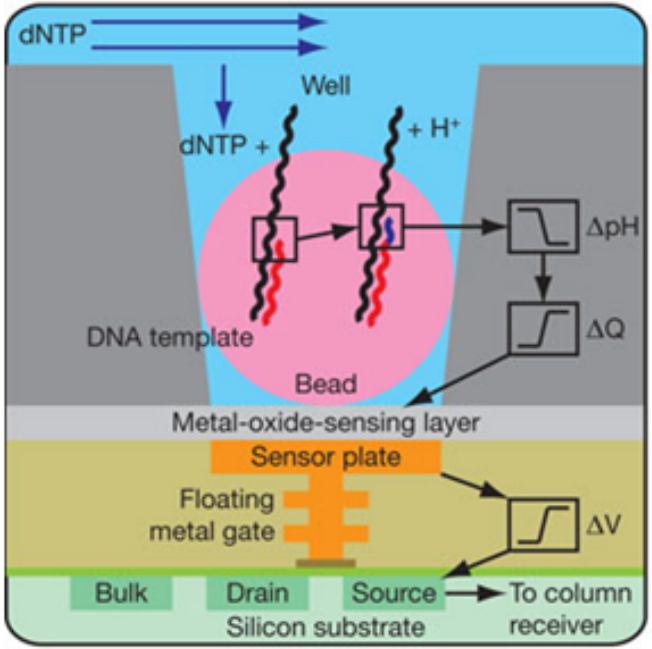


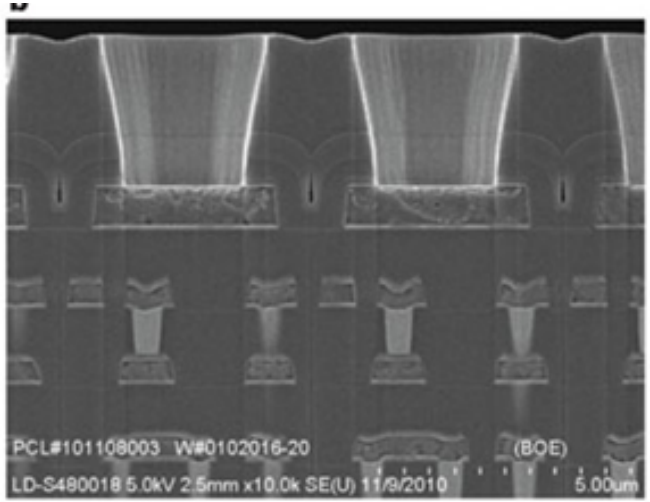


J. Craig Venter™ INSTITUTE

Detection of pH Change

"Nanowell" solid-state detection





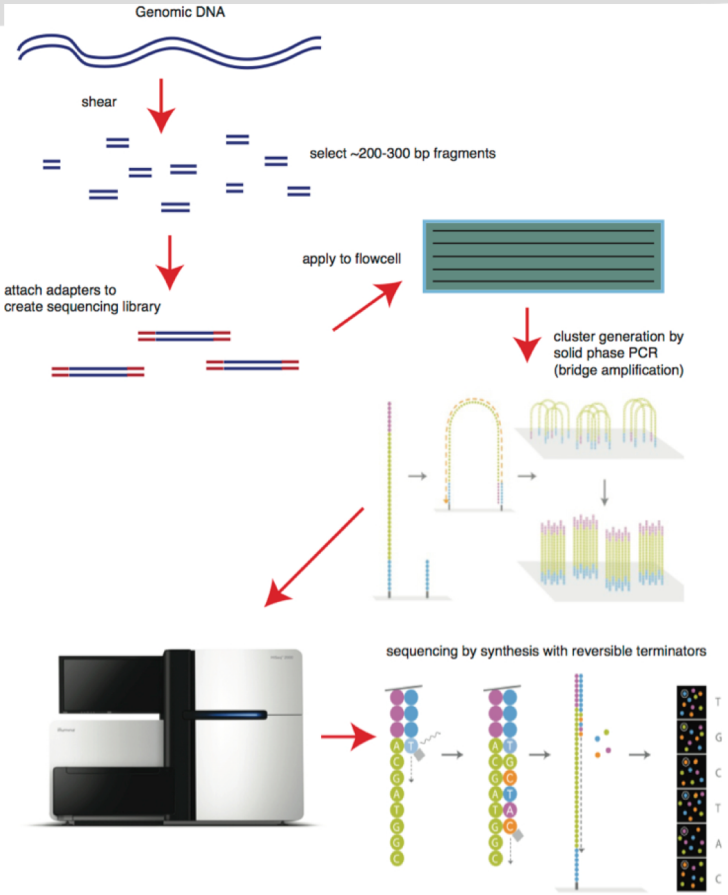
Rothberg et al. *Nature* (2011)

J. Craig Venter™ INSTITUTE

Illumina/Solexa

- Adapter ligation
- DNA cluster amplification
- Sequencing by synthesis with reversible dye termination

https://www.youtube.com/watch?v=fCd6B5HRaZ8



J. Craig Venter™ INSTITUTE

Illumina Benchtop Sequencers

	iSeq 100	MiniSeq	MiSeq Series	NextSeq 550 Series	NextSeq 1000 & 2000
Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application	Key Application
* Large Whole-Genome Sequencing (human, plant, animal)					
* Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●	●
Exome & Large Panel Sequencing (enrichment-based)				●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●	●	●
* Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)				●	●
* Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)				●	●
Targeted Gene Expression Profiling	●	●	●	●	●
miRNA & Small RNA Analysis	●	●	●	●	●
DNA-Protein Interaction Analysis (ChIP-Seq)			●	●	●
Methylation Sequencing				●	●
* 16S Metagenomic Sequencing		●	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)				●	●
Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11-48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	360 Gb*
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.2 billion*
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

J. Craig Venter™ INSTITUTE

Illumina Production-scale Sequencers



	NextSeq 550 Series	NextSeq 1000 & 2000	NovaSeq 6000
Popular Applications & Methods	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)			●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●
Exome & Large Panel Sequencing (enrichment-based)	●	●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	●	●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	●	●	●
Chromatin Analysis (ATAC-Seq, ChIP-Seq)	●	●	●
Methylation Sequencing	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)	●	●	●
Cell-Free Sequencing & Liquid Biopsy Analysis	●	●	●
Run Time	12–30 hours	11-48 hours	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	360 Gb*	6000 Gb
Maximum Reads Per Run	400 million	1.2 billion*	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 250**

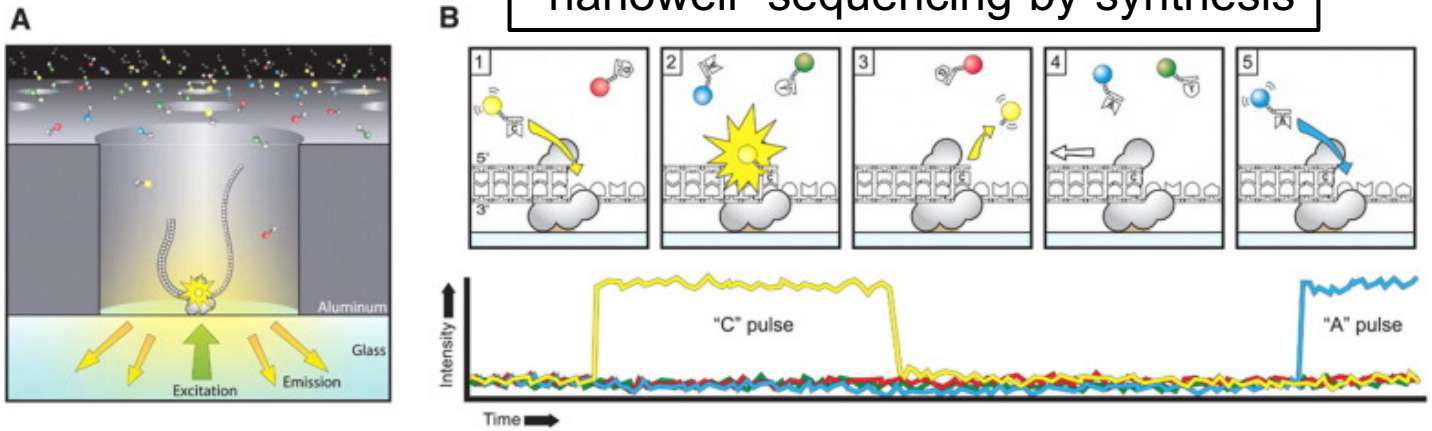
J. Craig Venter™ INSTITUTE

Third Generation

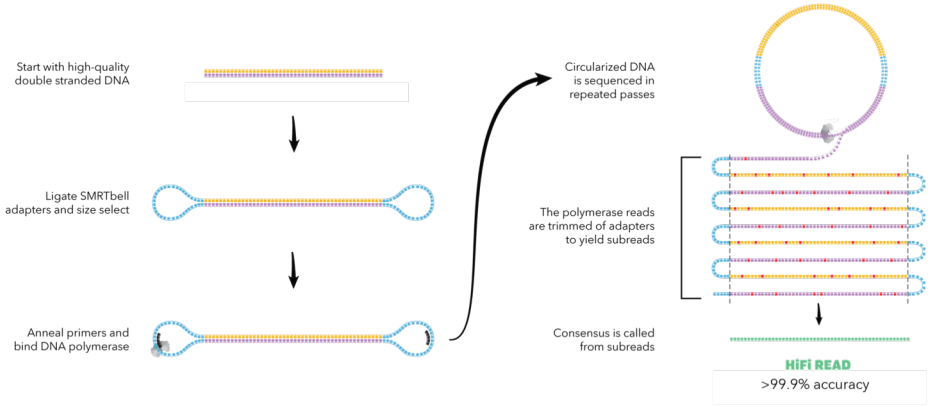
- Single molecule sequencing
- Helicos, PacBio, Oxford Nanopore

PacBio SMRT

“nanowell” sequencing-by-synthesis



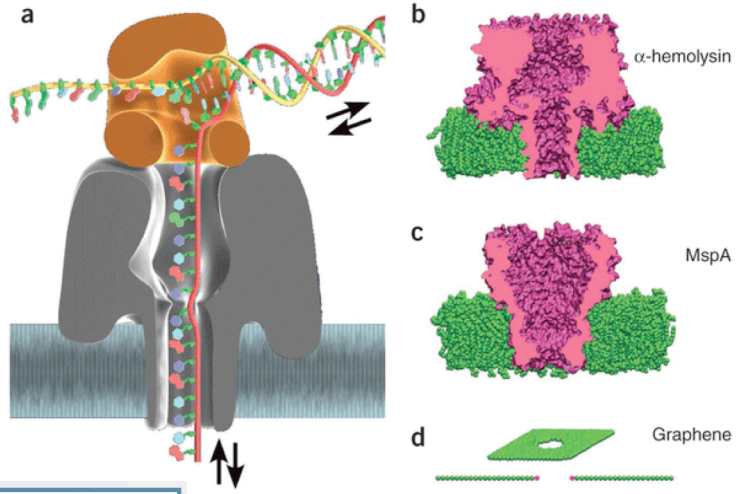
- Single molecule detection
- Sequencing by synthesis
- Single base incorporation

