

# Starting up a long read sequencing facility

\* Javier Temes, \* Daniel Garcia Souto, \* Jorge Rodríguez, Martín Santamarina, Andrea Lema & Jose MC Tubio  
Genomes & Disease - CIMUS - Universidade de Santiago de Compostela

## Abstract

1) After a decade of absolute domain of short read NGS sequencing, there are still unsolved questions about the structure and organization of the DNA that the new generation of long read sequencers are helping to address.

2) Perhaps one of the most astonishing technologies, both by its concept and accessibility is Oxford Nanopore Technologies (ONT) sequencing. This platform is based on disposable flowcells constituted by an artificial bilipidic membrane containing multiple nano-scale pores. These pores allow the flow of either DNA or RNA while registering the subsequent ionic current. The measurement of the subtle changes in this ionic current allows deciphering the molecule sequence at nucleotide level, generating long-read data.

3) In this work, we explain the process of establishing an ONT MinION facility at CIMUS.

4) Since the foundation of this facility, we've overcome four big challenges: The constant evolution and improvement of this technology affecting both chemistry and bioinformatic pipelines, achieving a suitable DNA isolation protocol for long-read sequencing, improving the already per se delicate library construction and dealing with bottlenecks on processing and storing the massive amount of data generated in each run.

5) To this day our group has undertaken the sequencing of more than a hundred DNA samples from a myriad of species, such as healthy and tumor human tissues and cell lines, dog, bivalve, yeast, Escherichia coli or lambda phage. Additionally, this technology has proven to be extremely versatile for the massive sequencing of high molecular weight PCR barcoded fragments.

Among our greatest achievements, this technology has allowed us the de-novo assembly of a non model animal (Cerastoderma edule), the validation of retraspersions insertions and movilizacions in cancer, the study of structural variations, otherwise unachievable with short reads, and the massive and simultaneous sequencing of hundreds of mitochondrial DNA amplicons to establish molecular phylogenies. Nowadays, we routinely obtain from 10 to 25 Gbases per each run, with average fragment lengths up to 20 K bases and some reads up to 400 Kbases.

## MinION (Oxford Nanopore Technologies)

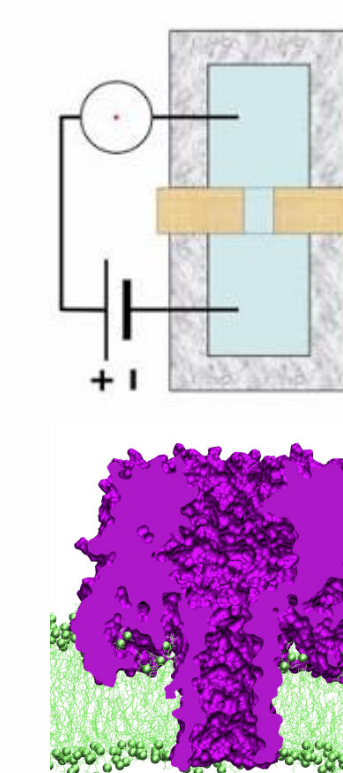
Based on the Coulter counter and ion channels. It detects an ionic current across a nanometer channel (nanopore) each time a molecule transverses it. Amongst the most widely employed nanopores are the  $\alpha$ -Hemolysin, Mycobacterium smegmatis porin (MspA), Bacteriophage phi29 and solid-state nanopores.

In 2014 Oxford Nanopore Technologies developed the first nanopore based commercial sequencer: the MinION. This small device plugs into a computer using a USB 3.0 cable, using flowcells with 2048 nanopores allowing a maximum of 512 simultaneous reads.

Using the MinION is extremely easy. Basically mixing the DNA or RNA sample with the kit reagents ONT provides, and putting it in the sequencing chamber of the MINION, connecting it to a computer and start the real time sequencing.

The reagents contains molecular proteins motors that once ligated to the DNA molecules, attach to a nanopore, unfold the DNA molecule and insert them into the nanopore one nucleotide at a time (450 Nucleotides/sec).

