Webinar – ICGEB, Delhi, 24.07.2020

Dr. Vipin Singh

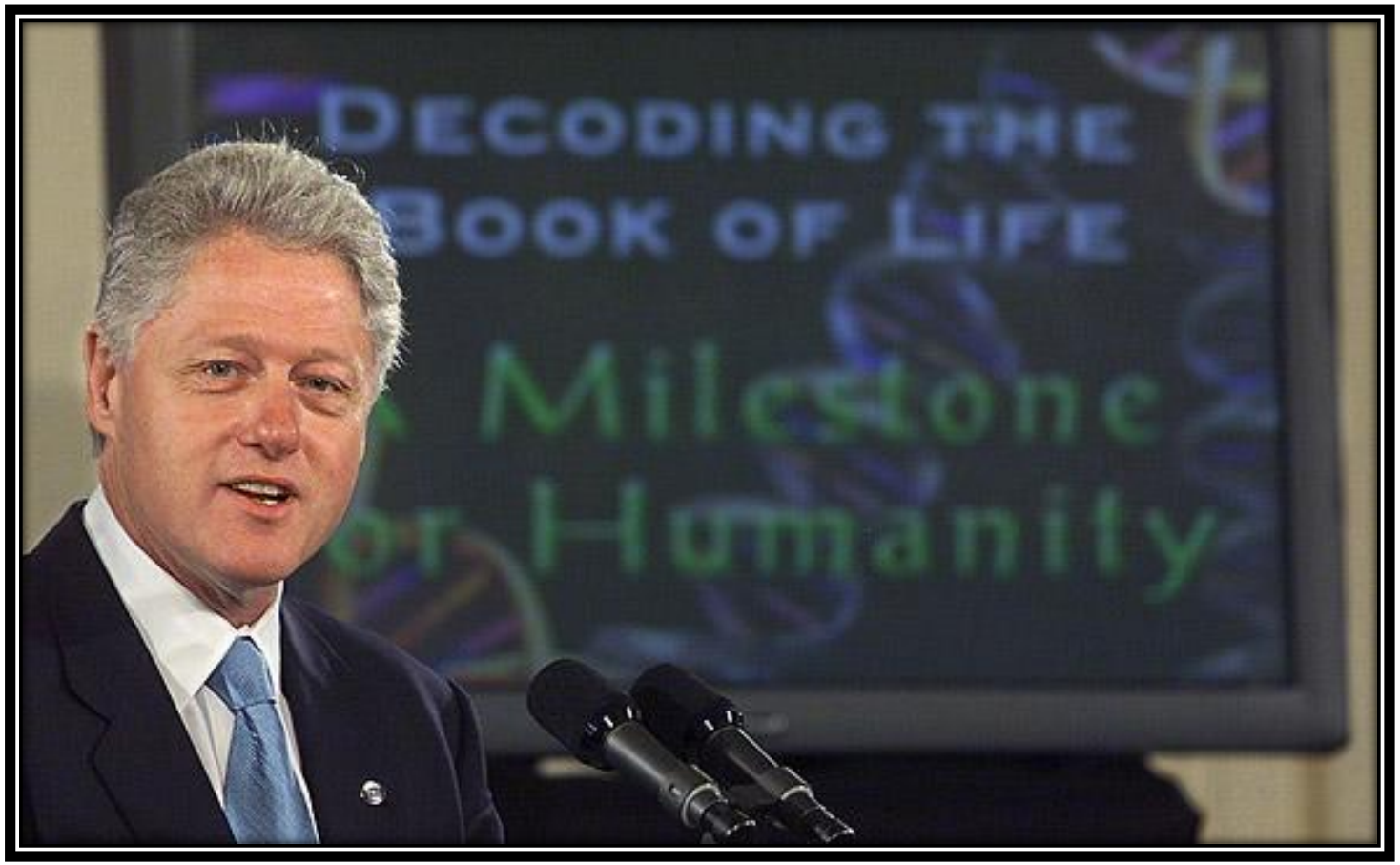
Post Doctoral Fellow

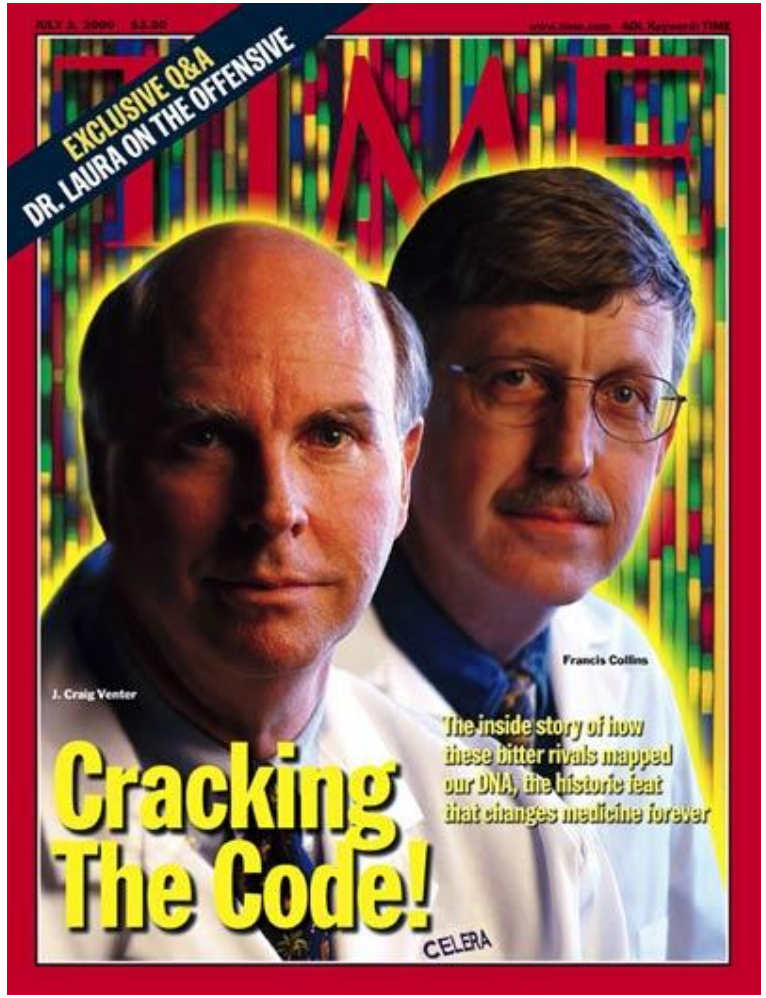
Institute of Biology, Ecole Normal Superior (IBENS), Paris

Associate Professor, University Institute of Biotechnology,

Chandigarh University

Cracking the Code





February 2001

Sequencing Techniques

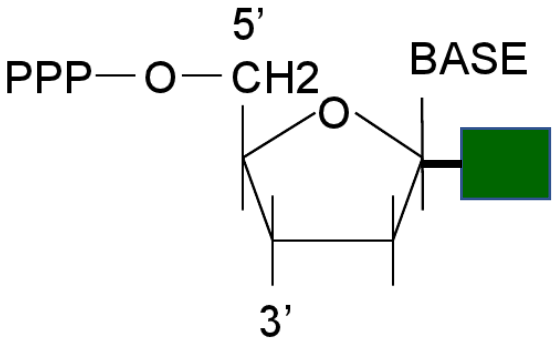
Polymerase based – Most sequencing techniques

