

Talleres Internacionales de Bioinformática (TIB)

Cuernavaca, 2017

<http://congresos.nnb.unam.mx/TIB2017/>

Analysis of high-throughput sequencing data using Galaxy (ChIP-seq and RNA-seq)

Denis Puthier, Claire Rioualen & Jacques van Helden
Aix-Marseille Univ, INSERM, TAGC lab, Marseille, France

Goals of the workshop

- Target audience
 - Biologists involved in NGS projects.
 - No prior experience of NGS bioinformatics.
- Approach
 - Practice-driven.
 - Elements of theory interspersed in the tutorials.
- Scope
 - Study cases from ChIP-seq and RNA-seq.
 - However many concepts and tools are also used by many other applications.
- Software environment
 - Mainly Galaxy
 - Visualisation with IGV
 - Web sites for specific resources.
 - R under RStudio convivial environment? To be discussed ...

Schedule

- **Days 1 - 2: ChIP-seq analysis**
 - NGS Technologies
 - ChIP-Seq analysis - Intro
 - Short read file formats
 - Quality control of the reads
 - Trimming
 - Read mapping
 - Data visualization (IGV)
 - Coverage normalisation
 - Peak calling
 - Peak annotation
 - Motif analysis
- **Days 3-4: RNA-seq**
 - RNA-Seq method intro
 - Preprocessing
(Quality control, Trimming)
 - Splice-aware alignment
 - Transcript discovery
 - Data visualization
 - Quantification
 - Differential analysis
 - Functional annotation
 - Motif analysis (continued)
- **Day 5: tutorship and/or R ?**
 - Customized analytic flow charts
+ playing with your own data.
 - Optional: first steps with R.

Presentation of the teachers

- **Denis Puthier**
 - Bioinformatics analysis of high-throughput data.
 - Teaching domains: bioinformatics, genomics, programming, statistics.
- **Claire Rioualen**
 - Bioinformatics analysis of high-throughput data.
 - Development of workflows for NGS data (ChIP-seq, RNA-seq).
- **Jacques van Helden**
 - Development of bioinformatics tools for the analysis of regulatory sequences and networks (<http://rsat.eu/>).
 - Teaching domains: bioinformatics, statistics, genomics.

Presentation of the participants

Participants introduce themselves in 4 sentences.

1. Name and affiliation
2. Background in biology/bioinformatics
3. Research project involving NGS / interest for NGS.
4. Prior experience with NGS bioinformatics?

Resources used during the training

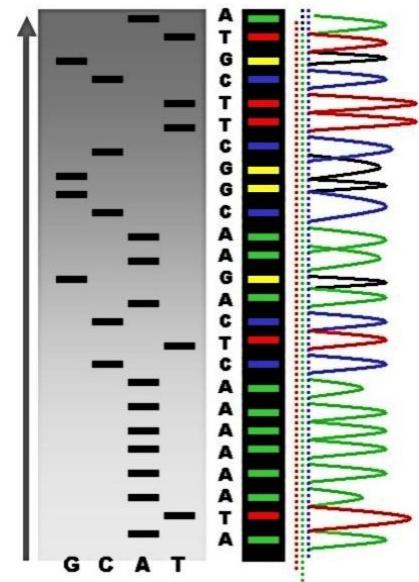
Resource	Description + URL	To install locally
TIB 2017 homepage	http://congresos.nnb.unam.mx/	
TIB 2017 Galaxy	http://congresos.nnb.unam.mx/TIB2017/galaxy/	
Galaxy server	Galaxy server for the TIB2017 training http://132.248.220.36/	
IGV	Integrative Genomics Viewer http://software.broadinstitute.org/software/igv/	X
R	R statistical package https://www.r-project.org/	X
RStudio	An environment to manage R programming and projects https://www.rstudio.com/	X
GEO	Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/geo/	
ArrayExpress	Gene expression database https://www.ebi.ac.uk/arrayexpress/	
RSAT	Regulatory Sequence Analysis Tools http://rsat.eu/	