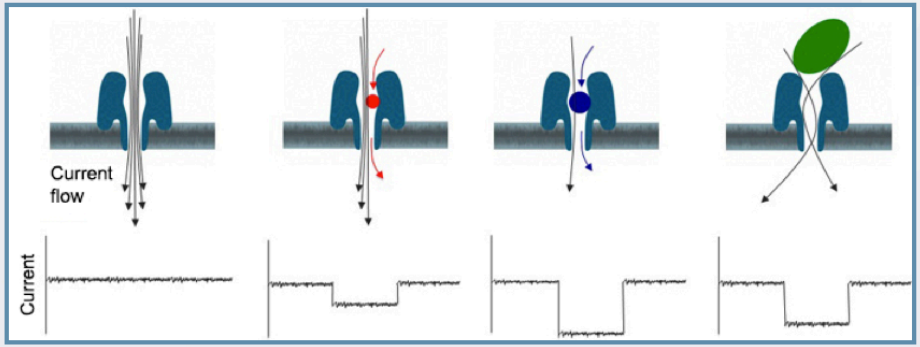


J. Craig Venter™ INSTITUTE

Oxford Nanopore sequencing

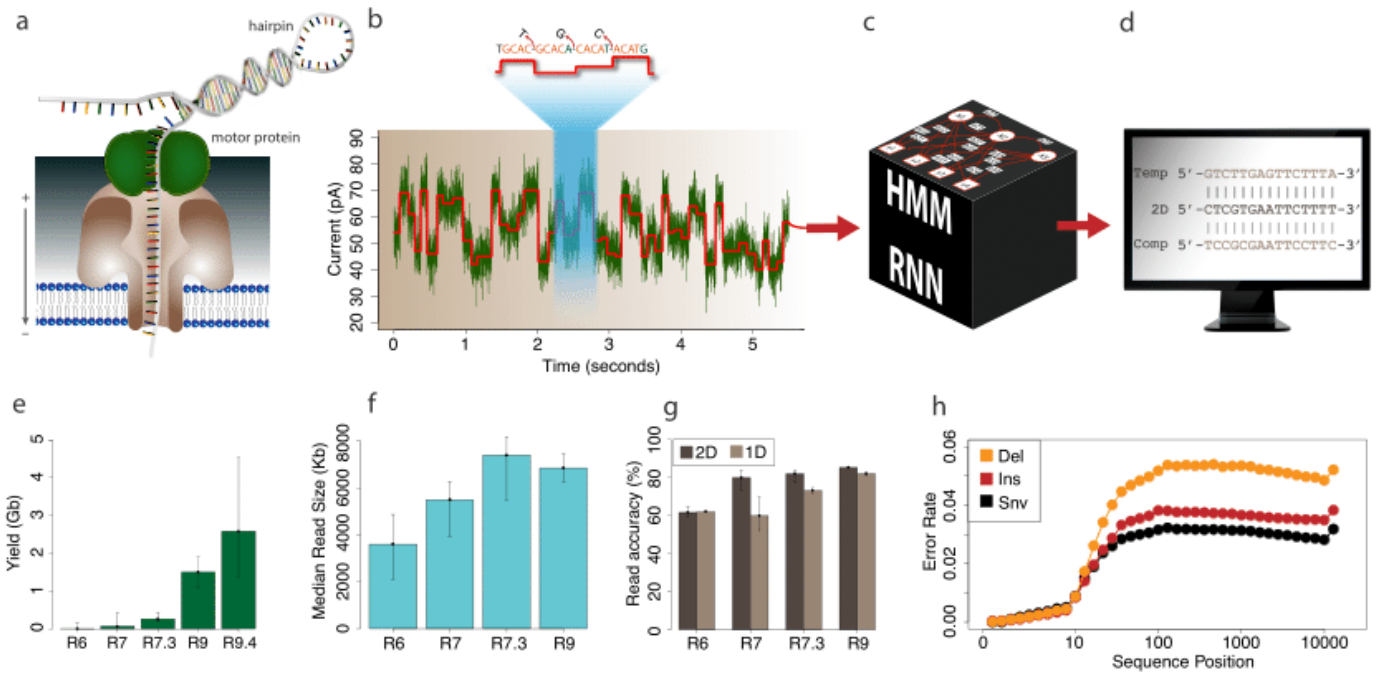
- DNA pushed through a nanopore in a lipid membrane
- Speed control provided by a Phi29 DNA polymerase
- Measure changes in the ionic current of an applied electric field





J. Craig Venter INSTITUTE

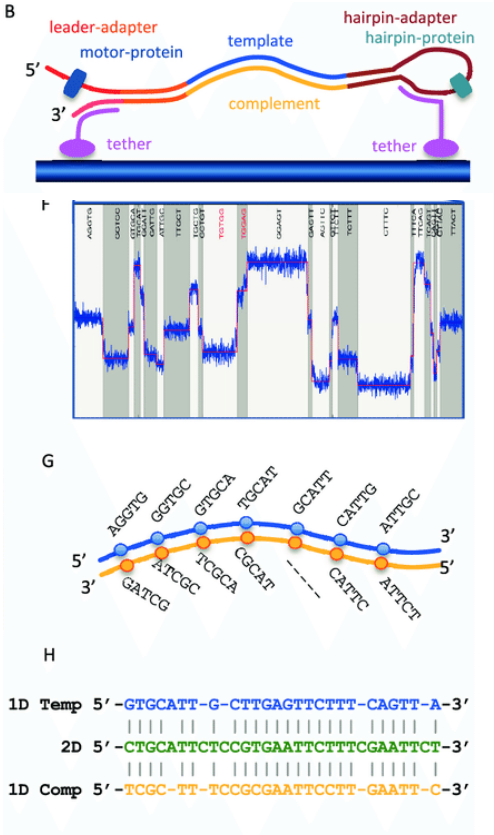
Base calling and accuracy

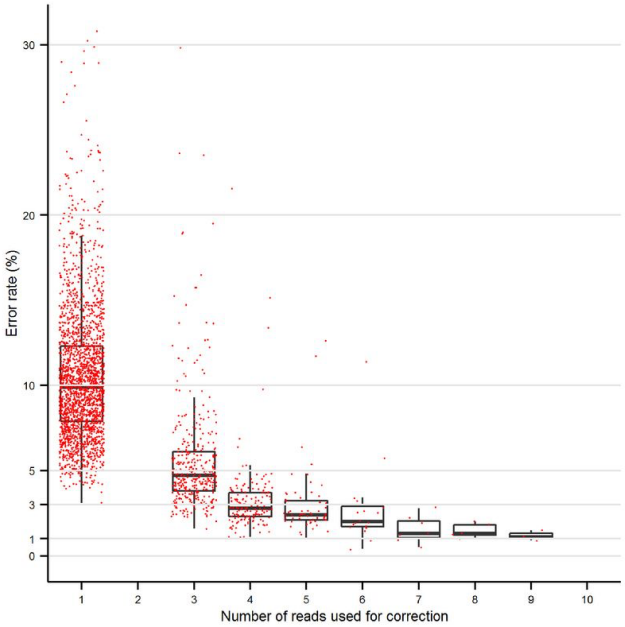


Magi et al. Briefings in Biotechnology, 2017, 1–17.

J. Craig Venter INSTITUTE

Accuracy improvements

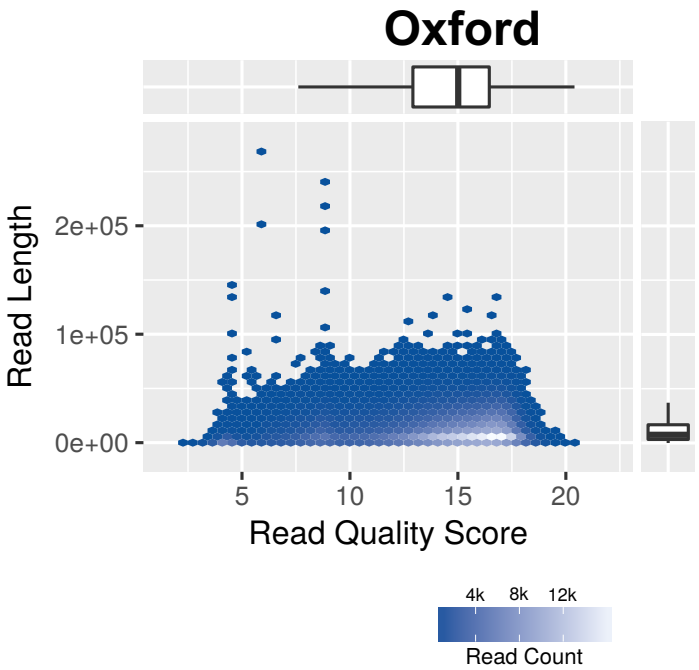


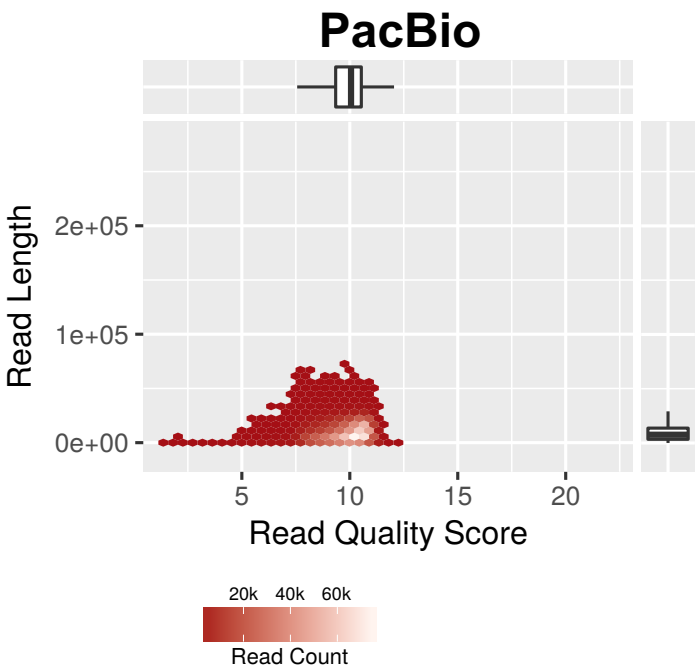


Karst SM, et al. Nature Biotechnology volume 36, pages 190–195 (2018)

J. Craig Venter™ INSTITUTE

Oxford produces high quality reads >50 kb; longest >800 kb



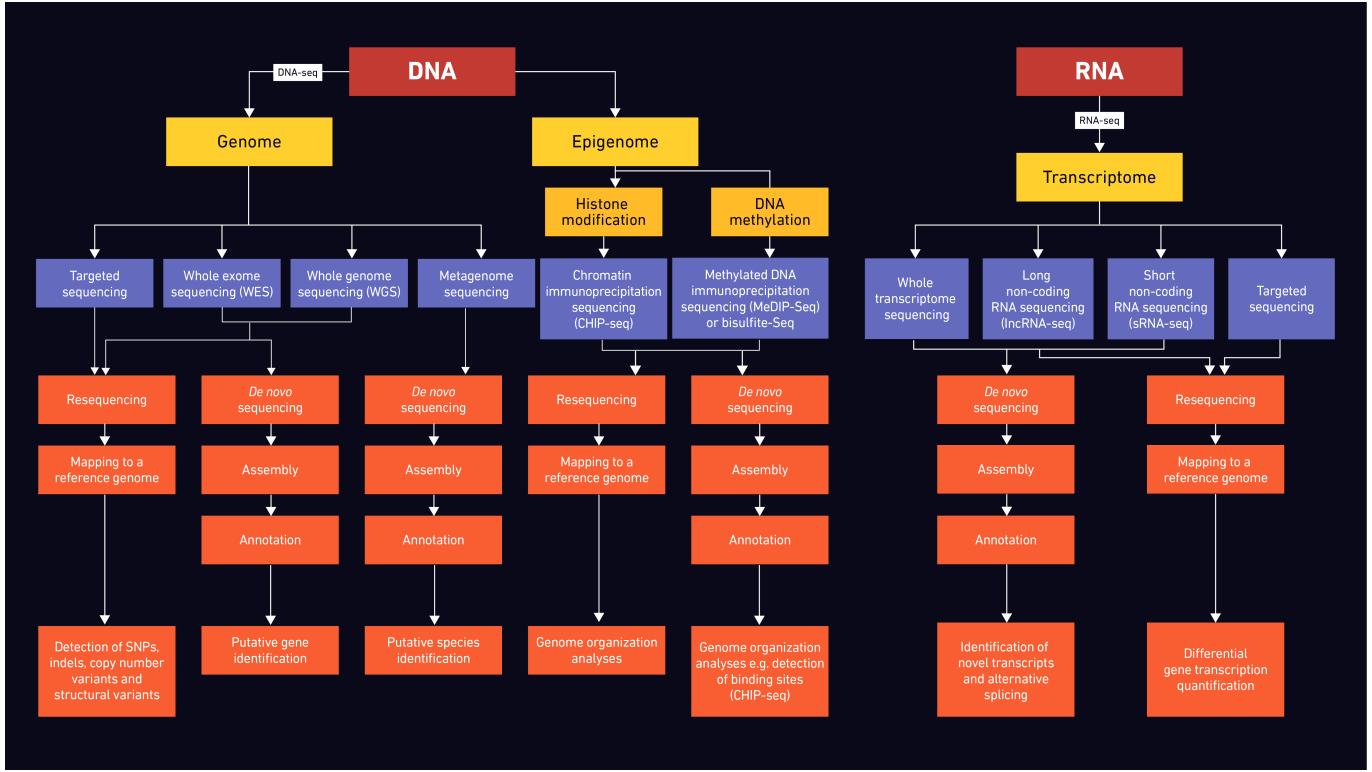


J. Craig Venter™ INSTITUTE

Long Read Technology Comparison

- Advantages
  - Full length transcriptomes, including splice variants
  - Resolution of long repeat regions in genomes
  - Genomic structural variants
  - Haplotype phasing
- Disadvantages
  - Higher error rates
  - Lower throughput

NGS applications



https://www.technologynetworks.com/genomics/articles/an-overview-of-next-generation-sequencing-346532

J. Craig Venter™ INSTITUTE

NGS Challenges

- Requires infrastructure to transfer, manage, store and process large amounts of data
- Some technologies are error prone and generate systematic biases and so data quality is an issue
- Assembling complete genomes from short sequence reads can be very difficult, especially in repeat regions and for shotgun sequences without reference genomes
- Difficult to establish phase of variants in diploid genomes
- Requires skills in bioinformatics that may not be readily accessible to many research labs

BIOINFORMATICS METHODS

Some More Trivia

- How many base pairs (bp) are there in a human genome?