Ligase based - SOLiD

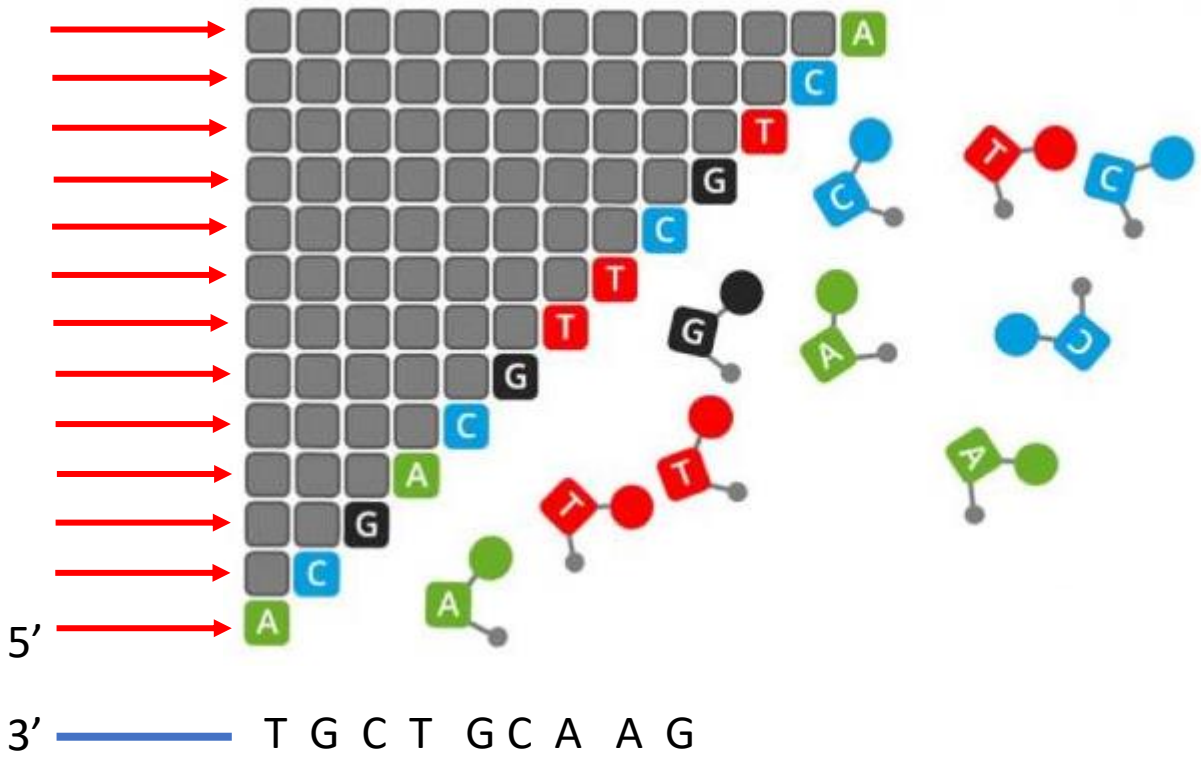
Helicase based - Nanopore

Sanger Sequencing - Dideoxy nucleotide

$$(DNA)n + \boxed{\textbf{ddNTP}} \xrightarrow[\text{Polymerase}]{\text{DNA}} (DNA)n+1 + \textbf{Ppi} + \textbf{H}^{+}$$



no hydroxyl group at 3' end
prevents strand extension



Differential flourochrome labeling of ddNTPs in automated Sanger sequencing

Automated Sanger Sequencer

Major Achievements

1995 – *Haemophylus influenzae*

1996 – Yeast

2000 – *Drosophila, Arabidopsis*

2001 – Human genome

The evolution of Sequencing technologies

Classical or first generation (1976 – 2002) -
Maxam Gilbert, Sanger Coulson

Second Generation –(2002 – till date) -
Pyrosequencing, Virtual terminator sequencing, SoLid

Third Generation – 2008 – evolving –
Nanopore, Ion torrent, SMRT

Rapidly evolution and fine tuning and extinction

Name	Company	PCR	Sequencing	Read-length
454	Roche	Emulsion	Polymerase- Pyrosequencing	250
Solexa	Illumina	Bridge	Polymerase- reversible terminators	36
SOLiD	Applied Biosystems	Emulsion	Ligase (octamers with 2 base encoding)	35

2009

High-end sequencing- Platform†	Sequencing chemistry	Read lengths/ through put	Run time	Template prep	Application
Roche 454 -Titanium FLX	Pyrosequencing	400 bp 400 Mb/run	10 hours	Emulsion PCR	Denovo WGS of microbes, pathogen discovery, Exome seq
Illumina/Solexa -HiSeq 2000	Reversible terminator chemistry	2×100bp 600 GB/ run (dual cell)	11.5 days	Solid-phase	Human WGS, exome seq, RNA-seq, Methylation
ABI/LifeTechnology-SOLiD 5550XL	Sequencing by ligation	2×60bp 15 GB/day	8 days	Emulsion PCR	Human WGS, exome seq, RNA-seq, Methylation

454 and SOLiD have been phased out

Contemporary Sequencing technologies

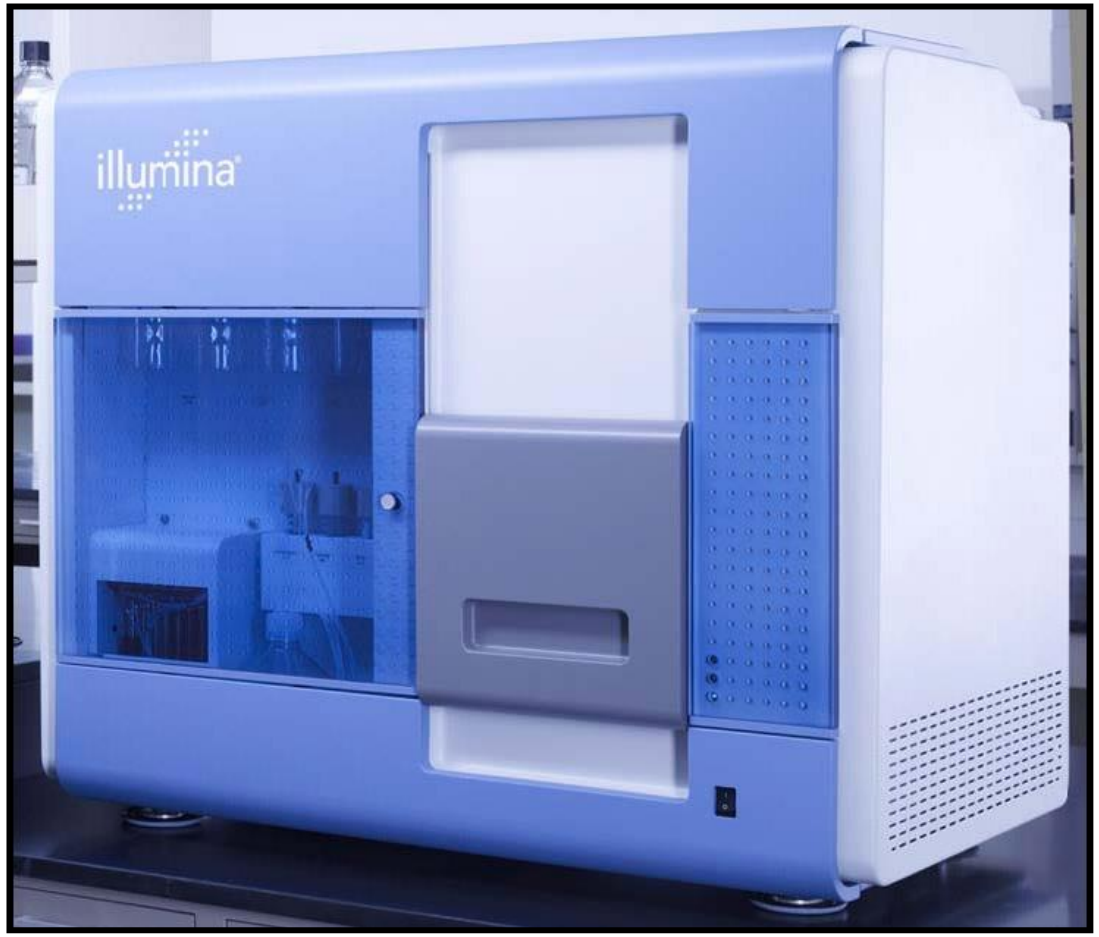
Short Read Sequencers

- Virtual Terminator Sequencing (Illumina)
- Ion Torrent (Roche)