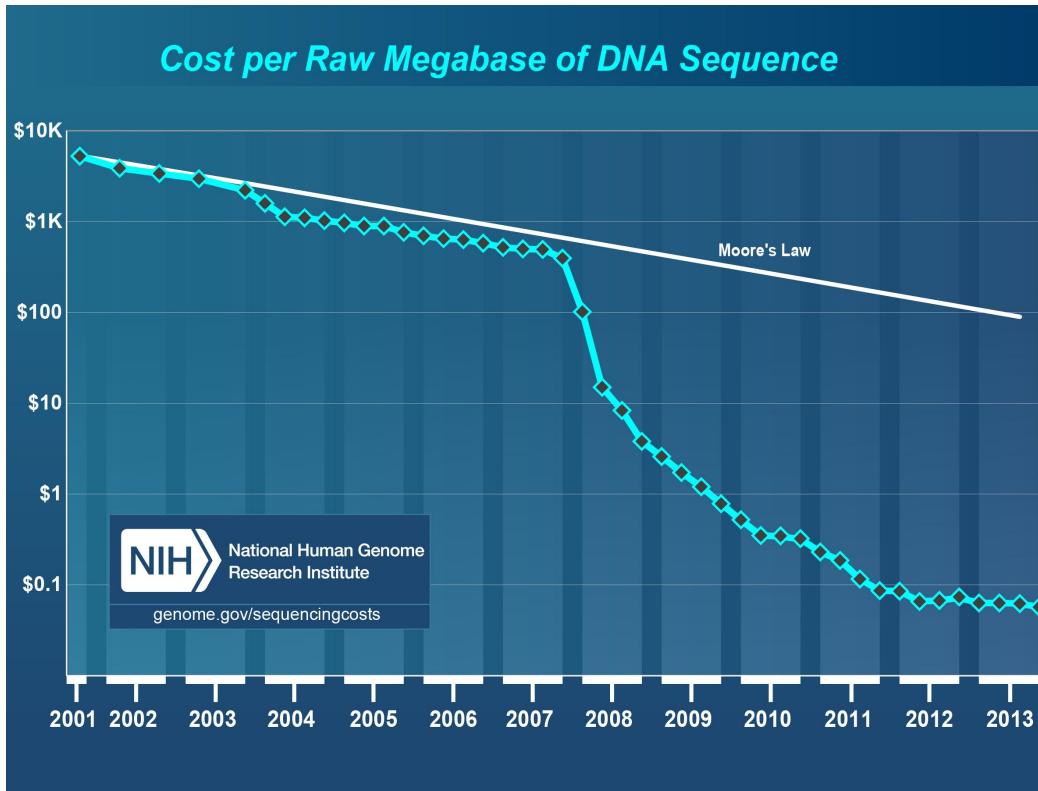
High-throughput sequencing

Breakthrough in DNA Sequencing

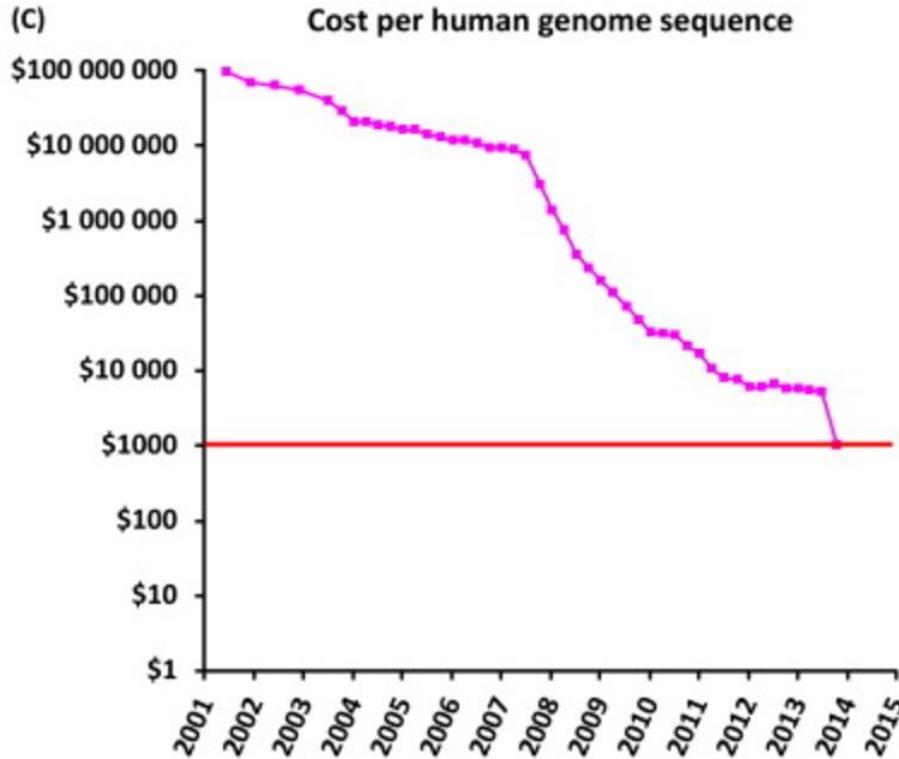
- 1977-1990, 500bp, manual analysis
- 1990-2000, 500bp, computer assisted analysis
(1D capillary sequencers)
- 2005-2014, 20-1000bp
(2D sequencers “Next Generation Sequencing.”)



Cost per megabase (1 million base)



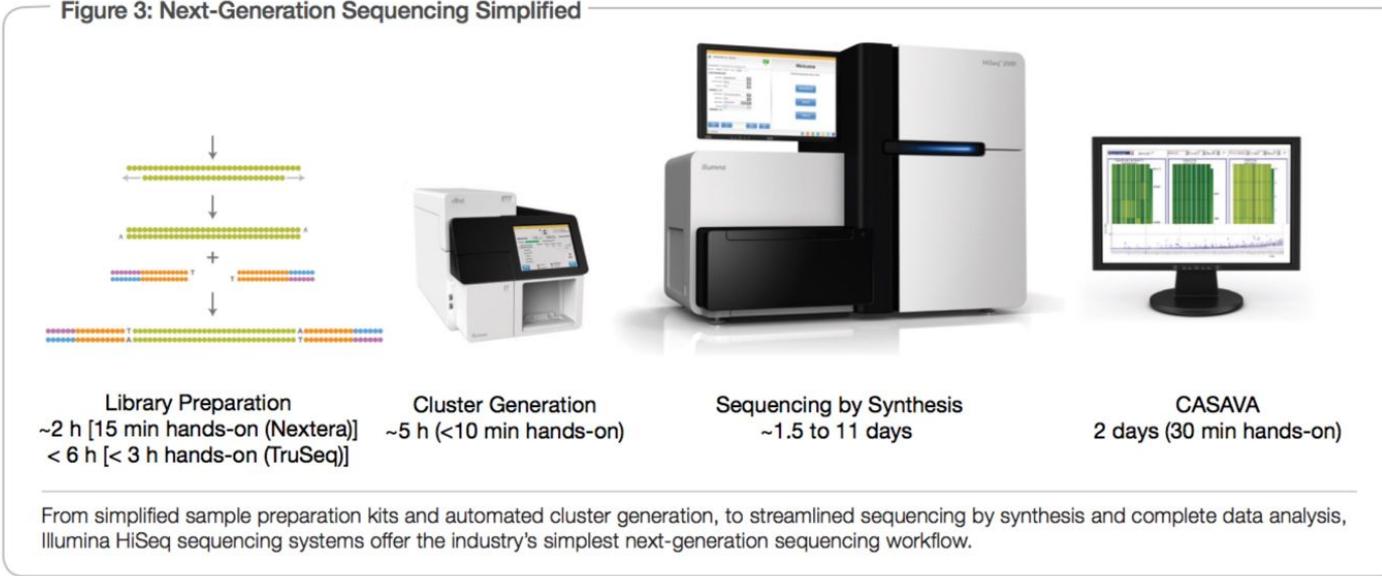
Cost per human genome



van Dijk,E.L., Auger,H., Jaszczyzyn,Y. and Thermes,C. (2014)
Ten years of next-generation sequencing technology. *Trends Genet*, **30**, 418–426.

NGS: a simplified view

Figure 3: Next-Generation Sequencing Simplified

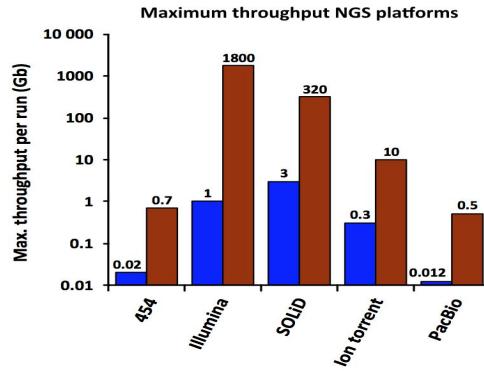


NB: most of the methods rely on fragmented DNA/RNA material.