Some More Trivia

- How many base pairs (bp) are there in a human genome?

  ~ 3 billion (haploid)

- When was the first human genome sequence completed?

  a) 1990 b) 1995 c) 2000 d) 2003

Some More Trivia

- How many base pairs (bp) are there in a human genome?

  ~ 3 billion (haploid)

- When was the first human genome sequence completed?

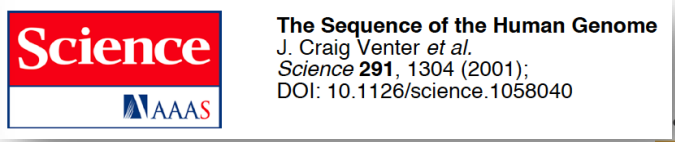
  a) 1990 b) 1995 c) 2000 d) 2003


articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

860 © 2001 Macmillan Magazines Ltd NATURE | VOL 409 | 15 FEBRUARY 2001 | www.nature.com


Science AAAS

The Sequence of the Human Genome
J. Craig Venter *et al.*
Science **291**, 1304 (2001);
DOI: 10.1126/science.1058040

J. Craig Venter™ INSTITUTE

Some More Trivia

- How many base pairs (bp) are there in a human genome?

  ~ 3 billion (haploid)

- When was the first human genome sequence completed?

  a) 1990 b) 1995 c) 2000 d) 2003

- How much did it cost to sequence the first human genome?

  a) $2 million b) $20 million c) $200 million d) $2 billion

- How long did it take to sequence the first human genome?

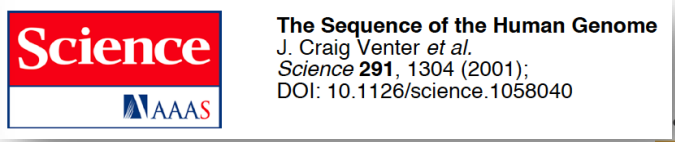
  a) 1 year b) 4 years c) 7 years d) 13 years


articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

860 © 2001 Macmillan Magazines Ltd NATURE | VOL 409 | 15 FEBRUARY 2001 | www.nature.com


Science AAAS

The Sequence of the Human Genome
J. Craig Venter *et al.*
Science **291**, 1304 (2001);
DOI: 10.1126/science.1058040

J. Craig Venter™ INSTITUTE

Some More Trivia

- How many base pairs (bp) are there in a human genome?

  ~ 3 billion (haploid)

- When was the first human genome sequence completed?

  a) 1990 b) 1995 c) 2000 d) 2003

- How much did it cost to sequence the first human genome?

  a) $2 million b) $20 million c) $200 million (private) d) $2.7 billion (public)

- How long did it take to sequence the first human genome?

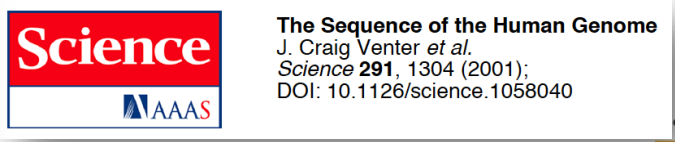
  a) 1 year b) 4 years (private) c) 7 years d) 13 years (public)


articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

860 © 2001 Macmillan Magazines Ltd NATURE | VOL 409 | 15 FEBRUARY 2001 | www.nature.com


Science AAAS

The Sequence of the Human Genome
J. Craig Venter *et al.*
Science **291**, 1304 (2001);
DOI: 10.1126/science.1058040

J. Craig Venter™ INSTITUTE

Some More Trivia

- How many base pairs (bp) are there in a human genome?

  ~ 3 billion (haploid)

- When was the first human genome sequence completed?

  a) 1990 b) 1995 c) 2000 d) 2003

- How much did it cost to sequence the first human genome?

  a) $2 million b) $20 million c) $200 million (private) d) $2.7 billion (public)

- How long did it take to sequence the first human genome?

  a) 1 year b) 4 years (private) c) 7 years d) 13 years (public)

- Whose genome was it?

J. Craig Venter™ INSTITUTE

Some More Trivia

- How many base pairs (bp) are there in a human genome?

  ~ 3 billion (haploid)

- When was the first human genome sequence completed?

  a) 1990 b) 1995 c) 2000 d) 2003

- How much did it cost to sequence the first human genome?

  a) $2 million b) $20 million c) $200 million (private) d) $2.7 billion (public)

- How long did it take to sequence the first human genome?

  a) 1 year b) 4 years (private) c) 7 years d) 13 years (public)

- Whose genome was it?

  several people (public); Craig Venter (private)

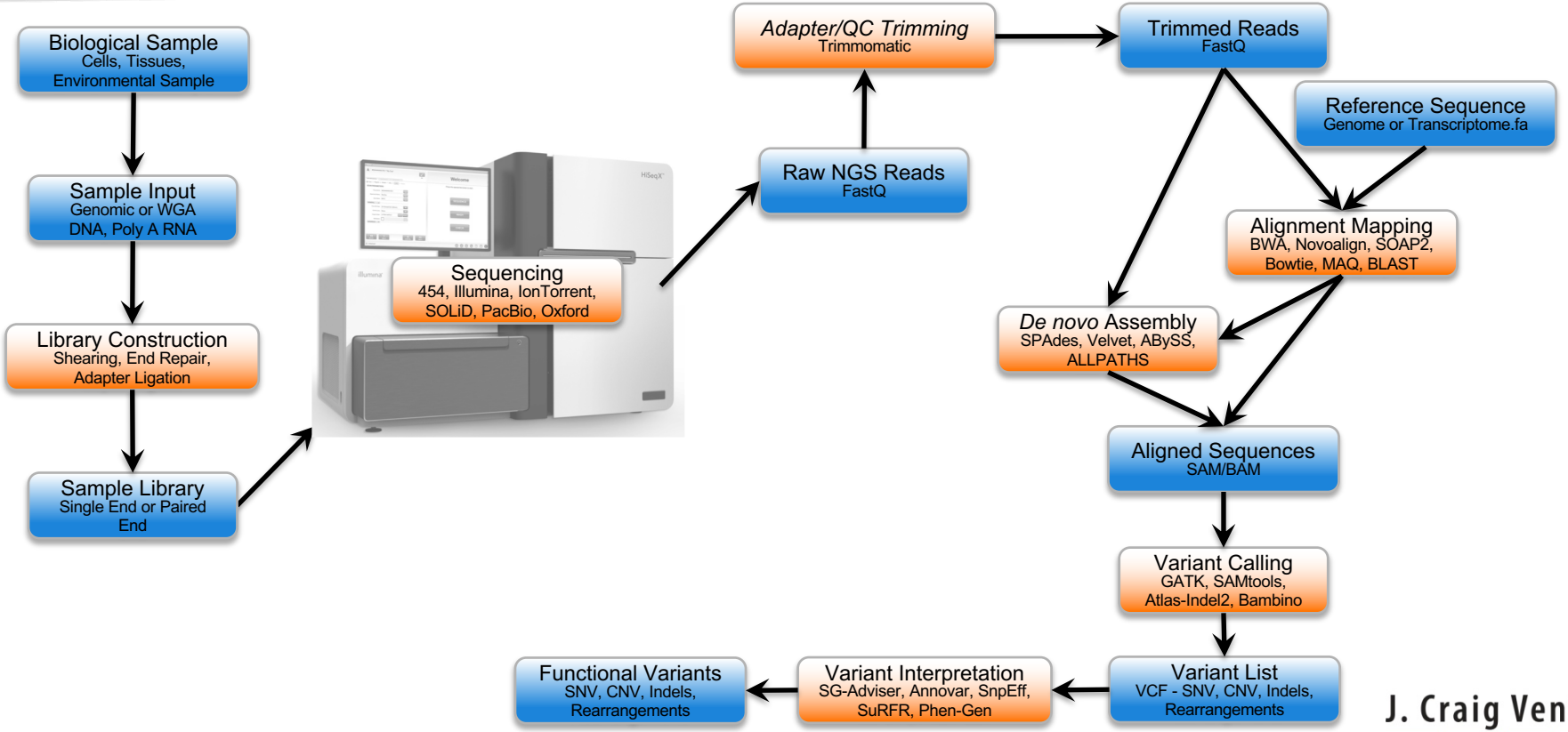
J. Craig Venter™ INSTITUTE

But was it?

- The Human Genome Project, completed in 2003, covered about 92% of the total human genome sequence. The technologies to decipher the gaps that remained didn't exist at the time.
- Since then, researchers have developed better laboratory tools, computational methods, and strategic approaches. The final, complete human genome sequence was described in a set of six papers in the April 1, 2022, issue of *Science*.
  - Telomere to Telomere (T2T) consortium led by researchers at NIH's National Human Genome Research Institute (NHGRI), the University of California, Santa Cruz, and the University of Washington, Seattle.
  - Long read sequencing technology used:
    - PacBio HiFi - 20,000 letters with nearly perfect accuracy
    - Oxford Nanopore - up to 1 million DNA letters at a time—with modest accuracy.
  - Added nearly 200 million nucleotides (8% of the genome) - newly added sequences were mainly in the centromeres and telomeres
  - https://www.nih.gov/news-events/nih-research-matters/first-complete-sequence-human-genome