Long Read Sequencers

- Nanopore (Oxford)
- SMRT (Pac Bio)

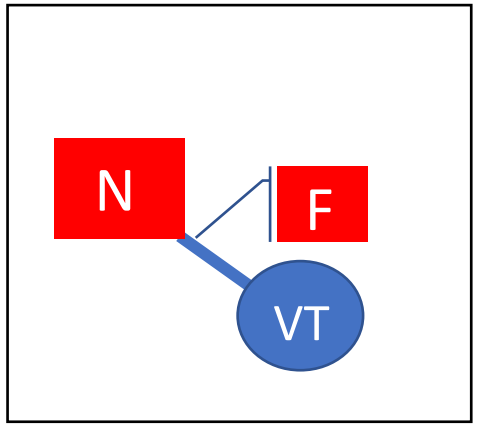
Virtual Terminator sequencing – Illumina, Solexa



Virtual terminator sequencing

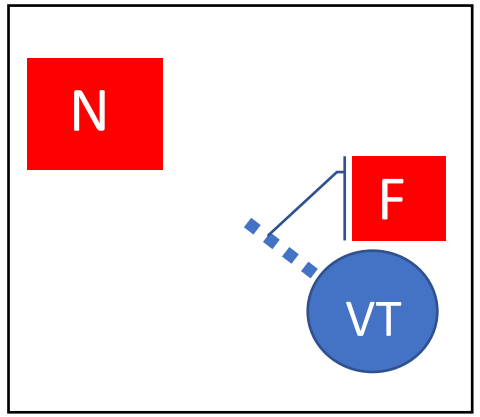
$$(DNA)n + \mathbf{ddNTP} \xrightarrow[\text{Polymerase}]{\text{DNA}} (DNA)n+1 + \mathbf{Ppi} + \mathbf{H^+}$$

The virtual terminator (VT) nucleotides are nucleotide analogues, contain a fluorescent dye and chemically cleavable group (VT) – **3'-O-azidomethyl group**.



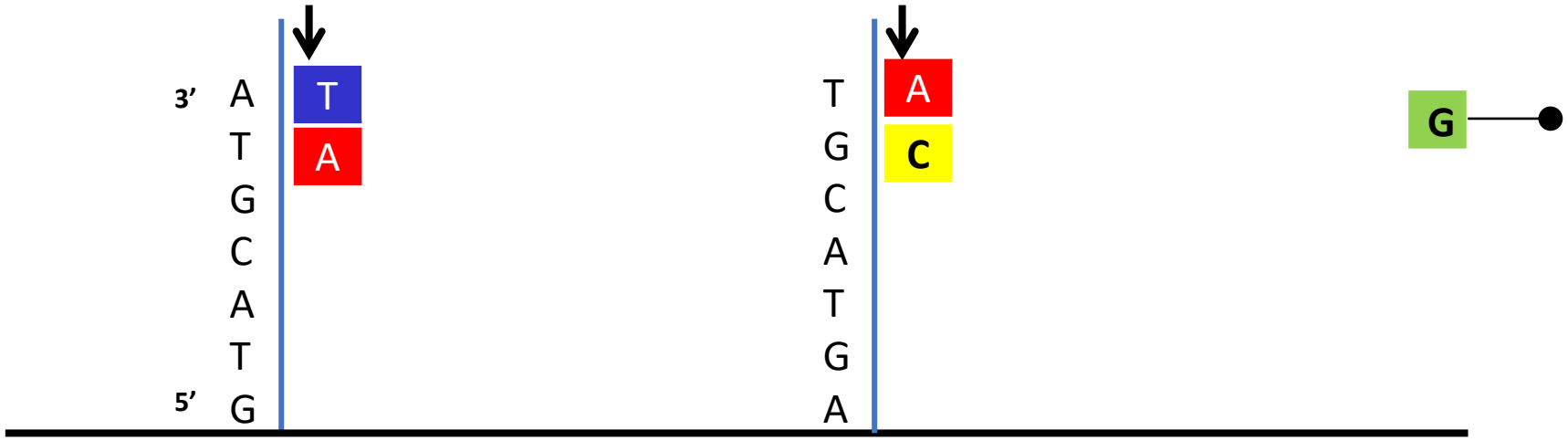
Once incorporated, the VT analogues block further incorporation untill the VT moiety is chemically removed

Hence the name – "**Virtual/ Reversible** terminators".



The virtual terminator label is removed before the next cycle **, regenerating the 3' OH group using reducing agent tris 2 carboxy ethyl phosphine (TCEP)**

Virtual terminator Sequencing



Cycle steps –

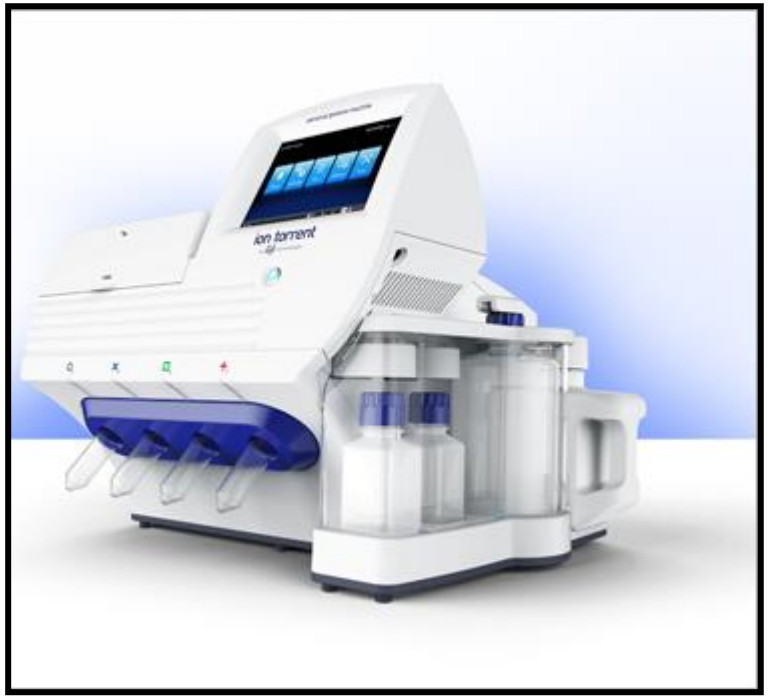
1. Primer is extended through virtual terminator nucleotides **by adding all four differentially labelled nucleotides simultaneously**

2. Because of the VT moitey **only one nucleotide is added at a time and recorded** for each sequence based on differential florescence tagging