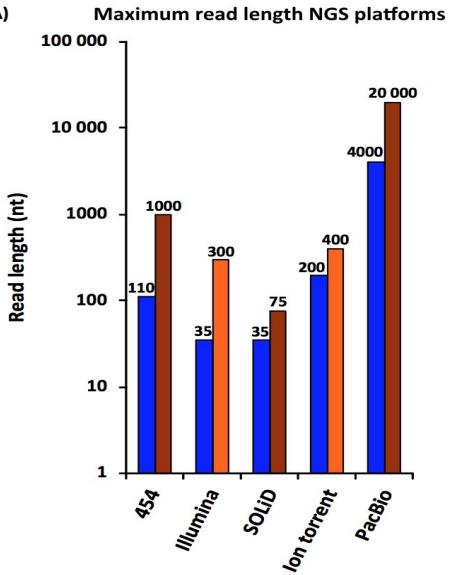
Important things to consider

- Sequencer throughput
 - Some application require good coverage
 - High dynamic range, sensibility
 - e.g transcriptome analysis, ChIP-Seq
 - May offer multiplexing
- Read length produced
 - May be important to resolve low complexity regions
 - i.e. a word of size 20 is more ambiguous than a word of size 500

(B)



(A)



Important things to consider (continued)

- Fidelity
 - Some sequencer may be error prone
 - Fidelity may be important for variant calling (...)
- With current technologies:
 - The longer the reads (*i.e* several kbs) the weaker the fidelity and coverage

Sequencing is continuously evolving

- Technologies are subject to rapid changes!
- From this 2011 table, only a few survived in 2016.

Table 1 2nd and 3rd Generation DNA sequencing platforms listed in the order of commercial availability

Platform	Current company	Former company	Sequencing method	Amplification method	Claim to fame	Primary applications
454	Roche	454	Synthesis (pyrosequencing)	emPCR	First Next-Gen Sequencer, Long reads	1* , 2 , 3* , 4 , 7 , 8*
Illumina	Illumina	Solexa	Synthesis	BridgePCR	First short-read sequencer; current leader in advantages†	1* , 2 , 3* , 4 , 5 , 6 , 7 , 8
SOLiD	Life Technologies	Applied Biosystems	Ligation	emPCR	Second short-read sequencer; low error rates	3* , 5 , 6 , 8
HeliScope	Helicos	N/A	Synthesis	None	First single-molecule sequencer	5 , 8
Ion Torrent	Life Technologies	Ion Torrent	Synthesis (H ⁺ detection)	emPCR	First Post-light sequencer; first system <\$100 000	1 , 2 , 3 , 4 , 8
PacBio	Pacific Biosciences	N/A	Synthesis	None	First real-time single-molecule sequencing	1 , 2 , 3 , 7 , 8
Starlight‡	Life Technologies	N/A	Synthesis	None	Single-molecule sequencing with quantum dots	1 , 2 , 7 , 8

Bold indicates applications that are most often used, economical or growing.

1 = *de novo* BACs, plastids, microbial genomes.

2 = transcriptome characterization.

3 = targeted re-sequencing.

4 = *de novo* plant and animal genomes.

5 = re-sequencing and transcript counting.

6 = mutation detection.

7 = metagenomics.

8 = other (ChIP-Seq, mRNA-Seq, Methyl-Seq, etc.; see Brautigam & Gowik 2010, Shendure & Ji 2008).

Field guide to next-generation DNA sequencers

TRAVIS C. GLENN

Department of Environmental Health Science and Georgia Genomics Facility, Environmental Health Science Building, University of Georgia, Athens, GA 30602, USA

Illumina sequencers



Illumina sequencers



	MiniSeq System	MiSeq Series +	NextSeq Series +
Popular Applications & Methods	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)			●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●
Exome Sequencing			●
Targeted Gene Sequencing (amplicon, gene panel)	●	●	●
Whole-Transcriptome Sequencing			●
Gene Expression Profiling with mRNA-Seq			●
Targeted Gene Expression Profiling	●	●	
miRNA & Small RNA Analysis	●	●	●
DNA-Protein Interaction Analysis		●	●
Methylation Sequencing			●
16S Metagenomic Sequencing		●	●
Run Time	4–24 hours	4–55 hours	12–30 hours
Maximum Output	7.5 Gb	15 Gb	120 Gb
Maximum Reads Per Run	25 million	25 million*	400 million
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp

NextSeq 500 Illumina sequencer

Key Methods	Everyday genome, exome, transcriptome sequencing, and more.		Production-scale genome, exome, transcriptome sequencing, and more.				
	 NextSeq 500	 HiSeq 2500	 HiSeq 3000	 HiSeq 4000			
Run Mode	Mid-Output	High-Output	Rapid Run	High-Output	N/A	N/A	
Flow Cells per Run	1	1	1 or 2	1 or 2	1	1 or 2	
Output Range	20-39 Gb	30-120 Gb	10-300 Gb	50-1000 Gb	125-750 Gb	125-1500 Gb	
Run Time	15-26 hours	12-30 hours	7-60 hours	<1-6 days	<1-3.5 days	<1-3.5 days	
Reads per Flow Cell†	130 million	400 million	300 million	2 billion	2.5 billion	2.5 billion	
Maximum Read Length	2 x 150 bp	2 x 150 bp	2 x 250 bp	2 x 125 bp	2 x 150 bp	2 x 150 bp	
System Overview	Speed and simplicity for everyday genomics.		Power and efficiency for large-scale genomics.		Maximum throughput and lowest cost for production-scale genomics.	Maximum throughput and lowest cost for production-scale genomics.	



HiSeq X 10: a factory scale sequencer

Population power. Extreme throughput. \$1,000 human genome.

The HiSeq X Ten is a set of ten ultra-high-throughput sequencers, purpose-built for large-scale human whole-genome sequencing.



Population Scale Studies

Learn how the HiSeq X Ten can benefit communities by enabling them to sequence their entire population.
[Read blog post »](#)

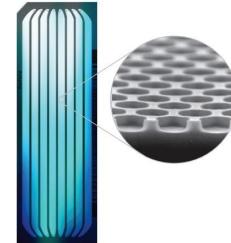


Figure 1: Patterned Flow Cell Design—Patterned flow cells contain billions of nanowells at fixed locations, providing even cluster spacing and uniform feature size to deliver extremely high cluster densities.

- Illumina
- A set of 10 sequencers.
 - Each producing 1,8 Terabases / 3 days
- 18,000 genome / year
 - Factory-scale sequencing technology

Some computing issues

<http://glennklockwood.blogspot.nl/>

- 18,000 / year ~ 340 / week
- 30-50 Tb storage / week
 - Cost of long term storage?
- 518 core hours / genome
- 175,000 core hours per week

But 1000\$ genome coming true....

Is the 1000 \$ genome for real ?



Home Clinical Research Ancestry Carrier Screening Company

My Cart Login

What's in my genes

Whole Exome
Sequencing

Illumina OMNI Express
GWAS Studies

mtDNA
Sequencing

Whole Genome
Sequencing

Whole Genome Sequencing

Description

Average 30X Coverage

Gene By Gene's whole genome sequencing service allows for a high degree of accuracy in identifying variants across the entire scope of the human genome.