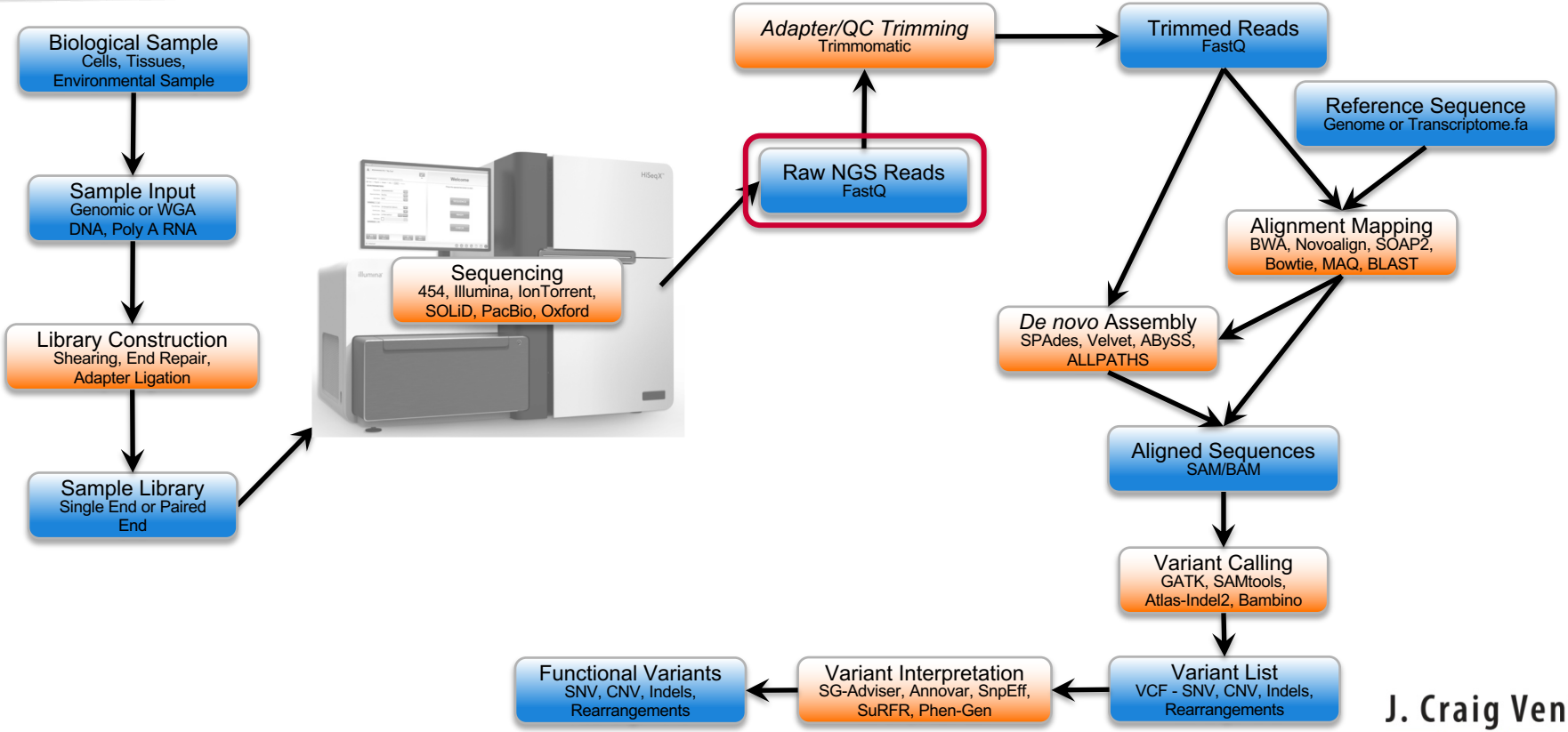
BIOINFORMATICS METHODS

NGS Processing Workflows



J. Craig Venter™ INSTITUTE

NGS Processing Workflows



J. Craig Venter™ INSTITUTE

FASTQ File Format

- The FASTQ format allows the storage of both sequence and quality information for each read.
- This is a compact text-based format that has become the *de facto standard for* storing data from next generation sequencing experiments.

FASTA File Format

```
>gb:ON369947|Organism:Influenza A virus|Strain Name:A/Alabama/09/2022|Segment:8|Subtype:H3N2|Host:Human
GTGACAAAGACATAATGGATTCCAACACTGTGTCAAGTTTCCAGGTAGATTGCTTCCTTTGGCATATCCG
GAAACAAGTTGTGGACCAAAAACTGAGTGATGCCCCATTCCTCGATCGGCTTCGCCGAGATCAGAGGTCC
CTAGGGGGAAGAGGCAATACTCTCGGTCTAGACATCAAATCAGCCACCCATGTTGGAAAGCAAATCGTAG
AAAAGATTCTGAAAGGAGAATCTGATGAGGCACTTAAAATGACCATGGTCTCAACACCTGCTTCGCGATA
CATAACTGACATGACTATTGAGGAATTGTCAAGAAACTGGTTCATGCTAATGCCCAAGCAGAAGGTGGAA
GGACCTCTTTGCATCAGAATGGACCAGGCAATCATGGAGAAAAACATCATGTTAAAAGCGAATTTCAATG
TGATTTTTGGCCGGCTAGAGACCATAGTATTGCTAAGGGCTTTCACCGAAGAGGGAGCAATTGTTGGCGA
AATCTCACCATTGCCTTCTTTTCCAGGACATACTATTGAGGATGTCAAAAATGCAATTGGGGTCCTCATC
GGAGGACTTGAATGGAATGATAACACAGTTCGAGTCTCTAAAAATCTACAGAGATTCGCTTGGAGAAGCA
GTCATGAGAGTGGGGGACCTCCACTTACTCCAAAACAGAAACGAGAAATGGCGAGAACAGCTAGGTCAGA
AGTTTGAAGAGATAAGATGGCTGATTGAAGAGGTGAGACACAGATTAAGAACAACTGAAAATAGCTTTGA
ACAAATAACATTCATGCAAGCATTACAACTACTATTTGAAGTGGAACAGGAGATAAGAACTTTCTCATTT
CAGCTTATTTGATGATAAAAAACAC

>gb:ON369979|Organism:Influenza A virus|Strain Name:A/Alaska/04/2022|Segment:8|Subtype:H3N2|Host:Human
GTGACAAAGACATAATGGATTCCAACACTGTGTCAAGTTTCCAGGTAGATTGCTTCCTTTGGCATATCCG
GAAACAAGTTGTGGACCAAAAACTGAGTGATGCCCCATTCCTCGATCGGCTTCGCCGAGATCAGAGGTCC
CTAGGGGGAAGAGGCAATACTCTCGGTCTAGACATCAAATCAGCCACCCATGTTGGAAAGCAAATCGTAG
AAAAGATTCTGAAAGGAGAATCTGATGAGGCACTTAAAATGACCATGGTCTCAACACCTGCTTCGCGATA
CATAACTGACATGACTATTGAGGAATTGTCAAGAAACTGGTTCATGCTAATGCCCAAGCAGAAGGTGGAA
GGACCTCTTTGCATCAGAATGGATCAGGCAATCATGGAGAAAAACATCATGTTAAAAGCGAATTTCAATG
TGATTTTTGGCCGGCTAGAGGCCATAGTATTGCTAAGGGCTTTCACCGAAGAGGGAGCAATTGTTGGCGA
AATCTCACCATTGCCTTCTTTTCCAGGACATACTATTGAGGATGTCAAAAATGCAATTGGGGTCCTCATC
GGAGGACTTGAATGGAATGATAACACAGTTCGAGTCTCTAAAAATCTACAGAGATTCGCTTGGAGAAGCA
GTCATGAGAATGGGGGACCTCCACTTACTCCAAAACAGAAACGAGAAATGGCGAGAACAGCTAGGTCAGA
AGTTTGAAGAGATAAGATGGCTGATTGAAGAGGTGAGACACAGATTAAGAACAACTGAAAATAGCTTTGA
ACAAATAACATTCATGCAAGCATTACAACTACTGTTTGAAGTGGAACAGGAGATAAGAACTTTCTCATTT
CAGCTTATTTAATGATAAAAAACAC

>gb:ON370011|Organism:Influenza A virus|Strain Name:A/Arizona/08/2022|Segment:8|Subtype:H3N2|Host:Human
GTGACAAAGACATAATGGATTCCAACACTGTGTCAAGTTTCCAGGTAGATTGCTTCCTTTGGCATATCCG
GAAACAAGTTGTGGACCAAAAACTGAGTGATGCCCCATTCCTCGATCGGCTTCGCCGAGATCAGAGGTCC
CTAGGGGGAAGAGGCAATACTCTCGGTCTAGACATCAAATCAGCCACCCATGTTGGAAAGCAAATCGTAG
AAAAGATTCTGAAAGGAGAATCTGATGAGGCACTTAAAATGACCATGGTCTCAACACCTGCTTCGCGATA
CATAACTGACATGACTATTGAGGAATTGTCAAGAAACTGGTTCATGCTAATGCCCAAGCAGAAGGTGGAA
GGACCTCTTTGCATCAGAATGGACCAGGCAATCATGGAGAAAAACATCATGTTAAAAGCGAATTTCAATG
TGATTTTTGGCCGGCTAGAGACCATAGTATTGCTAAGGGCTTTCACCGAAGAGGGAGCAATTGTTGGCGA
AATCTCACCATTGCCTTCTTTTCCAGGACATACTATTGAGGATGTCAAAAATGCAATTGGGGTCCTCATC
GGAGGACTTGAATGGAATGATAACACAGTTCGAGTCTCTAAAAATCTACAGAGATTCGCTTGGAGAAGCA
GTCATGAGAATGGGGGACCTCCACTTACTCCAAAACAGAAACGAGAAATGGCGAGAACAGCTAGGTCAGA
AGTTTGAAGAGATAAGATGGCTGATTGAAGAGGTGAGACACAGATTAAGAACAACTGAAAATAGCTTTGA
ACAAATAACATTCATGCAAGCATTACAACTACTGTTTGAAGTGGAACAGGAGATAAGAACTTTCTCATTT
CAGCTTATTTAATGATAAAAAACAC
```

J. Craig Venter™ INSTITUTE

FastQ Format

```
@HWUSI-EAS582_157:6:1:1:1501/1
NCACAGACACACACGAACACACAAAGACATGCCCATATGAAGAT
+
%.7786867:7785568587465750588 73/347777476035
@HWUSI-EAS582_157:6:1:1:1606/1
NCTGGCACCTTGATTTTGGACTTCCCAGCCTCCAGAACTGTGAG
+
%19489888887989883668988886489987888988888588
@HWUSI-EAS582_157:6:1:1:453/1
NCTGCTTGCACCCCTGAAGTCACTGATCACATTTCAGGGTCACC
+
%/868998988888867668888986644788988413488885
@HWUSI-EAS582_157:6:1:1:1844/1
NGATTGACATTGGCAAAGAGGACAACTGATTGCAAACTTCACAC
+
%-7;::::;86499;75574586::635:62687666887879
```

"Read" (sequence)

Quality scores (phred-33)

Header

: = phred of 25

; = phred of 26

Illumina sequence identifiers [edit]

Sequences from the Illumina software use a systematic identifier:

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

. Craig Venter™ . NSTITUTE

Assessing Quality: Phred scores

- Phred quality scores were originally produced by the Phred base calling program using a statistical analysis of Sanger chromatogram trace files in support of the Human Genome Project. Subsequently adapted to NGS technologies for judging qualities of sequences.

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

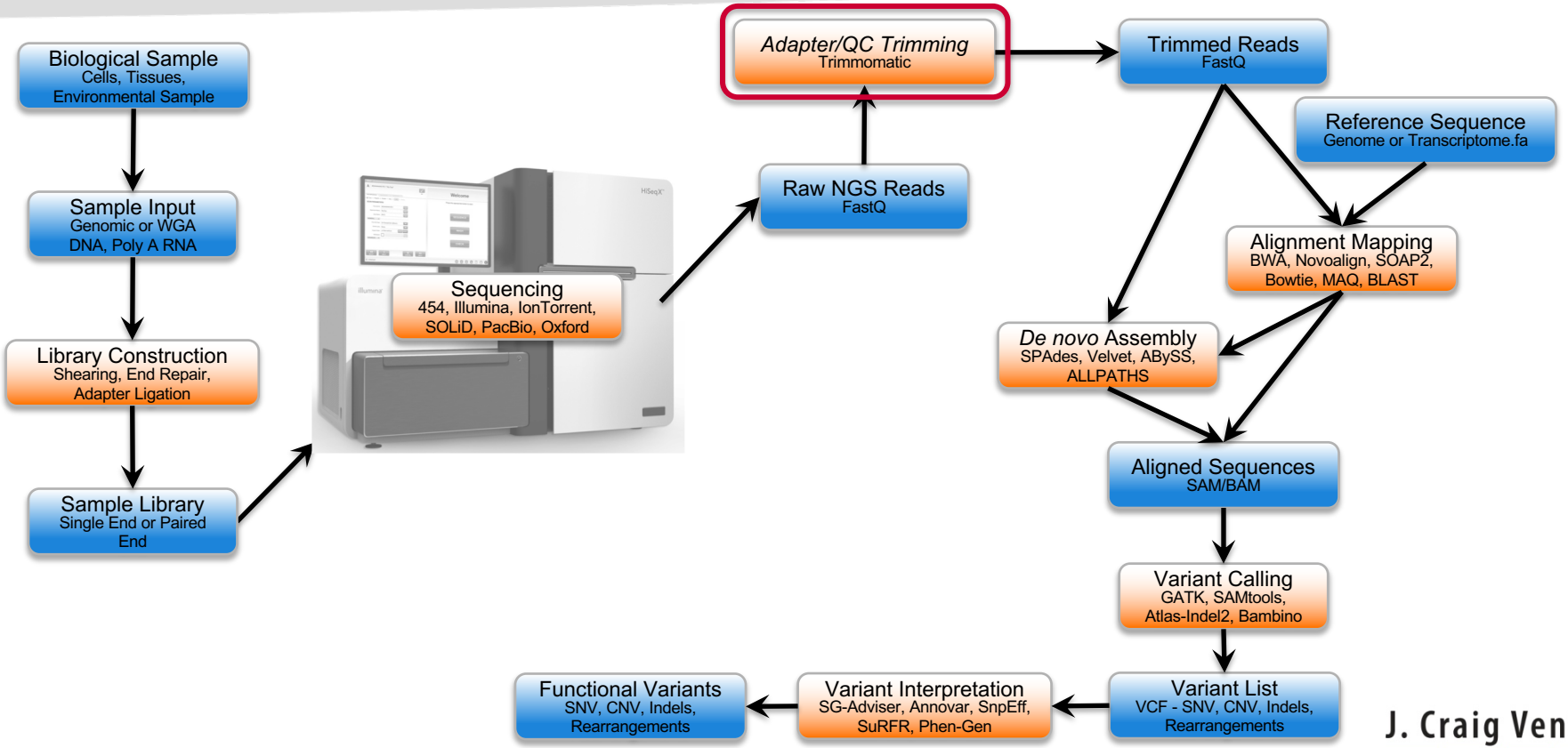
P_e = error probability of a given base call