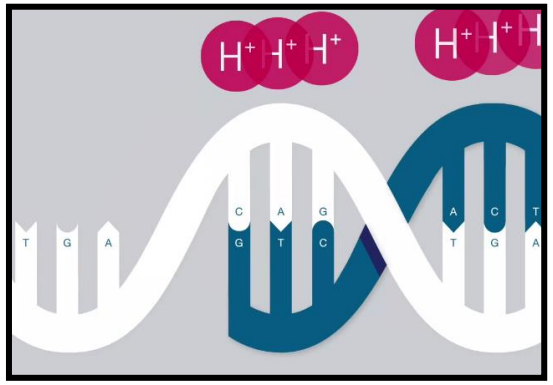
3. Next the **virtual terminator moitey is removed, regenerating the 3' OH group using reducing agent tris 2 carboxy ethyl phosphine (TCEP)**

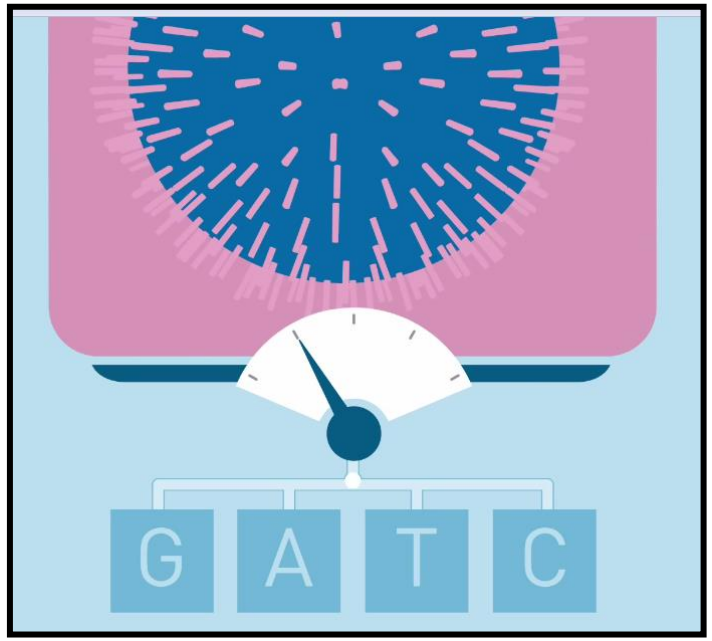
Steps one to three are repeated till the template is fully sequenced

Ion torrent



Ion torrent





The flow cell is flushed with a **single species of nucleotide at a time**

The semi-conductor chip records **changes in pH**, upon incorporation of complimentary nucleotide by polymerase