Gene By Gene is excited to announce the implementation of the Arpeggi engine pipeline. Arpeggi, Inc., now part of Gene By Gene, developed internal bioinformatics software for alignment and variant calling of next generation sequencing (NGS) data that is now exclusive to Gene By Gene. If you want us to complete the alignment, mapping, and variant calling on your exome sequencing, we will show you a full report comparing different pipeline options and why the Arpeggi engine delivers you the best possible VCF file so you have the highest quality data for your project.

Results are delivered to the customer via electronic FTP transfer and are only stored by Gene By Gene for 30-60 days.

Quantity

1

Analysis

None

Alignment and Variant Calling (\$400)

Price

\$9,995.00

(\$9,995.00 per sample)

Add to Cart

Genetic variation ongoing project

U.S. proposes effort to analyze DNA from 1 million people

WASHINGTON | BY TONI CLARKE AND SHARON BEGLEY



Human Longevity Inc. (HLI) Launched to Promote Healthy Aging Using Advances in Genomics and Stem Cell Therapies

HLI is Building World's Largest Genotype/Phenotype Database by Sequencing up to 40,000 Human Genomes/Year Combined with Microbiome, Metabolome and Clinical Data to Develop Life Enhancing Therapies



HLI has Purchased Two Illumina HiSeq X Ten Sequencing Systems

SAN DIEGO, CA (March 4, 2014) — Illumina's Jay Flatley at #PMWC14: Get Sequence of 1 million cancer patients in next 5 years

January 27, 2014 by nextgenseek • 1 Comment



Illumina's Jay Flatley said at #PMWC14 that Illumina wants to have cancer patients in a database in the next five years. And one of his cancer a "chronic" disease within 10 years. Jay Flatley said Illumina will work with researchers and clinicians to develop large population genomic datasets with researchers and clinicians talk at #PMWC14 happening right now at Mountain View, CA.

Thanks to awesome live tweets by [Kevin Davies](#), [@DivaBioTech](#), and links to the original tweets.



diseases associated with aging leading the development of c... decline in endogenous stem ce...



THE IRAN OPPORTUNITY By Farhad Zakeri / E-Cigarettes / \$20K Homes

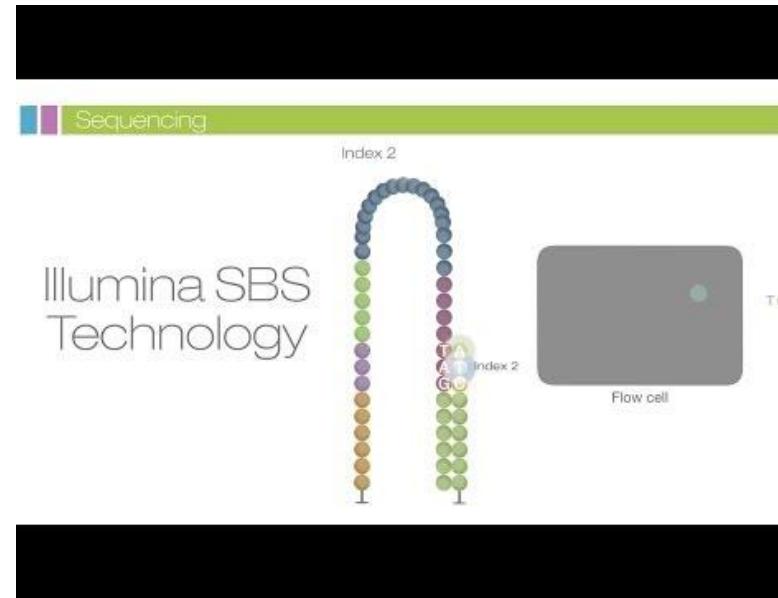
TIME
CAN
Google
SOLVE
DEATH?

The search giant is launching a venture to extend the human life span.
That would be crazy—if it weren't Google
By Harry McCracken and Lev Grossman