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30 ←	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

J. Craig Venter™ INSTITUTE

Trimming



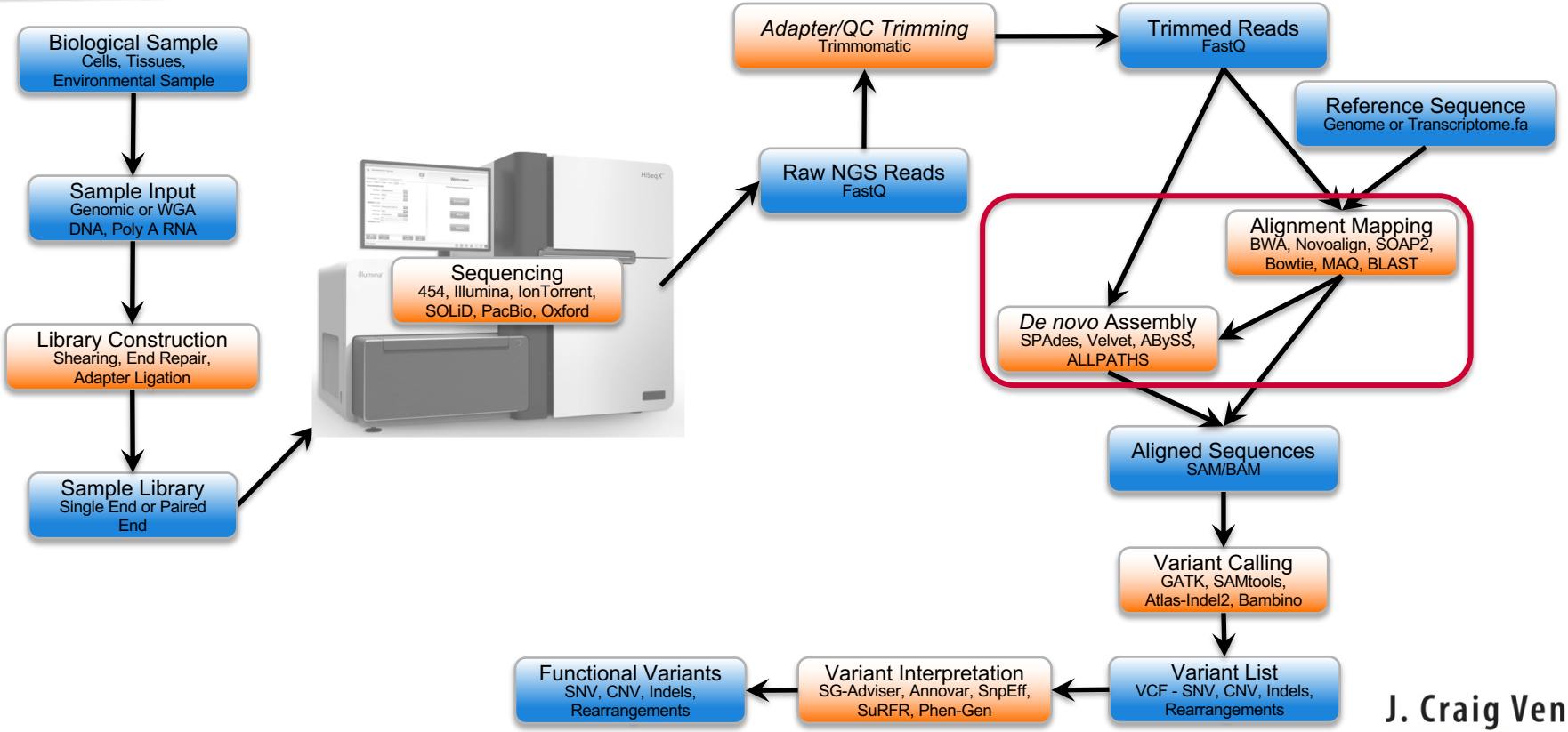
J. Craig Venter™ INSTITUTE

Trimming objectives and methods

- Objectives
  - Remove primers/adapters
  - Remove low quality positions and reads
- Methods
  - Trimmomatic
    - https://github.com/timflutre/trimmomatic
  - Cutadapt
    - https://github.com/marcelm/cutadapt
  - TrimGalore (Cutadapt + FastQC)
    - https://github.com/FelixKrueger/TrimGalore
  - Fastp
    - https://github.com/OpenGene/fastp

J. Craig Venter™ INSTITUTE

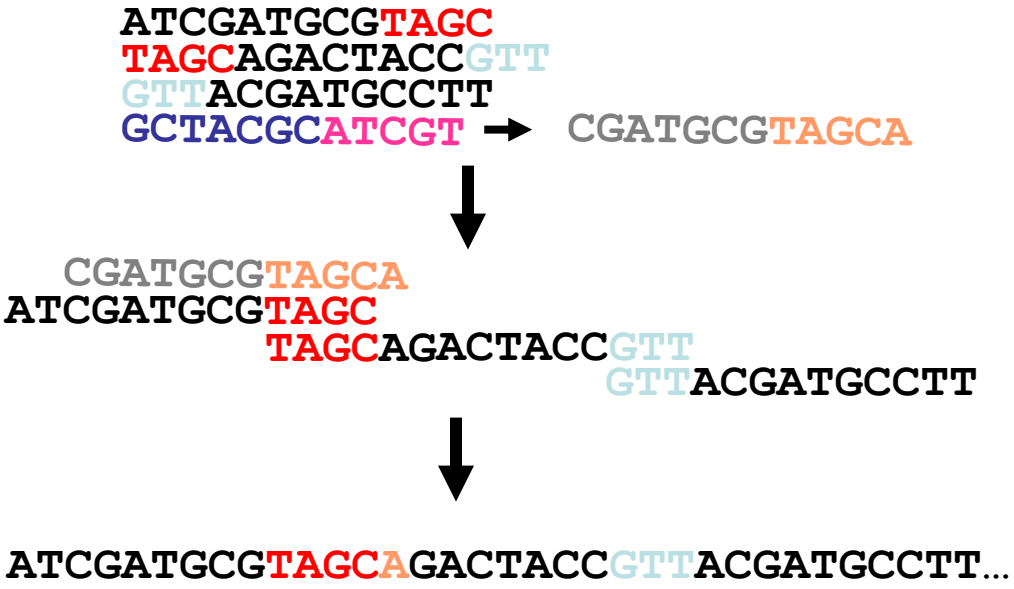
Alignment/Assembly



J. Craig Venter™ INSTITUTE

What is a sequence assembly?

- "An assembly is a hierarchical data structure that maps the sequence data to a putative reconstruction of the target"
- Miller JR, Koren S and Sutton G. 2010. **Assembly algorithms for next-generation sequencing data.** *Genomics* 95:315-32



J. Craig Venter INSTITUTE

Two Classes of Assembly

- Alignment-based mapping and assembly (aka reference-based alignment) refers to reconstruction of the underlying sequence facilitated by alignments to a previously resolved reference sequence.
- *De novo* assembly refers to reconstruction of the underlying sequence without a previously resolved reference sequence.

Alignment Mapping and Assembly of Short Reads

- ***Strategy:*** When a suitable reference sequence is available, index the reference genome sequence and search it efficiently
- For this purpose, map-alignment sequence assembly approaches generally use a computing strategy called Burrows–Wheeler transformation and indexing to notably reduce compute time and memory usage

  - **BWA – Burrows-Wheeler Aligner**
    http://bio-bwa.sourceforge.net

  - **Bowtie - An ultrafast memory-efficient short read aligner**
    Ben Langmead and Cole Trapnell, University of Maryland
    http://bowtie-bio.sourceforge.net/

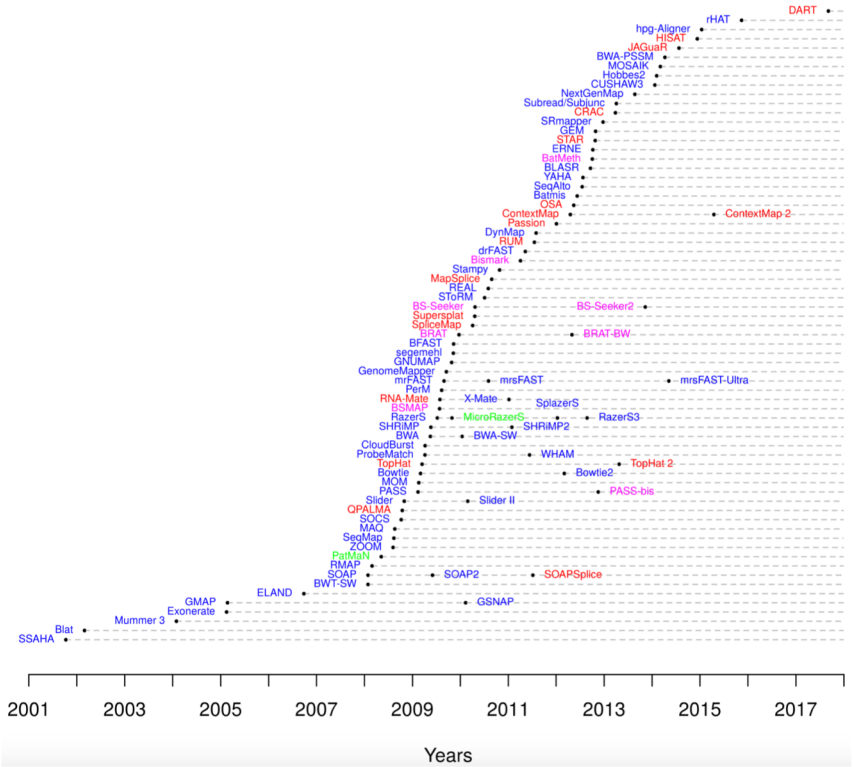
  - **MAQ – Mapping and Assembly with Quality**
    Heng Li, Sanger Centre http://maq.sourceforge.net/maq-man.shtml

  - **SOAPaligner/soap2 (Short Oligonucleotide Analysis Package)** http://soap.genomics.org.cn/soapaligner.html

  - Also see https://en.wikibooks.org/wiki/Next_Generation_Sequencing_(NGS)/Alignment

J. Craig Venter™ INSTITUTE

Current List of Alignment Mappers



DNA mappers are plotted in blue, RNA mappers in red, miRNA mappers in green, and bisulfite mappers in purple.

https://www.ebi.ac.uk/~nf/hts_mappers/

J. Craig Venter™ INSTITUTE

Considerations

- The short reads do not come with position information, that is, we do not know what part of the genome they came from; we need to use the sequence of the read itself to find the corresponding region in the reference sequence.
- The reference sequence can be quite long (~3 billion bases for human), making it a daunting task to find a matching region.
- Since our reads are short, there may be several, equally likely places in the reference sequence from which they could have been read. This is especially true for repetitive regions.
- If we were only looking for perfect matches to the reference, we would never see any variation. Therefore, we need to allow some mismatches and small structural variation (InDels) in our reads.
- Any sequencing technology produces errors. Similar to the "real" variations, we need to tolerate a low level of sequencing errors in our reads and separate them from the "real" variations later.
- We need to do that for each of the millions of reads in our sequencing data

Indexing and searching reference

- Short reads with long reference
- Preprocess reference
- Book – index => topics (in alphabetical order) and locations
- Instead of words, we'll use k-mer substrings
- Offset for position
- Query => index hits
- Extend by verification => match

```
Reference:
TACCTTCCCAGGTA

K-mers (k=5):
TACCT 1
 ACCTT 2
  CCTTC 3
   CTTCC 4
    TTCCC 5
     TCCCA 6
      CCCAG 7
       CCAGG 8
        CAGGT 9
         AGGTA 10
```

```
K-mer index:
ACCTT 2
AGGTA 10
CAGGT 9
CCAGG 8
CCCAG 7
CCTTC 3
CTTCC 4
TACCT 1
TCCCA 6
TTCCC 5

Read: CCAGGTA

Match: 8
```