The **change** in pH **is proportional to the number of nucleotides added**

Current read length 100 bp

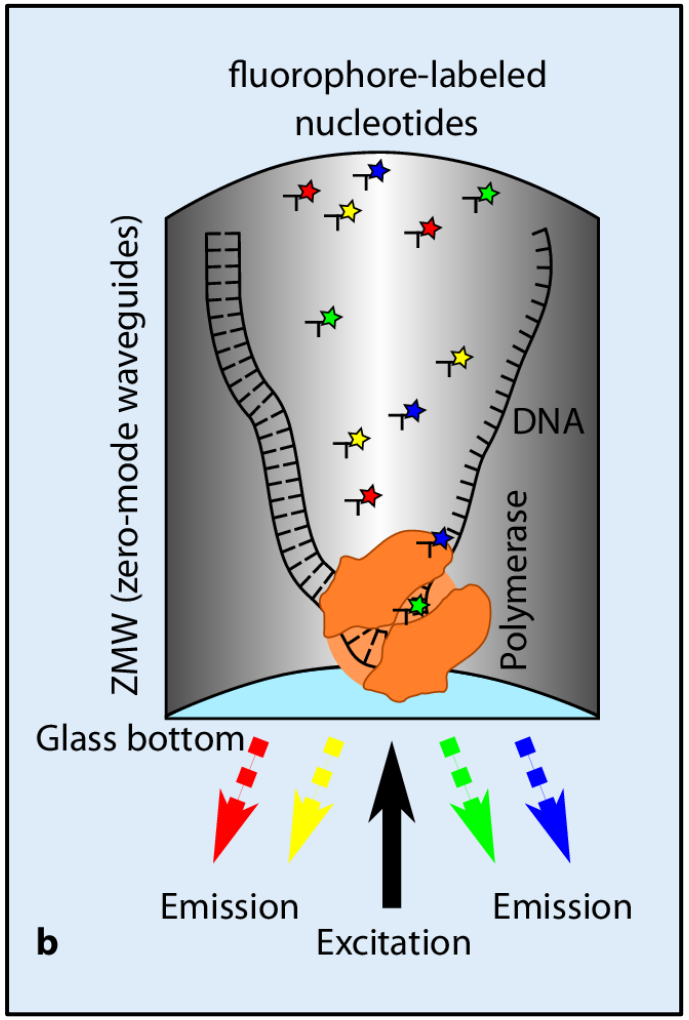
No imaging

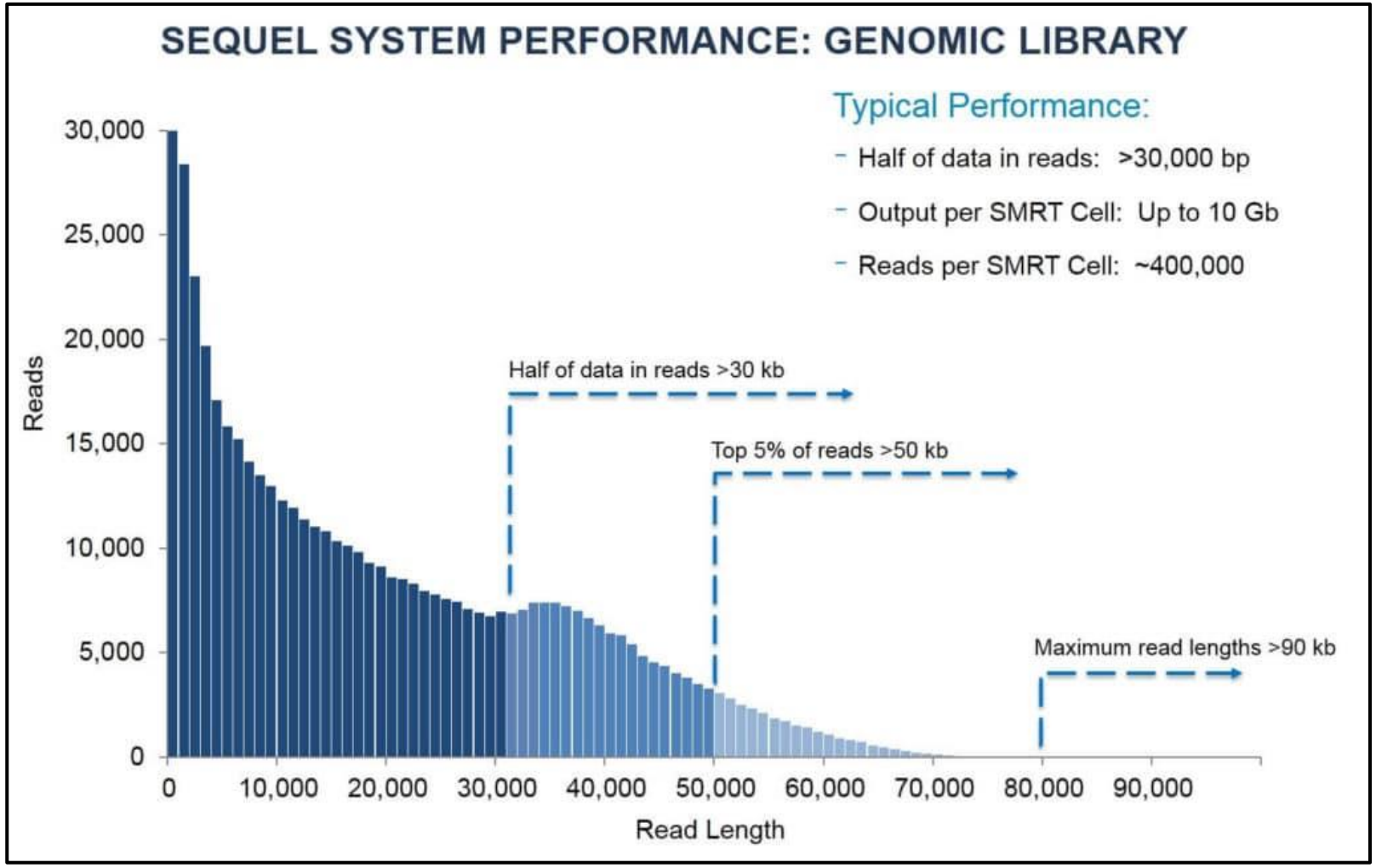
Long read sequencers

Long Read Sequencers

- SMRT(Pac Bio)
- Nanopore (Oxford)

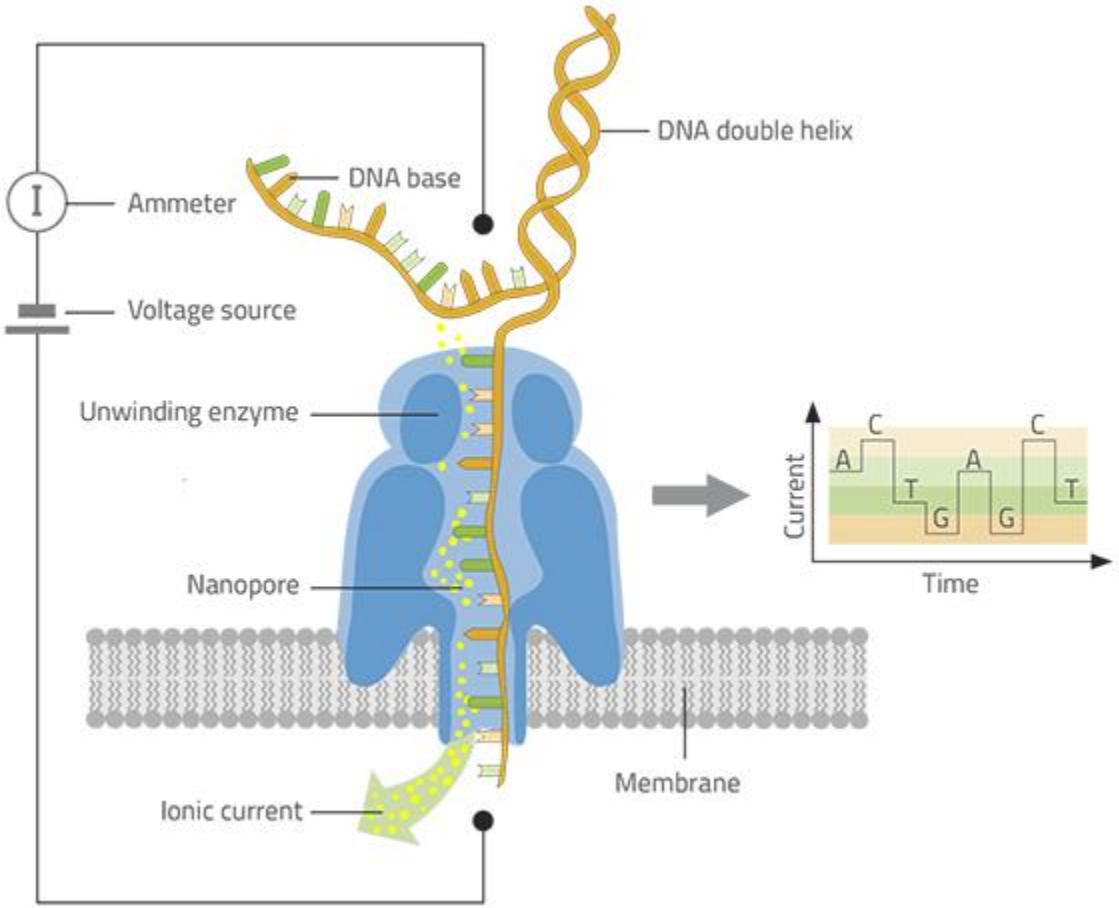
SMRT – Single Molecule Real Time Sequencing -PacBio





Nanopore – Helicase based sequencing

Oxford Nanopore



Marketed by Oxford Nanopore

Time to sequence human genome – 15 to 20 minutes

Size of sequencer – pen drive, scalable

Cost of sequencing whloe human genome - < 1000$

The long and short of sequencing

Table 1 | **Data type, length, accuracy, throughput and cost across long-read and short-read technologies and platforms**

Sequencing technology	Platform	Data type	Read length (kb)		Read accuracy (%)	Throughput per flow cell (Gb)		Estimated cost per Gb (US$)	Maximum throughput per year (Gb)[a]
			N50	Maximum		Mean	Maximum		
Pacific Biosciences (PacBio)	RS II[b]	CLR	5–15	>60	87–92	0.75–1.5	2	333–933[c]	4,380
	Sequel	CLR	25–50	>100		5–10	20	98–195[d]	17,520
	Sequel II	CLR	30–60	>200		50–100	160	13–26[e]	93,440
		HiFi	10–20	>20	>99	15–30	35	43–86[e]	10,220
Oxford Nanopore Technologies (ONT)	MinION/ GridION	Long	10–60	>1,000	87–98	2-20	30	50–500[f]	21,900 (MinION) 109,500 (GridION)
		Ultra-long	100–200	>1,500		0.5–2	2.5	500–2,000[f]	913 (MinION) 4,563 (GridION)
	PromethION	Long	10–60	>1,000		50–100	180	21–42[f]	3,153,600
Illumina	NextSeq 550	Single-end	0.075–0.15	0.15	>99.9	16–30	>30	50–63[g]	>47,782
		Paired-end	0.075–0.15 (×2)	0.15 (×2)		32–120	>120	40–60[g]	>70,080
	NovaSeq 6000	Single-end	0.05–0.25	0.25		65–3,000	>3,000	10–35[h]	>1,194,545
		Paired-end	0.05–0.25 (×2)	0.25 (×2)					

https://www.nature.com/articles/s41576-020-0236-x

The Game changer – ultra long reads

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain[1,13], Sergey Koren[2,13], Karen H Miga[1,13], Josh Quick[3,13], Arthur C Rand[1,13], Thomas A Sasani[4,5,13], John R Tyson[6,13], Andrew D Beggs[7], Alexander T Dilthey[2], Ian T Fiddes[1], Sunir Malla[8], Hannah Marriott[8], Tom Nieto[7], Justin O'Grady[9], Hugh E Olsen[1], Brent S Pedersen[4,5], Arang Rhie[2], Hollian Richardson[9], Aaron R Quinlan[4,5,10], Terrance P Snutch[6], Louise Tee[7], Benedict Paten[1], Adam M Phillippy[2], Jared T Simpson[11,12], Nicholas J Loman[3] & Matthew Loose[8]

We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30× theoretical coverage, were produced. Reference-based alignment enabled detection of large structural variants and epigenetic modifications. *De novo* assembly of nanopore reads alone yielded a contiguous assembly (NG50 ~3 Mb). We developed a protocol to generate ultra-long reads (N50 > 100 kb, read lengths up to 882 kb). Incorporating an additional 5× coverage of these ultra-long reads more than doubled the assembly contiguity (NG50 ~6.4 Mb). The final assembled genome was 2,867 million bases in size, covering 85.8% of the reference. Assembly accuracy, after incorporating complementary short-read sequencing data, exceeded 99.8%. Ultra-long reads enabled assembly and phasing of the 4-Mb major histocompatibility complex (MHC) locus in its entirety, measurement of telomere repeat length, and closure of gaps in the reference human genome assembly GRCh38.