An overview of Illumina technology - sequencing by synthesis

Illumina sequencing: general principle

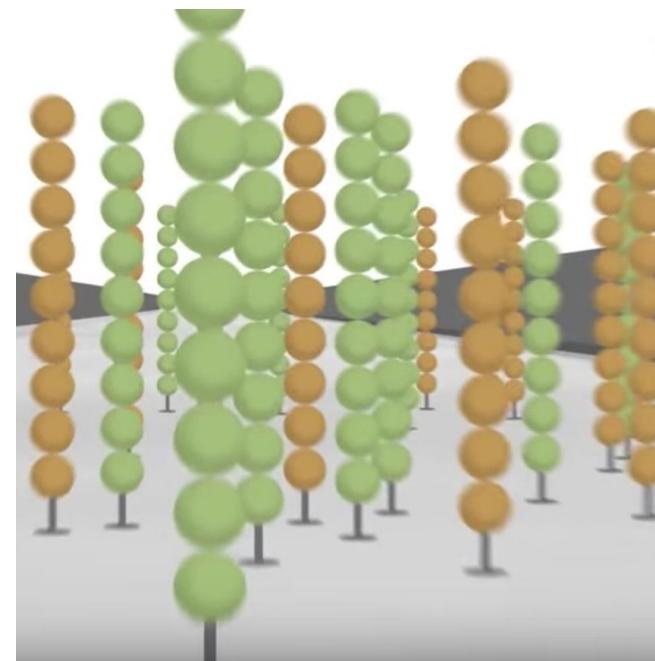


<http://www.illumina.com/company/video-hub/HMyCqWhwB8E.html>

Illumina sequencing: general principle

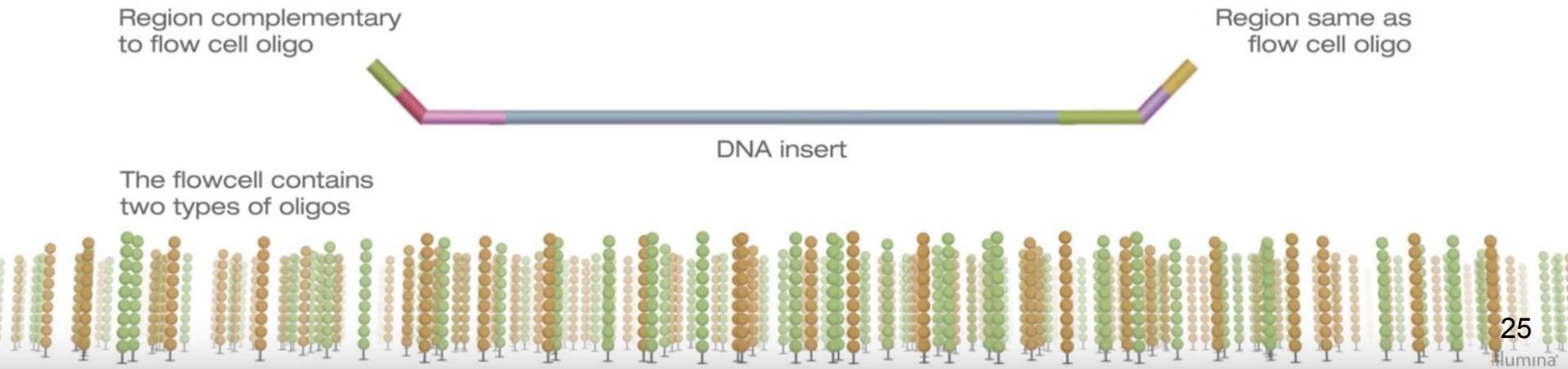


Flow cell



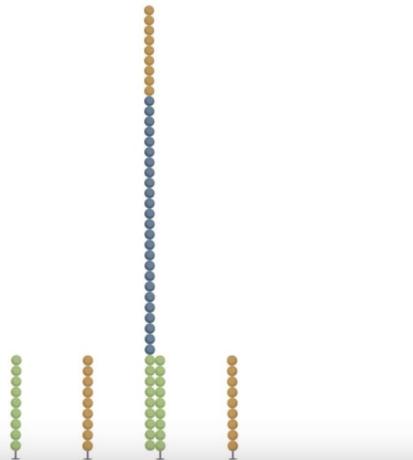
Starting with a fragment

- Terminology:
 - “Fragment” a piece of DNA
 - “Read” the sequence(s) associated to this fragment