J. Craig Venter INSTITUTE

De novo Assembly of Short Reads

- ***Strategy:*** When a suitable reference sequence is not available, assemble reads *de novo*
- For this purpose, assembly approaches generally use a computing strategy called de Bruijn graph of k-mers to notably reduce compute time and memory usage, but de Bruijn graphs are inherently very large due to the observed number of distinct k-mers in the target sequences and hence require significant computer memory to hold the constructed graph
  - **SPAdes**
    - *Max A. Alekseyev and Pavel Pezner*
    - https://github.com/ablab/spades
  - **Velvet**
    - *Daniel Zerbino and Ewan Birney, EMBL-EBI*
    - http://www.ebi.ac.uk/~zerbino/velvet/
  - **ABySS**
    - *Inanç Birol, Shaun Jackman, Steve Jones and others, GSC*
    - http://www.bcgsc.ca/platform/bioinfo/software/abyss
  - **ALLPATHS-LG**
    - *Jaffe et al CRD, Broad Institute*
    - http://www.broadinstitute.org/software/allpaths-lg/blog/
  - **SOAPdenovo**
    - *Li et al. Beijing Genome Institute*
    - http://soap.genomics.org.cn/soapdenovo.html
- Additional software listed in the Earl DA et al. 2011. **Assemblathon 1: A competitive assessment of de novo short read assembly methods.** http://genome.cshlp.org/content/early/2011/09/16/gr.126599.111

J. Craig Venter™ INSTITUTE

Current Methods

Name	Description / Methodology	Technologies	Author	Presented / Last updated	Licence*	Homepage
ABySS	parallel, paired-end sequence assembler designed for large genome assembly of short reads (genomic and transcriptomic), employ a Bloom filter to De Bruijn graph	Illumina	[8][9]	2009 / 2017	OS	link
DISCOVAR	paired-end PCR-free reads (successor of ALLPATHS-LG)	Illumina (MiSeq or HiSeq 2500)	[10]	2014	OS	link
DNA Baser Sequence Assembler	DNA sequence assembly with automatic end trimming & ambiguity correction. Includes a base caller.	Sanger, Illumina	Heracle BioSoft SRL	2018.09	C ($69)	NA
DNASTAR Lasergene Genomics	(large) genomes, exomes, transcriptomes, metagenomes, ESTs	Illumina, ABI SOLiD, Roche 454, Ion Torrent, Solexa, Sanger	DNASTAR	2007 / 2016	C	link
Newbler	genomes, ESTs	454, Sanger	454 Life Sciences	2004/2012	C	link
Phrap	genomes	Sanger, 454, Solexa	Green, P.	1994 / 2008	C / NC-A	link
Plass	Protein-level assembler: assembles six-frame-translated sequencing reads into protein sequences	Illumina	[11]	2018 / 2019	OS	link
Ray	a suite of assemblers including de novo, metagenomic, ontology and taxonomic profiling; uses a De Bruijn graph		[12]	2010	OS	link
SPAdes	(small) genomes, single-cell	Illumina, Solexa, Sanger, 454, Ion Torrent, PacBio, Oxford Nanopore	[13]	2012 / 2021	OS	link
Velvet	(small) genomes	Sanger, 454, Solexa, SOLiD	[14]	2007 / 2011	OS	link
HGAP	Genomes up to 130 MB	PacBio reads	[15]	2011 / 2015	OS	link
Falcon	Diploid genomes	PacBio reads	[16]	2014 / 2017	OS	link
Canu	Small and large, haploid/diploid genomes	PacBio/Oxford Nanopore reads	[17]	2001 / 2018	OS	link
MaSuRCA	Any size, haploid/diploid genomes	Illumina and PacBio/Oxford Nanopore data, legacy 454 and Sanger data	[18]	2011 / 2018	OS	link
Hinge	Small microbial genomes	PacBio/Oxford Nanopore reads	[19]	2016 / 2018	OS	link
Trinity	transcriptome assemblies by de Bruijn graph	Illumina RNA-seq	[20]	2011		link

***Licences:** OS = Open Source; C = Commercial; C / NC-A = Commercial but free for non-commercial and academics

https://en.wikipedia.org/wiki/Sequence_assembly
https://en.wikipedia.org/wiki/De_novo_sequence_assemblers

J. Craig Venter™ INSTITUTE

Hybrid Approach

- Align reads to reference if you can
- *De novo* assemble remaining reads for identification of novel regions/genomes

SAM/BAM File Format

- Sequence Alignment Map (SAM) format
- BAM compressed binary version

Example Header Lines

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/data/local/ref/GATK/human_glk_v37.fasta        M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/data/local/ref/GATK/human_glk_v37.fasta        M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/data/local/ref/GATK/human_glk_v37.fasta        M5:fdfd811849cc2fadebc929bb925902e5
@RG     ID:UM0098:1     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L001 LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@RG     ID:UM0098:2     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L002 LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@PG     ID:bwa  VN:0.5.4
@PG     ID:GATK TableRecalibration      VN:1.0.3471     CL:Covariates=[ReadGroupCovariate, QualityScoreCovariate, CycleCovariate, DinucCovariate, TileCovariate], default_read_group=null, default
```

In the alignment examples below, you will see that the 2nd alignment maps back to the RG line with ID UM0098.1, and all of the alignments point back to the SQ line with SN:1 because their RNAME is 1.

Example Alignments

This is what the alignment section of a SAM file looks like:

```
1:497:R:-272+13M17D24M  113     1       497     37      37M     15      100338662       0       CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG   0;==-==9;>>>>>>=>>>>>>>>>>>>=>>>>>>>>>>   XT:A:U  NM:i:0  SM
19:20389:F:275+18M2D19M 99      1       17644   0       37M     =       17919   314     TATGACTGCTAATAATACCTACACATGTTAGAACCAT   >>>>>>>>>>>>>>>>>>>>>>><>>>><<>>>4:;>>:<9   RG:Z:UM0098:1    XT:A:R  NM
19:20389:F:275+18M2D19M 147     1       17919   0       18M2D19M        =       17644   -314    GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT   ;44999;499<8<8<<8<<><<<<><7<;<<<>><<   XT:A:R  NM:i:2  SM
9:21597+10M2I25M:R:-209 83      1       21678   0       8M2I27M =       21469   -244    CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT   <;9<<5><<<<><<<>><<><>><<9>><>>>9>>><>   XT:A:R  NM:i:2  SM:i:0  AM
```

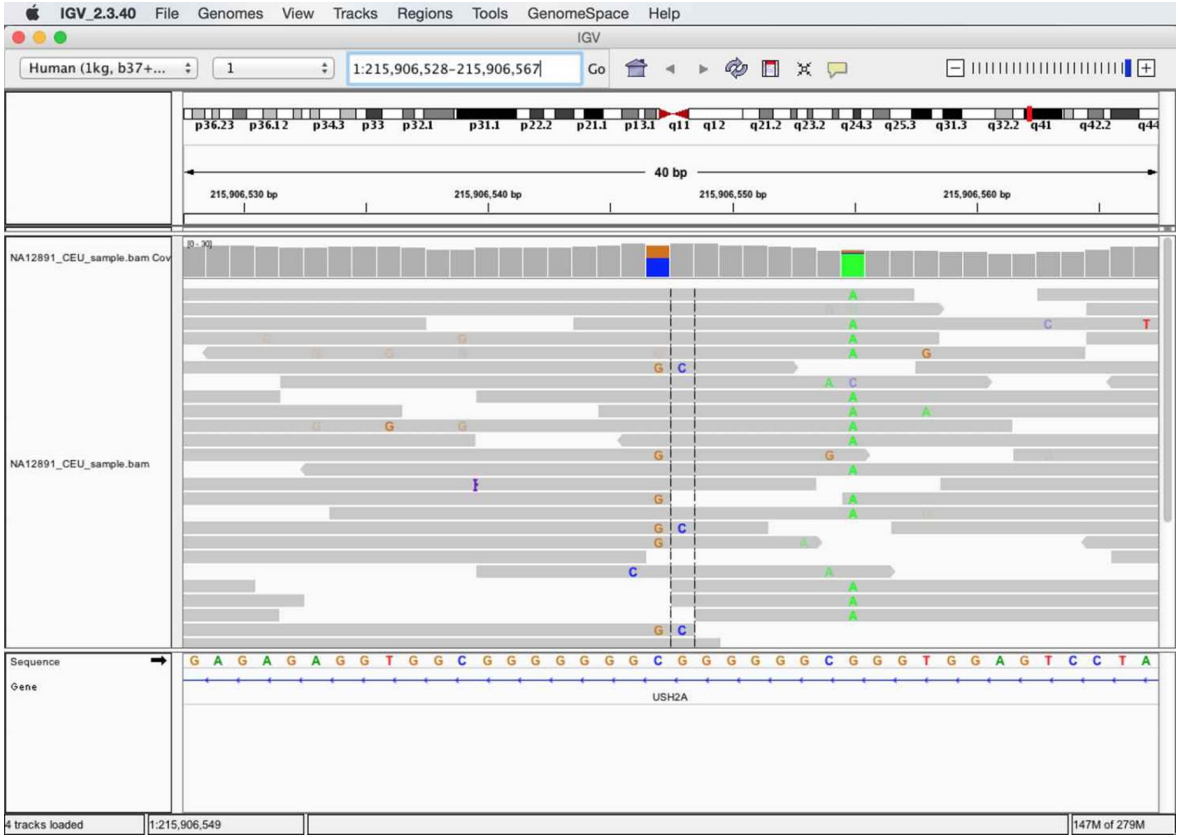
In this example, the fields are:

Field	Alignment 1	Alignment 2	Alignment 3	Alignment 4
QNAME	1:497:R:-272+13M17D24M	19:20389:F:275+18M2D19M	19:20389:F:275+18M2D19M	9:21597+10M2I25M:R:-209
FLAG	113	99	147	83
RNAME	1	1	1	1
POS	497	17644	17919	21678
MAPQ	37	0	0	0
CIGAR	37M	37M	18M2D19M	8M2I27M
MRNM/RNEXT	15	=	=	=
MPOS/PNEXT	100338662	17919	17644	21469
ISIZE/TLEN	0	314		
SEQ	CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG	TATGACTGCTAATAATACCTACACATGTTAGAACCAT	GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT	CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT
QUAL	0;==-==9;>>>>>>=>>>>>>>>>>>>>=>>>>>>>>>>	>>>>>>>>>>>>>>>>>>>>>><>>><>>>4::>>:<9	;44999;499<8<8<<8<<><<<><7<;<<<>><<	<;9<<5><<<<><<<>><<><>><<9>><>>>9>>><>
TAGs	XT:A:U NM:i:0 SM:i:37 AM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37	RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37	XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19	XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:5 X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:35

http://genome.sph.umich.edu/wiki/SAM
https://en.wikipedia.org/wiki/SAM_(file_format)

J. Craig Venter™ INSTITUTE

Visualization – IGV Pile Up



J. Craig Venter™ INSTITUTE

ONT base calling, quality control, and methylation/modification profiles

- Changes in current as a molecule is pulled through a pore is changed into "squiggle plots" by the *Minknow* software - this can also be used to determine methylation or BrdU incorporation (fast5 files)
- *Guppy* converts the fast5 files into fastq files, i.e., standard base calls with quality scores in either real time or post run
  - Real time Guppy combined with K-mer analysis allows for "adaptive sequencing" where pores will recognize unwanted DNA in real time, reverse polarity, and reject the untargeted sequences
- *Nanofilt* is used for quality trimming of fastq files
- The fastq and fast5 are used in combination to determine methylation profiles using *Nanopolish*

What can you do with these data?

- High quality genome assembly – *Flye*
- Viral variant detection- *Variabel*
- De-novo bulk transcript assembly with splice variants- *Pinfish*
- 10X single cell transcript assembly- *Sockeye*
- Differential methylation analysis- *pycoMeth*

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-022-28852-1 OPEN

Rescuing low frequency variants within intra-host viral populations directly from Oxford Nanopore sequencing data

Yunxi Liu[1,3], Joshua Kearney[1,3], Medhat Mahmoud[2], Bryce Kille[1], Fritz J. Sedlazeck[1,2] & Todd J. Treangen[1]

Article | Published: 01 April 2019

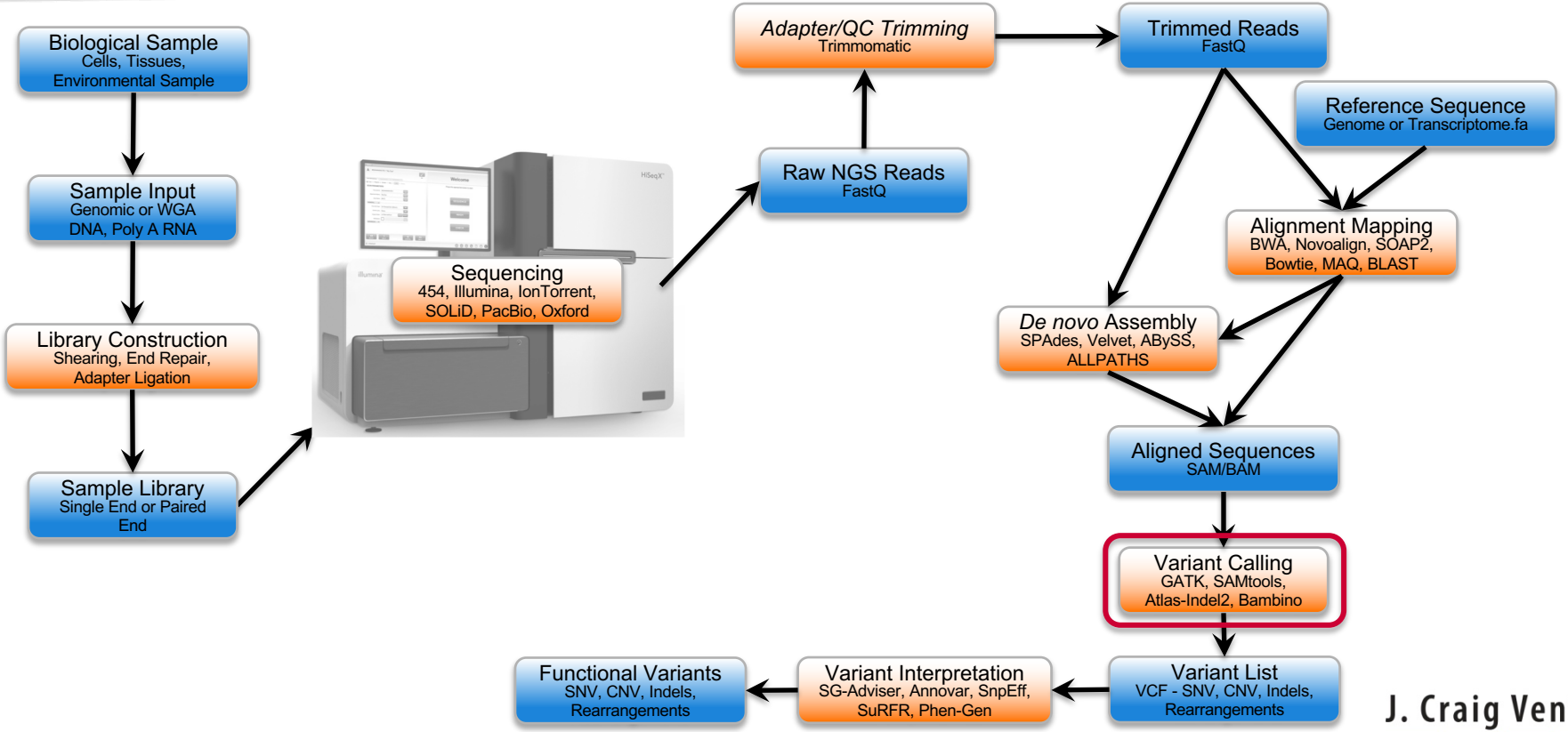
Assembly of long, error-prone reads using repeat graphs

Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin & Pavel A. Pevzner

Nature Biotechnology **37**, 540–546 (2019) | Cite this article

J. Craig Venter™ INSTITUTE

NGS Processing Workflows



J. Craig Venter™ INSTITUTE

Variant Calling

- ***Strategy:*** Determine the presence of sequence variations within a sample (alleles in a diploid organism, quasi-species in population samples, somatic mutation heterogeneity in cancer tissues) and sequence variations between sets of samples or between a sample and a reference
- Single nucleotide variants (SNVs), copy number variants (CNVs), insertions and deletions (Indels), rearrangements
  - **GATK** - *Broad Institute* https://www.broadinstitute.org/gatk/
  - **SAMtools** - *Wellcome Trust Sanger Institute* https://github.com/samtools/samtools
  - **Atlas-Indel2 -** *Baylor College of Medicine* http://sourceforge.net/p/atlas2/wiki/Atlas-Indel/
  - **Bambino -** *National Cancer Institute* https://github.com/NCIP/cgr-bambino

Variant detection through NGS

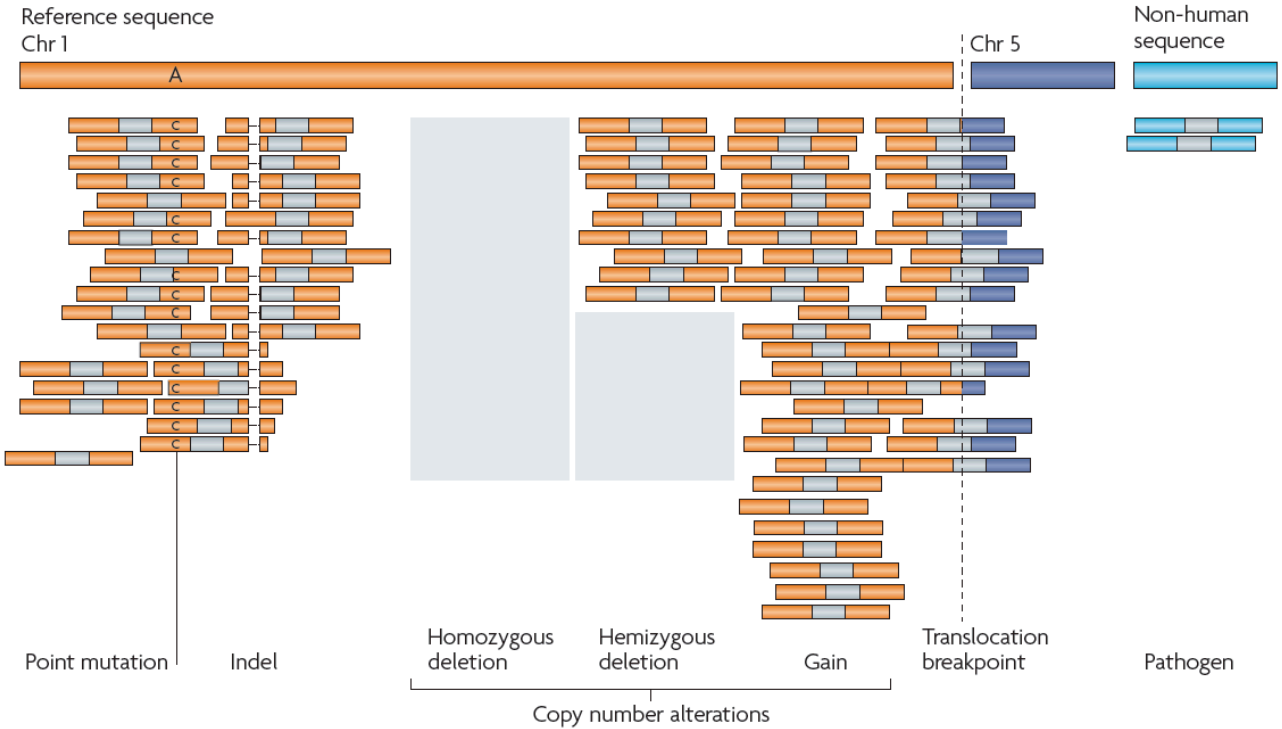


Figure 3 | **Types of genome alterations that can be detected by second-generation sequencing.** Sequenced

Meyerson et al. NRG 2010

J. Craig Venter INSTITUTE

VCF File Format

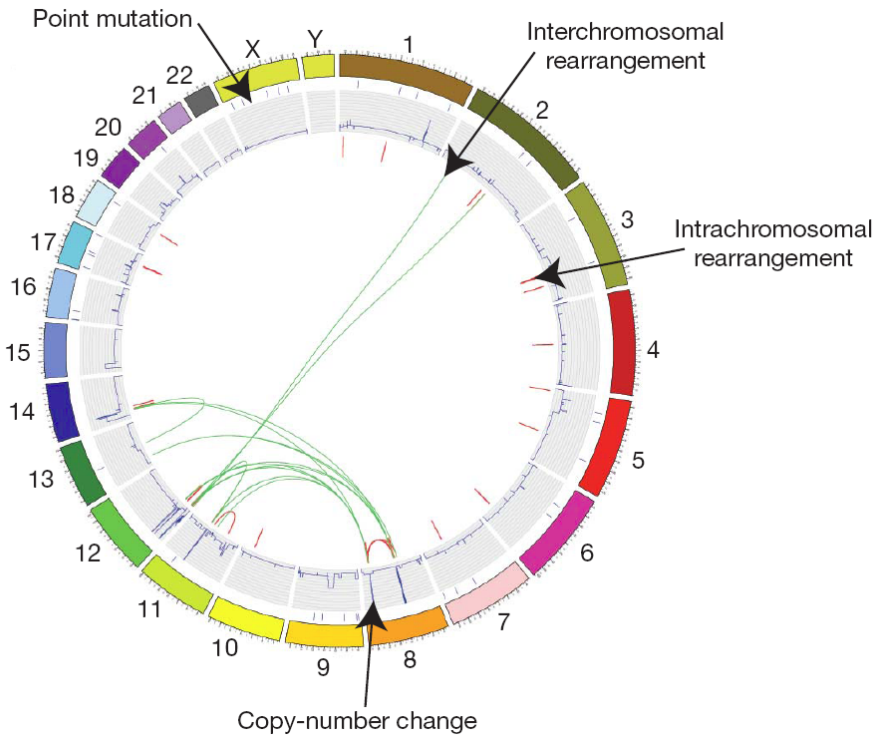
- Developed for the 1000 Genomes Project
- Store only the variant information – SNVs and Indels

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT     QUAL FILTER INFO                              FORMAT      Sample1         Sample2         Sample3
2      4370    rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:.,.
2      7330    .         T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3   0/0:41:3
2      110696  rs6055    A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
2      130237  .         T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51 0/0:61:2
2      134567  microsat1 GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4        0/2:17:2        1/1:40:3
```