Nature Biotechnology - 29 January 2018; doi:10.1038/nbt.4060

REVIEWS

Check for updates

Long-read human genome sequencing and its applications

Glennis A. Logsdon[1], Mitchell R. Vollger[1] and Evan E. Eichler[1,2]

Abstract | Over the past decade, long-read, single-molecule DNA sequencing technologies have emerged as powerful players in genomics. With the ability to generate reads tens to thousands of kilobases in length with an accuracy approaching that of short-read sequencing technologies, these platforms have proven their ability to resolve some of the most challenging regions of the human genome, detect previously inaccessible structural variants and generate some of the first telomere-to-telomere assemblies of whole chromosomes. Long-read sequencing technologies will soon permit the routine assembly of diploid genomes, which will revolutionize genomics by revealing the full spectrum of human genetic variation, resolving some of the missing heritability

Nature Reviews Genetics
https://www.nature.com/articles/s41576-020-0236-x

NGS Applications

NGS Broad Applications

1. Variant calling (SNPs or structural variations)

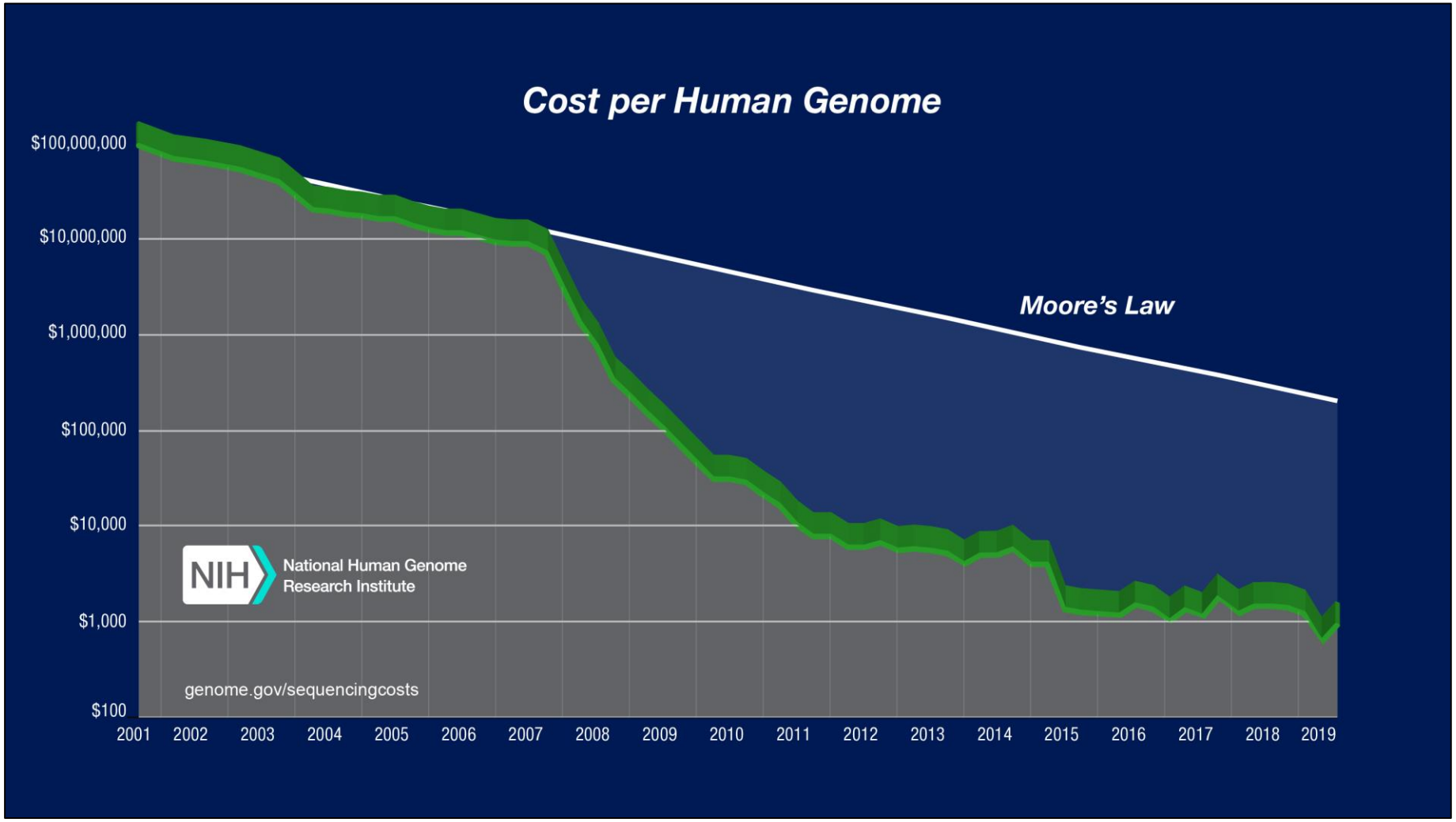
2. Epigenomics study - Methylation status of Cytosines in CG, CHG and CHH contexts

3. Transcriptome analysis – reconstitution of mRNA, identification of exon intron boundaries, alternative splicing, differential expression analysis

4. Sequence specific binding of proteins (DNA-Protein interactions)

DNA-Seq **BS-Seq** **RNA-Seq** **ChIP-Seq**

Cost and time efficient



https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

High resolution - Mendelian disease genes from exome sequencing

Table 2. Mendelian disease gene identifications by exome or genome sequencing

Disorder	Inheritance	Gene identified	Scope	References
Congenital chloride diarrhea	Recessive	*SLC26A3*	Exome	Choi *et al.* [16]
Miller syndrome	Recessive	*DHODH*	Exome	Ng *et al.* [14]
Charcot-Marie-Tooth neuropathy	Recessive	*SH3TC2*	Genome	Lupski *et al.* [20]
Metachondromatosis	Dominant	*PTPN11*	Genome	Sobreira *et al.* [23]
Schinzel-Giedion syndrome	Dominant	*SETBP1*	Exome	Hoischen *et al.* [29]
Nonsyndromic hearing loss	Recessive	*GPSM2*	Exome	Walsh *et al.* [69]
Perrault syndrome	Recessive	*HSD17B4*	Exome	Pierce *et al.* [25]
Hyperphosphatasia mental retardation syndrome	Recessive	*PIGV*	Exome	Krawitz *et al.* [68]
Sensenbrenner syndrome	Recessive	*WDR35*	Exome	Gilissen *et al.* [26]
Cerebral cortical malformations	Recessive	*WDR62*	Exome	Bilguvar *et al.* [70]
Kaposi sarcoma	Recessive	*STIM1*	Exome	Byun *et al.* [71]
Spinocerebellar ataxia	Dominant	*TGM6*	Exome	Wang *et al.* [72]
Combined hypolipidemia	Recessive	*ANGPTL3*	Exome	Musunuru *et al.* [40]
Complex I deficiency	Recessive	*ACAD9*	Exome	Haack *et al.* [52]
Autoimmune lymphoproliferative syndrome	Recessive	*FADD*	Exome	Bolze *et al.* [73]