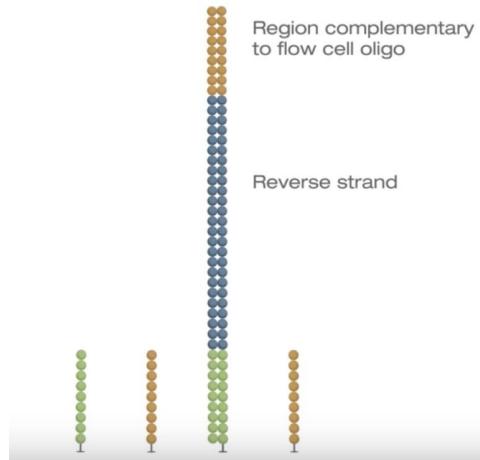


First-strand synthesis

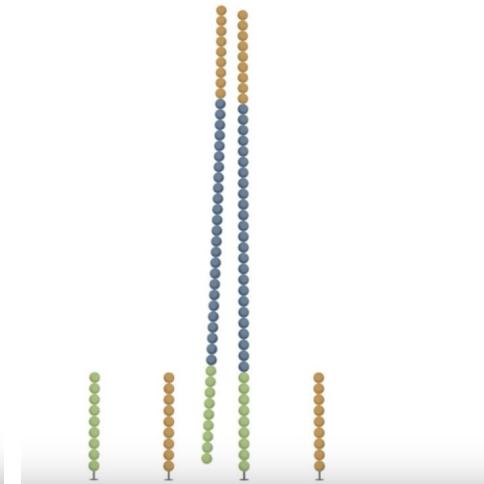
Annealing



Reverse-strand synthesis

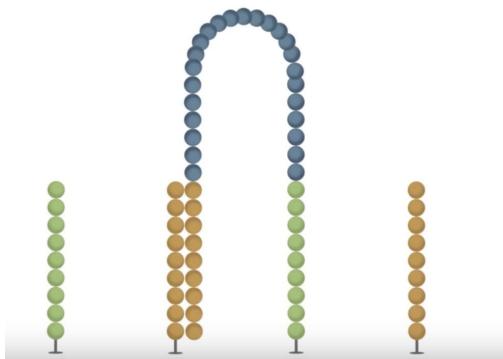


Denaturation
Fragment released

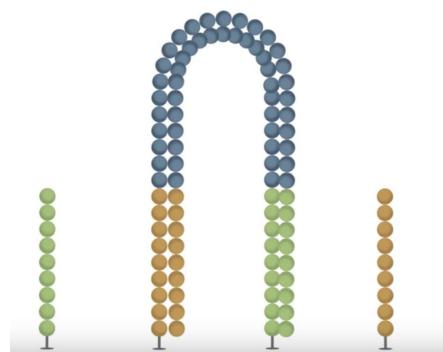


Bridge-PCR

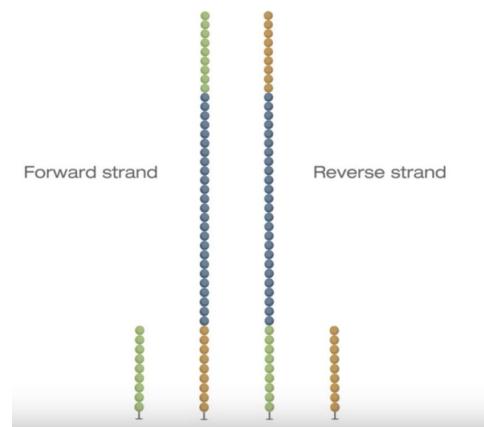
Annealing



Reverse-strand
synthesis

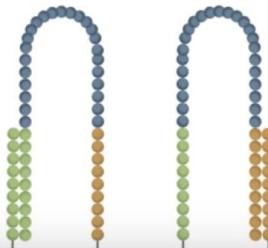


Denaturation

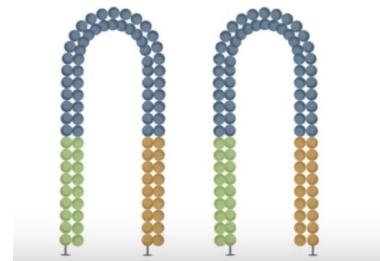


Bridge-PCR cycles

Annealing

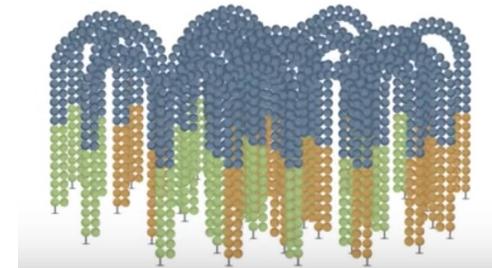


2-copies

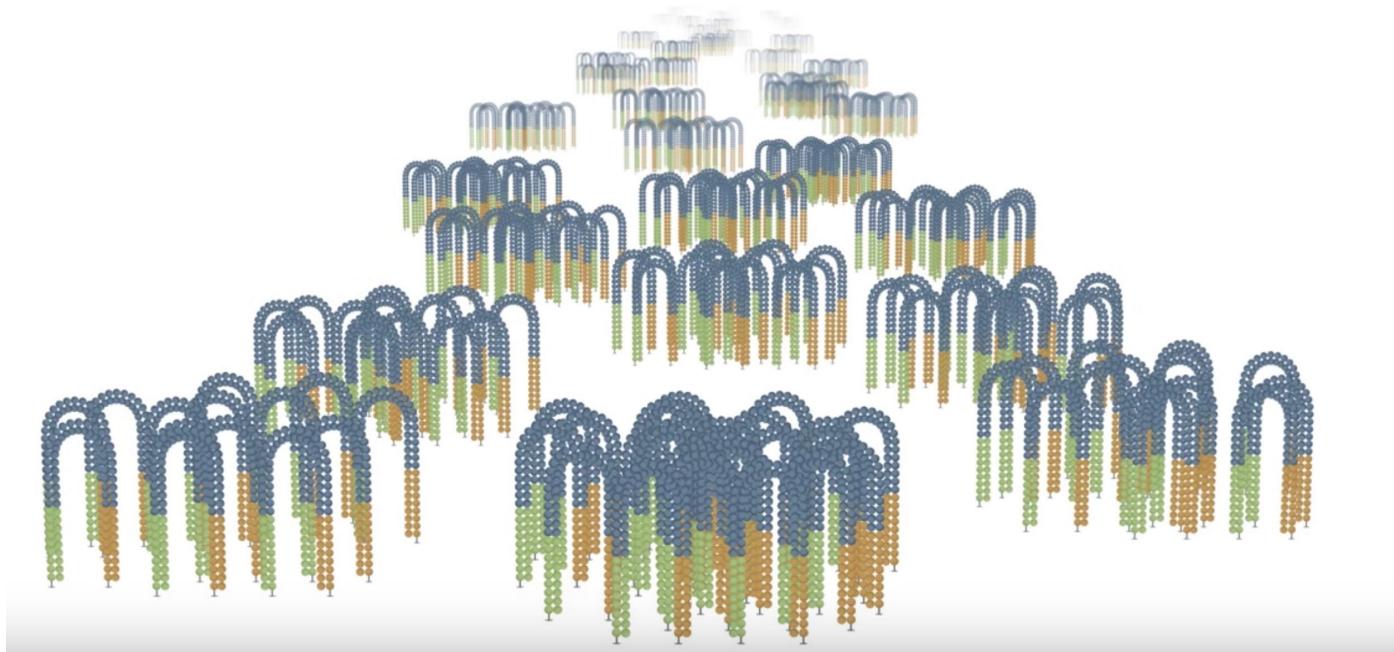


N-copies

- Cluster
- Polymerase colonies
(Polonies)

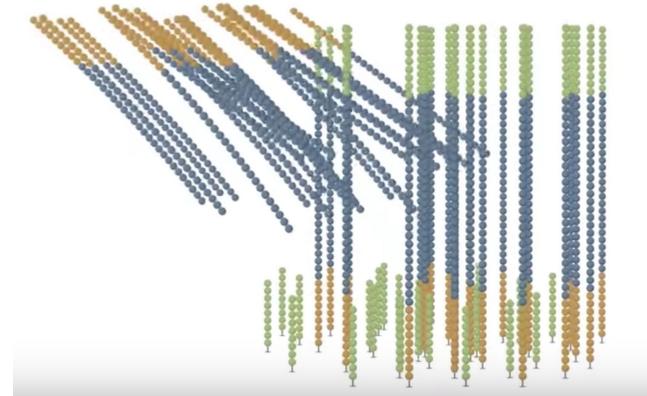


A population of DNA colonies



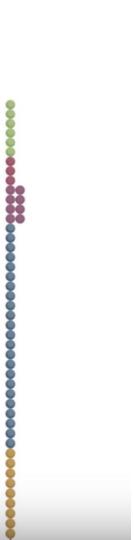
Getting single-stranded colonies

Reverse strand cleavage



First-end sequencing

Primer
annealing

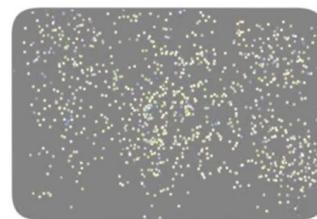


Synthesis/extension
Record color at each step



Flow cell

This is a parallelized
process !

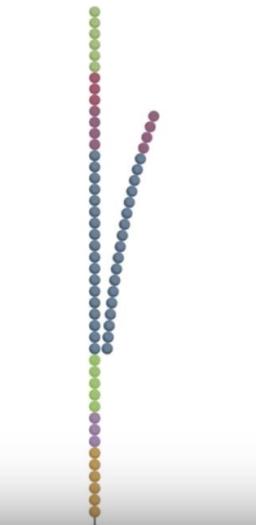


Flow cell

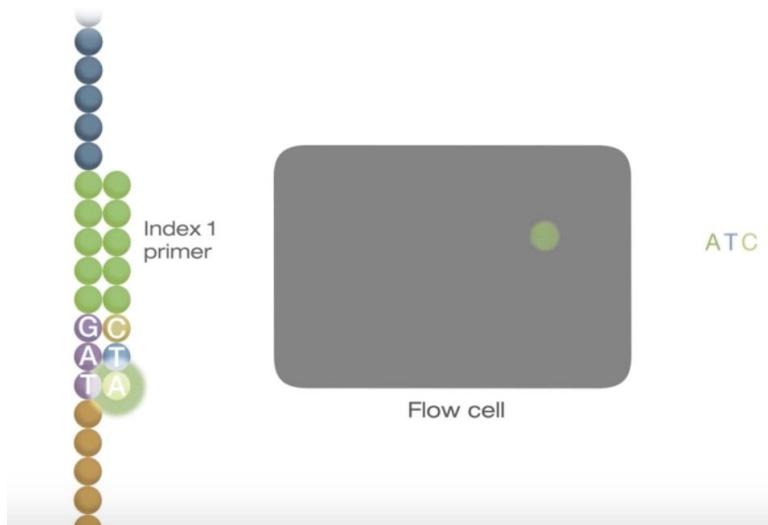
ACAAAAGCAATTGACAAAC
ACGCCGTACTACCTCAGCA
AAGAAACAAAAGCAATTGA
CCGTACTACCTCAGCAGTA
CAGCAGTAGTAAGAAACAA
ACAAAAGCAATTGACAAAC
ACGCCGTACTACCTCAGCA
AAGAAACAAAAGCAATTGA
CCGTACTACCTCAGCAGTA
CAGCAGTAGTAAGAAACAA
ACAAAAGCAATTGACAAAC
ACGCCGTACTACCTCAGCA
AAGAAACAAAAGCAATTGA
CCGTACTACCTCAGCAGTA
CAGCAGTAGTAAGAAACAA
ACAAAAGCAATTGACAAAC