https://vcftools.github.io/specs.html

J. Craig Venter INSTITUTE

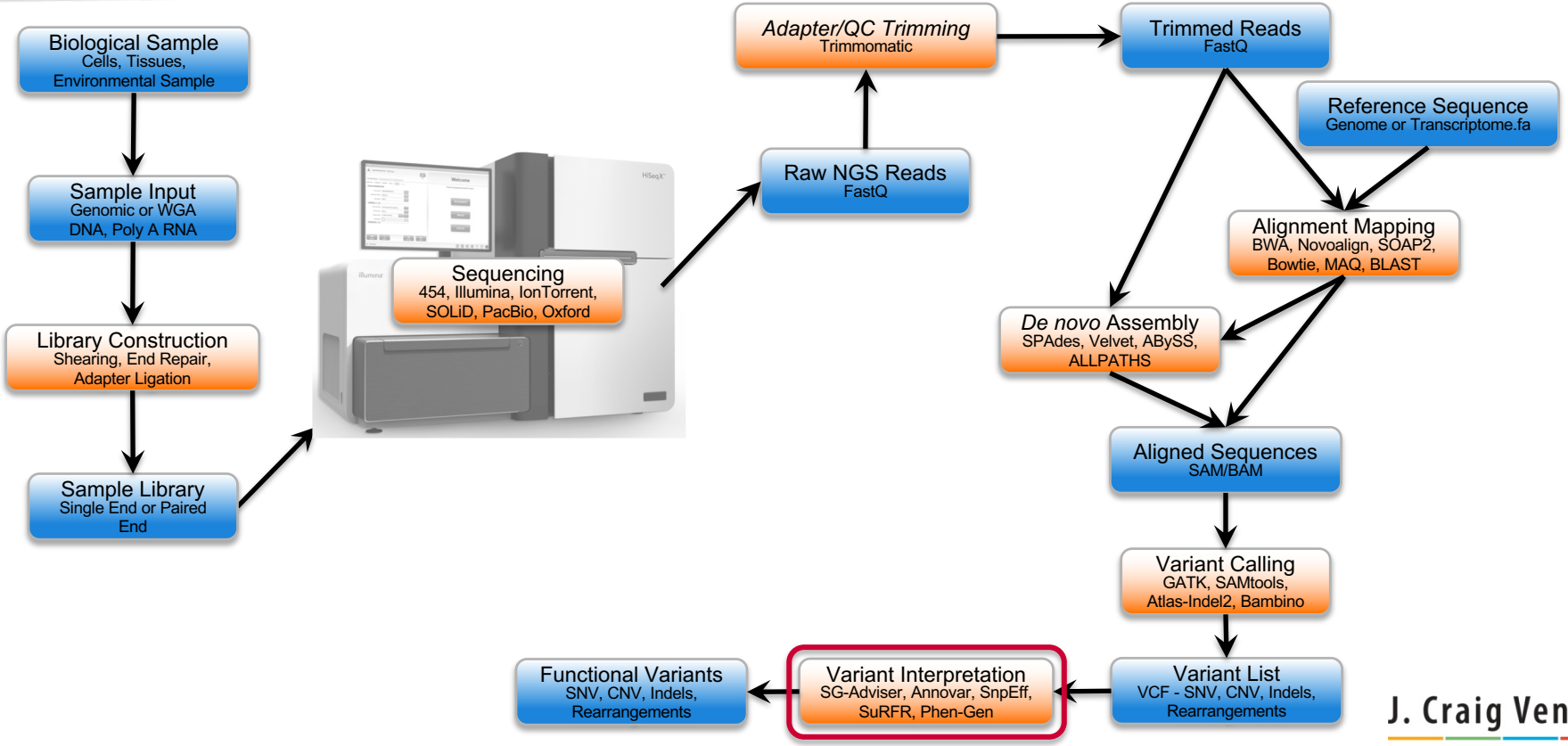
Visualization – Circos Plots



M. Stratton *et al. Nature,* 458 (2009)

J. Craig Venter INSTITUTE

NGS Processing Workflows



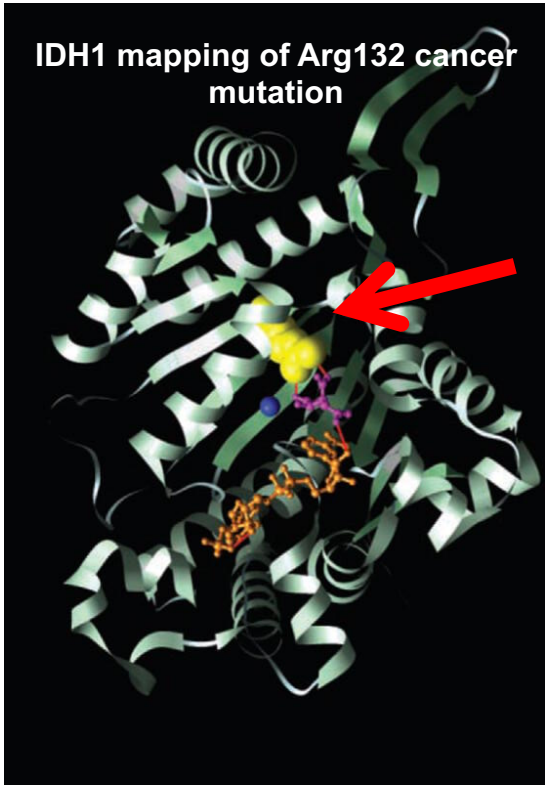
J. Craig Venter™ INSTITUTE

Variant Interpretation

- ***Strategy:*** Compare against databases of known variants with functional information and predict functional effects of novel variants
- Databases of known variants – dbSNP, ClinVar, OMIM, Cosmic, Ensembl Variation
- Functional predictions – SIFT and PolyPhen-2
  - **SG-Adviser** - *Scripps Research Institute,* https://genomics.scripps.edu/ADVISER/
  - **Annovar** – *University of Pennsylvania,* http://www.openbioinformatics.org/annovar/
  - **SnpEff** – *Wayne State University,* http://snpeff.sourceforge.net/
  - **SuRFR** – *University of Edinburgh,* http://www.cgem.ed.ac.uk/resources/SuRFR/SuRFR_Sweave_v1.pdf
  - **Phen-Gen** - *Genome Institute of Singapore,* http://phen-gen.org/
  - **Promethease** – *SNPedia,* http://www.snpedia.com/index.php/Promethease/

Interpretation of genetic variants

- Substitutions translated bioinformatically
- SIFT - probability that a substitution is tolerated
  - < 0.05 is deleterious.
- PolyPhen – categorical definitions
  - "benign", "possibly damaging" and "probably damaging"
- Protein structural mapping



Parson et al., *Science*, (2008)

J. Craig Venter INSTITUTE

SG-Adviser Example

Chromosom	Begin	End	VarType	Reference	Allele	Gene_Type	Location	Coding_Impact	Protein_Impact_Prediction(Polyphen)	Protein_Impact_Prediction(SIFT)	COSMIC_Gene~NumSamples	Gene_Symbol
chr3	140419772	140419773	snp	G	T	Protein_Coding	Exon_5	Nonsynonymous	probably damaging	INTOLERANT	carcinomaNSmalignant_me	TRIM42
chr5	79030319	79030320	snp	C	G	Protein_Coding	Exon_2	Nonsynonymous	benign	TOLERANT	NS$carcinoma~41	CMYA5
chr11	5011043	5011044	snp	A	G	Protein_Coding	Exon_2	Nonsynonymous	benign	TOLERANT	carcinoma~6	MMP26
chr14	102514936	102514937	snp	G	A	Protein_Coding	Exon_74	Nonsynonymous	benign	TOLERANT	glioma$carcinoid-endocrine_t	DYNC1H1
chr16	3255108	3255109	snp	T	A	Protein_Coding	Exon_1	Nonsynonymous	probably damaging	INTOLERANT	carcinoma~2	OR1F1
chr20	39991131	39991132	snp	C	A	Protein_Coding	Exon_4	Nonsynonymous	probably damaging	INTOLERANT	carcinoma~3	EMILIN3
chr19	54697485	54697486	snp	C	T	Protein_Coding/	Exon_6//	Nonsynonymous,	possibly damaging///-///-///-	INTOLERANT///-///-///-	carcinoma~1///carcinoma~1/,	TSEN34
chr17	4689235	4689236	snp	C	G	Protein_Coding/	Exon_2//	Nonsynonymous,	probably damaging///-///probably dar	INTOLERANT///-///INTOLERANT/	carcinoma~5///carcinoma~5/,	VMO1
chr7	144077014	144077015	snp	T	C	Protein_Coding/	Exon_15/	Nonsynonymous,	probably damaging///probably damagi	INTOLERANT///TOLERANT	carcinomaNSmalignant_me	ARHGEF5
chr15	31220826	31220827	snp	G	A	Protein_Coding/	Exon_11/	Nonsynonymous,	possibly damaging///possibly damagin	TOLERANT///INTOLERANT	-~-///-~-	MTMR15
chr18	28588004	28588005	snp	A	T	Protein_Coding/	Exon_11/	Nonsynonymous,	possibly damaging///possibly damagin	INTOLERANT///INTOLERANT	carcinoma$malignant_melanc	DSC3
chr6	47650081	47650082	snp	T	C	Protein_Coding/	Exon_6//	Nonsynonymous,	probably damaging///probably damagi	INTOLERANT///INTOLERANT///-	malignant_melanoma$carcinc	GPR111
chr12	88514807	88514808	snp	A	G	Protein_Coding/	Exon_14/	Nonsynonymous,	benign///benign///-	N/A///TOLERANT///-	NS$carcinoma~18///NS$carci	CEP290
chr6	97726730	97726731	snp	C	A	Protein_Coding/	Exon_3//	Nonsynonymous,	probably damaging///probably damagi	INTOLERANT///INTOLERANT///-/	-~-///-~-///-~-///-~-///-~-	C6orf167
chr2	241702740	241702741	snp	C	A	Protein_Coding/	Exon_20/	Nonsynonymous,	benign///benign///benign	INTOLERANT///TOLERANT///INT	NS$carcinoma$lymphoid_neo	KIF1A
chr7	43351425	43351426	snp	G	A	Protein_Coding/	Exon_4//	Nonsynonymous,	benign///benign///benign	TOLERANT///TOLERANT///TOLER	primitive_neuroectodermal_t	KIAA0322///HE
chr5	176520167	176520168	snp	C	T	Protein_Coding/	Exon_9//	Nonsynonymous,	benign///benign///benign///-///-	TOLERANT///TOLERANT///TOLER	malignant_melanomaNScar	FGFR4
chr1	103548371	103548372	snp	T	C	Protein_Coding/	Exon_2//	Nonsynonymous,	unknown///unknown///unknown///ur	N/A///N/A///N/A///N/A	NS$carcinoma$glioma~45///N	COL11A1
chr1	180063725	180063726	snp	C	T	Protein_Coding/	Exon_34/	Nonsynonymous,	probably damaging///probably damagi	INTOLERANT///INTOLERANT///IN	NS$carcinoma$lymphoid_neo	CEP350
chr6	76023775	76023776	snp	C	T	Protein_Coding/	Exon_5//	Nonsynonymous,	probably damaging///probably damagi	INTOLERANT///INTOLERANT///IN	carcinoma$malignant_melanc	FILIP1

Gene_Symbol	DrugBank	Reactome_Pathway	Gene_Onotology	Disease_Ontology	ACMG_Score_Research~Disease_Entry~Explanation
TRIM42	-	-	-	-	-~-~-
CMYA5	-	-	-	-	-~-~-
MMP26	DB00786	-	GO:0006508~proteolysis	DOID:5616~intraepithelial	-~-~-
DYNC1H1	-	Cell Cycle~Cell Cycle,	GO:0006810~transport$(	-	4~Charcot-Marie-Tooth disease, axonal(DYNC1H1)$Mental retardation(DYNC1H1)~Rare, Amino Acid Change Predicted Neutral
OR1F1	-	GPCR downstream si	GO:0007165~signal trans	-	-~-~-
EMILIN3	-	-	-	DOID:162~cancer	-~-~-
TSEN34	-	-	GO:0000379~tRNA-type i	-	3~Pontocerebellar hypoplasia(TSEN34)~Rare, Amino Acid Change Predicted Damaging with Low Confidence
VMO1	-	-	GO:0030704~vitelline me	-	-~-~-
ARHGEF5	-	-	GO:0035023~regulation	DOID:684~hepatocellular c	-~-~-
MTMR15	-	-	-	-	-~-~-
DSC3	-	-	GO:0007155~cell adhesic	DOID:9256~colorectal canc	3~Hypotrichosis & recurrent skin vesicles(DSC3)~Rare, Amino Acid Change Predicted Damaging with Low Confidence
GPR111	-	-	GO:0007186~G-protein c	-	-~-~-
CEP290	-	Cell Cycle~Cell Cycle,	GO:0015031~protein tra	DOID:14791~Leber congen	4~Joubert syndrome(CEP290)$Joubert syndrome, Senior-Loken type(CEP290)$Senior-Loken syndrome(CEP290)$Meckel syndron
C6orf167	-	-	-	-	-~-~-
KIF1A	DB03431~I	-	GO:0007018~microtubul	-	3~Intellectual disability, nonsyndromic(KIF1A)$Spastic paraparesis(KIF1A)~Rare, Amino Acid Change Predicted Damaging with Lo
KIAA0322///HE	-///-	-///-	-///GO:0006464~protein	-///-	-~-~-
FGFR4	DB00039	Downstream signalin	GO:0006468~protein pho	DOID:10008~malignant ne	3~Cancer, accelerated progression, association with(FGFR4)~Rare, Amino Acid Change Predicted Damaging with Low Confidence
COL11A1	-	-	GO:0007155~cell adhesic	DOID:657~adenoma$DOID	4~Fibrochondrogenesis(COL11A1)$Marshall / Stickler syndrome(COL11A1)$Osteoarthritis, early-onset ?(COL11A1)$Lumbar disc
CEP350	-	-	-	-	-~-~-
FILIP1	-	-	-	-	-~-~-

Craig Venter™ INSTITUTE