Gilissen et al. Genome Biology 2011, 12:228

Exome sequencing is revolutionizing Mendelian disease gene identification. This results in **improved clinical diagnosis, more accurate genotype-phenotype correlations and new insights into the role of rare genomic variation in disease**

NGS based diagnostics - Targeted gene sequencing panels

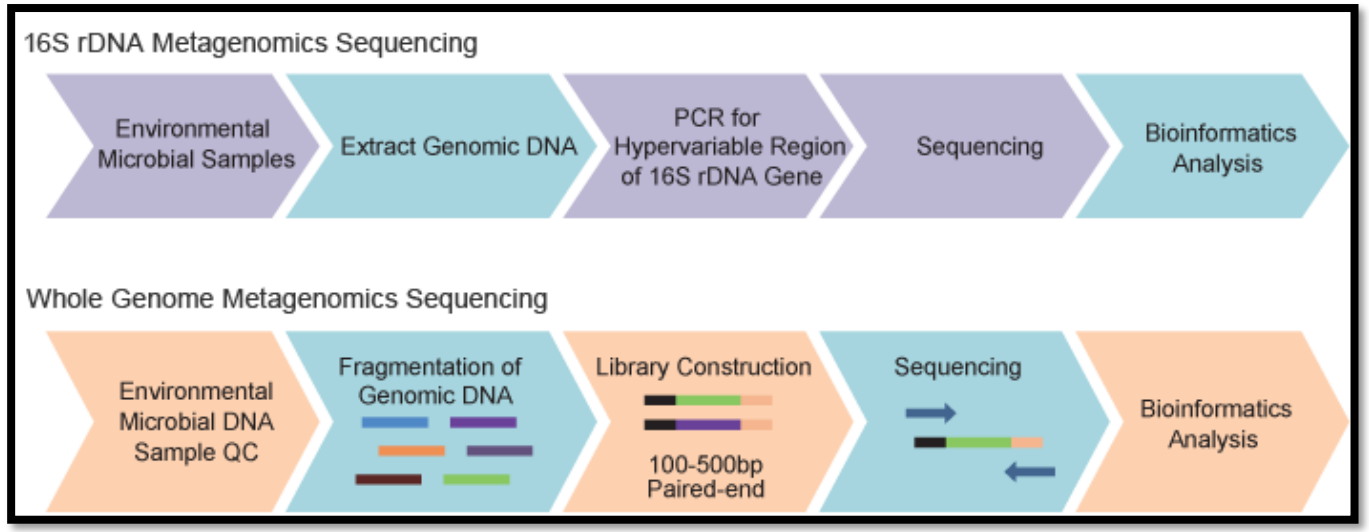
Targeted gene sequencing panels - Focused panels contain a **select set of genes** or gene regions that have known or suspected associations with the disease or phenotype under study. Gene panels can be purchased with preselected content or custom designed to include genomic regions of interest.

Multiple genes can be assessed across many samples in parallel, saving time and reducing costs associated with running multiple separate assays.

Targeted gene sequencing also produces a **smaller, more manageable data set** compared to broader approaches such as whole-genome sequencing, making analysis easier.

Metagenomics

Metagenomics (**Environmental Genomics** or **Community Genomics**) is the studyof genomes recovered from environmental samples **without** the need for **culturing** them **(cultured fraction represents only 1% of biodiversity)**



Metagenomics: DNA sequencing of environmental samples

•Susannah Green Tringe & Edward M. Rubin Nature Reviews Genetics,**6**,805–814(2005)

This technology — genomics on a huge scale — enables a survey of the different microorganisms present in a specific environment, such as water or soil, to be carried out.

By integrating the information gleaned with information about biological functions within the community, the structure of microbial communities can potentially be probed.

Helps identify massive **uncultured microbial diversity** present in the environment to provide new molecules for **therapeutic and biotechnological applications**

Microbiome project

Human microbiome project

Microbiome numbers an order of magnitude higher than total number of human cells

Many microbial interactions endow or enhance human physiology including development, nutrition, immunity and resistance to pathogens

Majority of the human microbiome is largely unknown – 250+ healthy human samples

Earth microbiome project

To systematically approach the problem of characterizing microbial life on earth

Explore microbes in environmental para space

Define microbial community structure and the protein universe

Genome Sequencing Milestones

Genome milestones

1977: Bacteriophage ΦX174 (ref. 72)
1982: Bacteriophage lambda[13]
1995: *Haemophilus influenzae*[26]
1996: *Saccharomyces cerevisiae*[27]
1998: *Caenorhabditis elegans*[28]
2000: *Drosophila melanogaster*[32]
2000: *Arabidopsis thaliana*[146]
2001: *Homo sapiens*[29–31]
2002: *Mus musculus*[147]
2004: *Rattus norvegicus*[148]
2005: *Pan troglodytes*[149]
2005: *Oryza sativa*[150]
2007: *Cyanidioschyzon merolae*[126]
2009: *Zea mays*[151]
2010: Neanderthal[88]
2012: Denisovan[145]
2013: The HeLa cell line[152,153]
2013: *Danio rerio*[154]
2017: *Xenopus laevis*[155]

ARTICLE

OPEN
doi:10.1038/nature25458

The axolotl genome and the evolution of key tissue formation regulators

Sergej Nowoshilow[1,2,3]†*, Siegfried Schloissnig[4]*, Ji-Feng Fei[5]*, Andreas Dahl[3,6], Andy W. C. Pang[7], Martin Pippel[4], Sylke Winkler[1], Alex R. Hastie[7], George Young[8], Juliana G. Roscito[1,9,10], Francisco Falcon[11], Dunja Knapp[3], Sean Powell[4], Alfredo Cruz[11], Han Cao[7], Bianca Habermann[12], Michael Hiller[1,9,10], Elly M. Tanaka[1,2,3]† & Eugene W. Myers[1,10]

REVIEW

doi:10.1038/nature24286

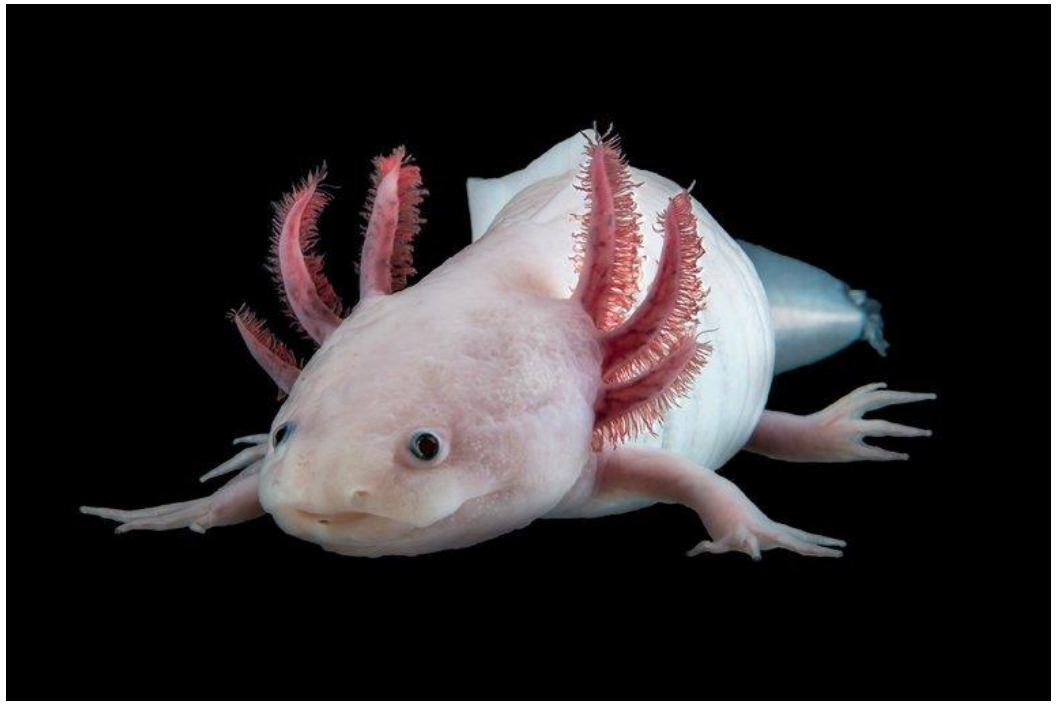
DNA sequencing at 40: past, present and future