Barcode analysis

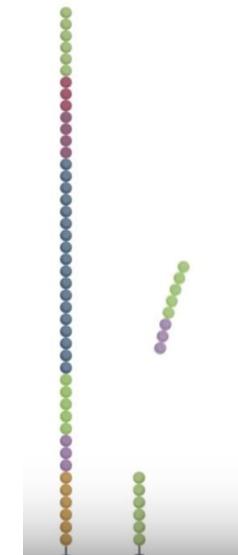
Release of the read



Read the barcode
(for subsequent de-multiplexing)

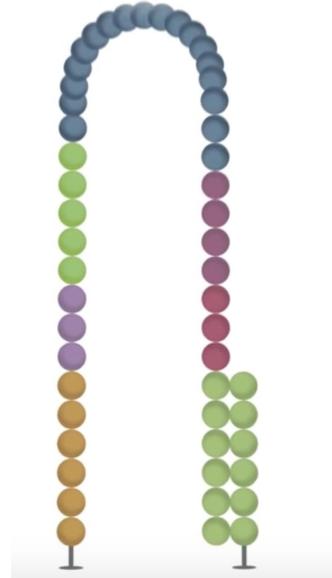


Release of the read

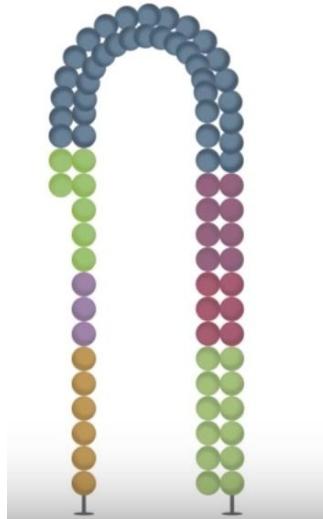


Paired-end sequencing

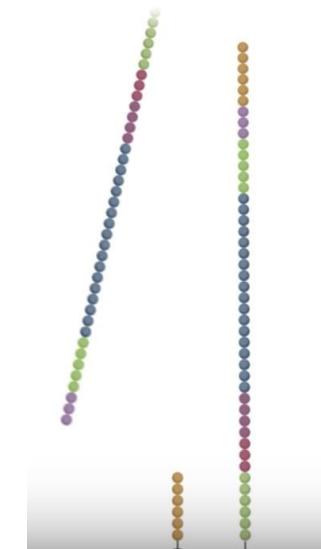
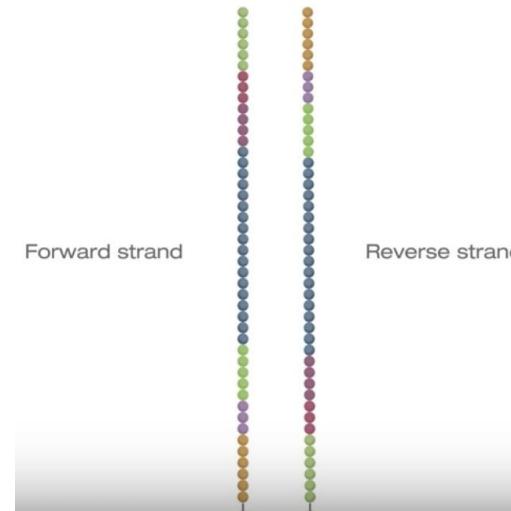
Bridged-annealing



Reverse-strand synthesis

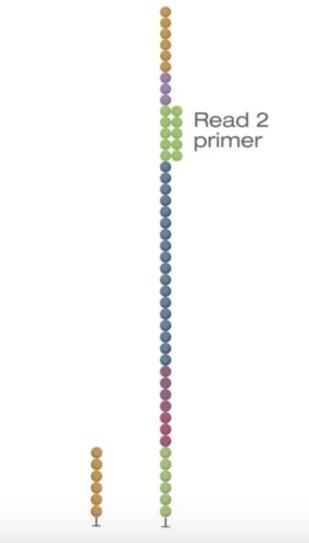


Cleavage of the forward strand

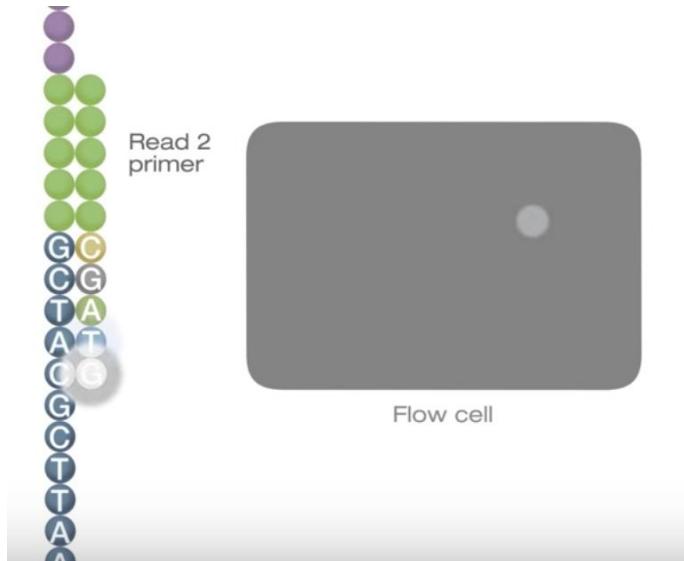


Paired-end sequencing

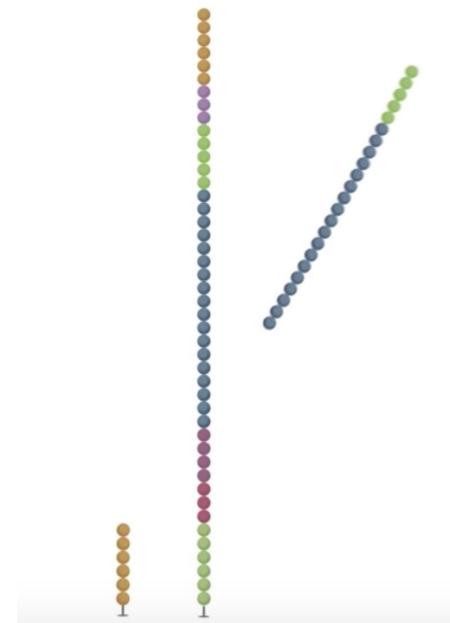
Bridged-annealing



Parallel sequencing



Denaturation



Single-end vs paired-end

- Paired-end sequencing: sequence both ends of a fragment
 - Facilitate alignment
 - Facilitate gene fusion det
 - Better to reconstruct transcript model from RNA-seq





Other technologies

Example application of MinION

Real-time, portable genome sequencing for Ebola surveillance.

Nature doi:10.1038/nature16996

[View it >](#)

Joshua Quick, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Stephan Günther, Miles W. Carroll *et al*

Abstract

The Ebola virus disease epidemic in West Africa is the largest on record, responsible for over 28,599 cases and more than 11,299 deaths. Genome sequencing in viral outbreaks is desirable to characterize the infectious agent and determine its evolutionary rate. Genome sequencing also allows the identification of signatures of host adaptation, identification and monitoring of diagnostic targets, and characterization of responses to vaccines and treatments. The Ebola virus (EBOV) genome substitution rate in the Makona strain has been estimated at between 0.87×10^{-3} and 1.42×10^{-3} mutations per site per year. This is equivalent to 16–27 mutations in each genome, meaning that sequences diverge rapidly enough to identify distinct sub-lineages during a prolonged epidemic. Genome sequencing provides a high-resolution view of pathogen evolution and is increasingly sought after for outbreak surveillance. Sequence data may be used to guide control measures, but only if the results are generated quickly enough to inform interventions. Genomic surveillance during the epidemic has been sporadic owing to a lack of local sequencing capacity coupled with practical difficulties transporting samples to remote sequencing facilities. To address this problem, here we devise a genomic surveillance system that utilizes a novel nanopore DNA sequencing instrument. In April 2015 this system was transported in standard airline luggage to Guinea and used for real-time genomic surveillance of the ongoing epidemic. We present sequence data and analysis of 142 EBOV samples collected during the period March to October 2015. We were able to generate results less than 24 h after receiving an Ebola-positive sample, with the sequencing process taking as little as 15–60 min. We show that real-time genomic surveillance is possible in resource-limited settings and can be established rapidly to monitor outbreaks.

And now the Smidgion...



SmidgION: nanopore sensing for use with mobile devices

Using the same core technology as the handheld MinION device, we are now starting to develop an even smaller device.

In early development

The MinION portable sequencer



Long read lengths

The Oxford Nanopore system processes the reads that are presented to it rather than generating specific read lengths. The longest read reported by a MinION user to date is more than 200Kb, but it can process the spectrum of read lengths.

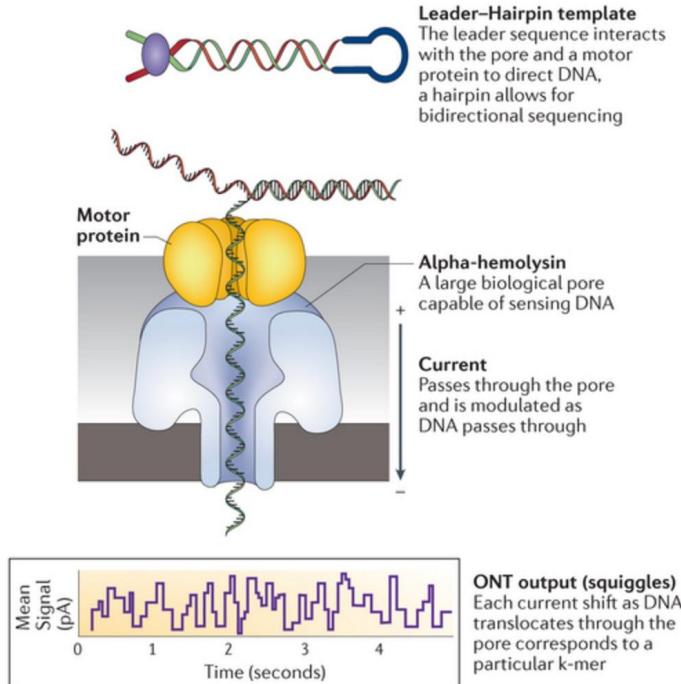
Long read lengths

Real-time data

Direct molecular analysis

Portability

Ab Oxford Nanopore Technologies



- Alpha-hemolysine
 - A nanopore from bacteria that causes lysis of red blood cells
- Molecules that enter the nanopore cause characteristic disruption of the current.
- Potentially offers read lengths of tens of kilobases (kb) limited only by the length of DNA molecules presented to it.”
- ~1Gb to 2 Gb of sequence per minION.
- Detect DNA modifications.

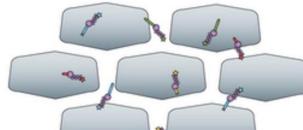
Single-molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio).

Aa Pacific Biosciences

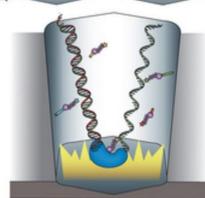
SMRTbell template
Two hairpin adapters allow continuous circular sequencing



ZMW wells
Sites where sequencing takes place



Labelled nucleotides
All four dNTPs are labelled and available for incorporation



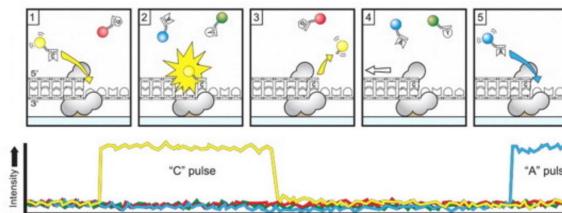
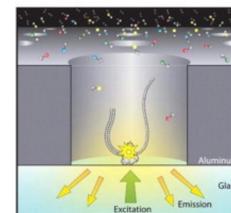
Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light

PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base



- Zero-mode waveguides (ZMW)

- Each ZMW well is several nanometres in diameter
- The size of each well does not allow for light propagation
- The fluorophores bound to bases can only be visualized through the glass substrate in the bottom-most portion of the well, a volume in the zeptolitre range
- Polymerase is fixed to the bottom of the well.
- dNTP incorporation on each single-molecule template per well is continuously analyzed by a laser and
- The polymerase cleaves the dNTP-bound fluorophore during incorporation, allowing it to diffuse away
- High error-rate, high cost per base



Coming of age: ten years of next-generation sequencing technologies 40

Sara Goodwin¹, John D. McPherson² and W. Richard McCombie¹

Review

PacBio Sequencing and Its Applications

Anthony Rhoads^{1, a}, Kin Fai Au^{1, 2}, , , 



Applications of high-throughput sequencing

High-throughput sequencing: so much applications...



Sequencing Methods Review

A review of publications featuring Illumina® Technology

<http://tinyurl.com/znrb9jc>



Merci