Jay Shendure[1,2], Shankar Balasubramanian[3,4], George M. Church[5], Walter Gilbert[6], Jane Rogers[7], Jeffery A. Schloss[8] & Robert H. Waterston[1]

This review commemorates the 40th anniversary of DNA sequencing, a period in which we have already witnessed multiple technological revolutions and a growth in scale from a few kilobases to the first human genome, and now to millions of human and a myriad of other genomes. DNA sequencing has been extensively and creatively repurposed, including as a 'counter' for a vast range of molecular phenomena. We predict that in the long view of history, the impact of DNA sequencing will be on a par with that of the microscope.

http://www.nature.com/doifinder/10.1038/nature24286

Largest genome ever sequenced



Scientists have decoded the genome of the **axolotl**, the Mexican amphibian.

Sergej Nowoshilow. Nature, 554, 50–55(2018)

It has **32 billion base pairs**, which makes it **ten times the size of the human genome**

Desired skill set

NGS Data is typically Big Data and requires computational and data analytics skills

Perl/ Python (Python is preferred)

R / Matlab

Biostatistics

Awk

Picard, SAMtools, Bedtools, Bismark(for BS Seq data)

Familiarity with Unix and Linux and High Performance Computing

Google ‘vipin’s classroom’

Next Webinar - 'Principles of Next Generation Sequencing Techniques and applications' at ICGEB, Delhi, 24.7.2020

Vipin's e-Classroom

About me

Vipin's Webinar

Vipin's Crackitts ...

Student mails

AIB Students' Seminar Series

Vipin's Soft Skills classes

Vipin's GATE TO NET

Vipin's Study Material

Vipin's NET

Dr. Vipin Singh

Associate Professor,

Coordinator - Corporate Resource Centre,

Coordinator Admissions

University Institute of Biotechnology,

Chandigarh University, Mohali - June 2017

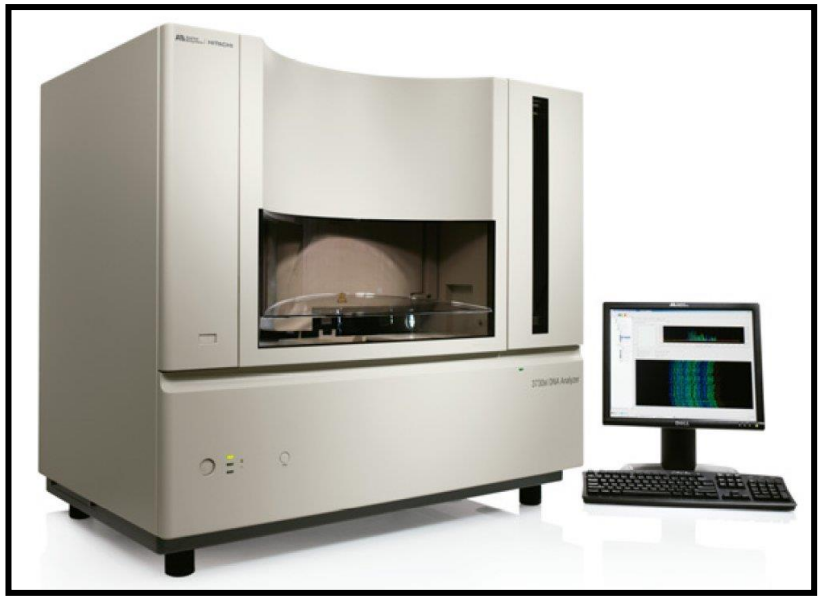
- "It is in a 'class' that magic truly happens - where 'nobodies' transform into 'somebodies'. In that sense teachers have immense power to transform an individual, the society and the Nation." Vipin Singh.

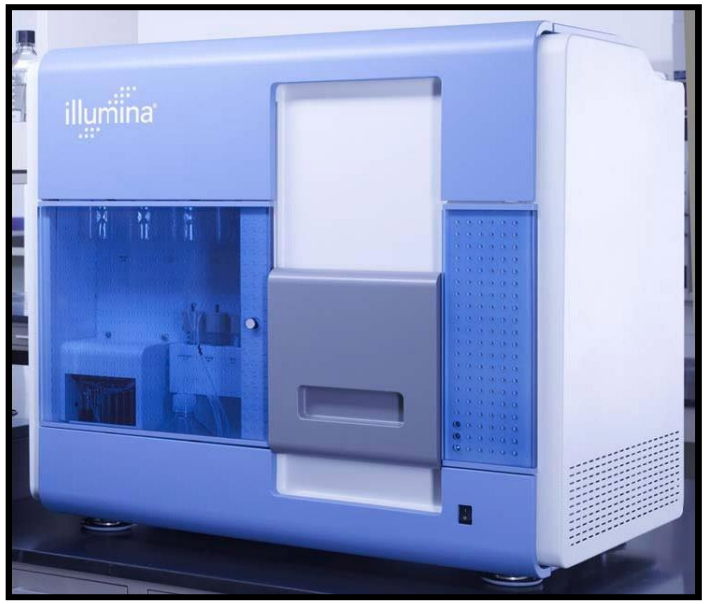
Educational Background

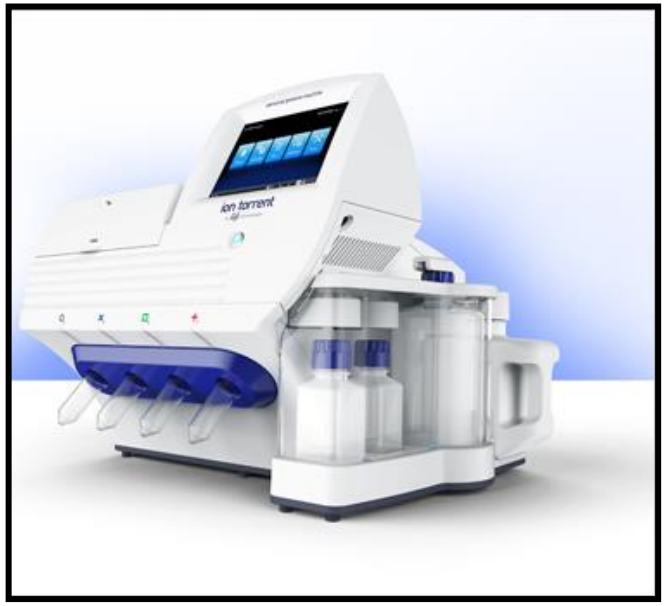
CSIR-UGC-NET- JRF, 2002,

Ph.D. – Life Sciences, CSIR-CCMB, Hyderabad - Thesis title - ‘Genomic alterations: Sequence changes associated with repeats'

Post Doctoral Fellow - Institute of Biology, Ecole Normal Superior, Paris (April 2019-April 2020)







Thank you !!!