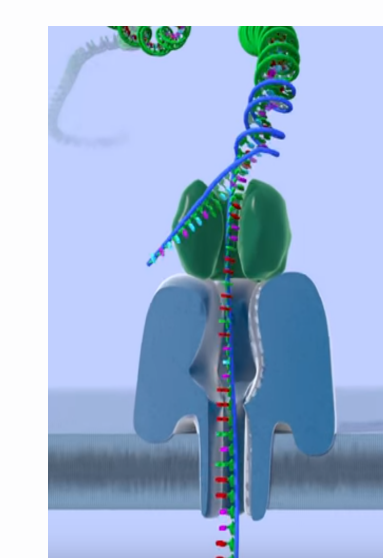
## 2 Technology



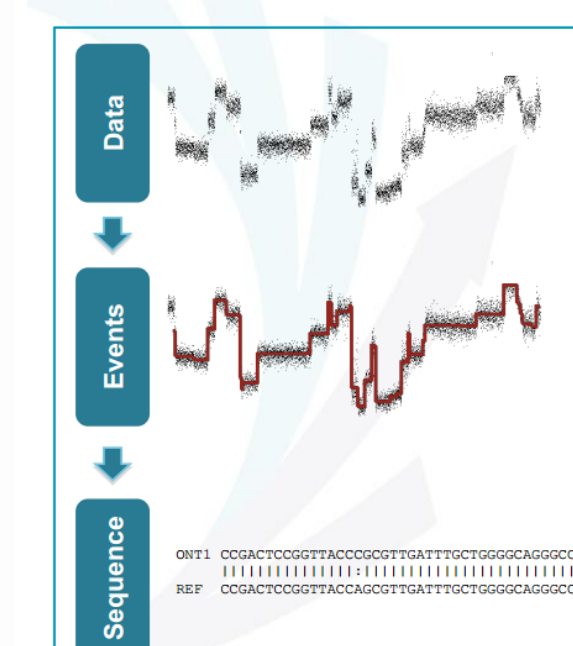
Coulter count: the transient current drop is proportional to the particle volume.



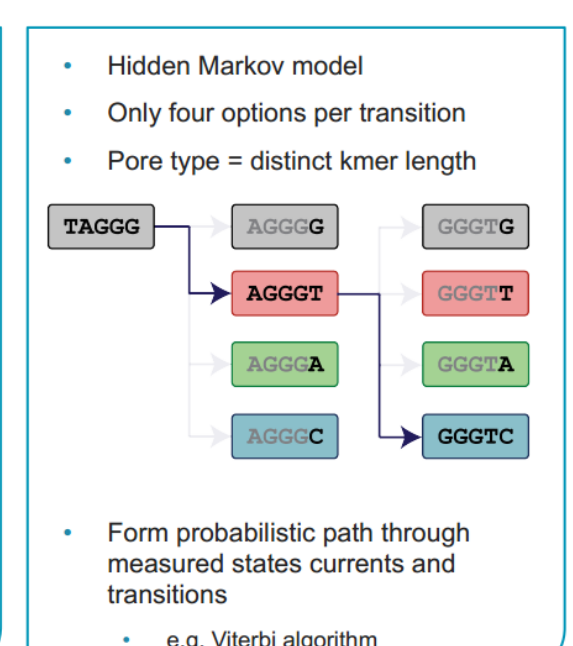
Alpha-Hemolysin: A typical nanopore self-assembled bacterial protein.



Oxford Nanopore Technologies nanopores and helicase



Signal conversion to fastq files (basecalling)



Hidden Markov model  
Only four options per transition  
Pore type = distinct kmer length  
Form probabilistic path through measured states currents and transitions  
e.g. Viterbi algorithm

- \* Up to 10 simultaneous runs
- \* 10 - 25 Gbases per run N50 15 Kbases
- \* Input : 10ng -1500ng of sample
- \* Parallel barcode sequencing: PCR products (96 samples) or native DNA/RNA (24 samples)
- \* In house or Cesga supercomputer processing
- \* Data available in less than a week
- \* Tape cartridge massive storage backup.

**Pros**  
\* It's relatively inexpensive and compact  
\* Produces exceptionally long reads.  
\* Real time  
\* In house Sequencing.  
\* Low initial capital investment

**Cons:**  
\* The error rate is higher than other technologies.  
\* In constant development

### Applications:

- \* Detection of structural variations.
- \* Genome Scaffolding
- \* Resolution of repeated sequences, haplotypes and ambiguous regions
- \* Real-time medical diagnostics and forensics (near future)
- \* Prospective applications as an environmental DNA sensor. (near future)

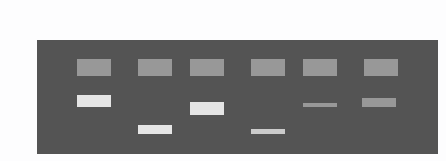
### Interesting facts

- \* Ultra-long read lengths (up to hundreds of kb). The lecture length depends on fragment length.
- \* It has been used up a mountain, in a jungle, in the arctic and at the International Space Station.

Current sequencing landscape

## 1

Amplicon validations



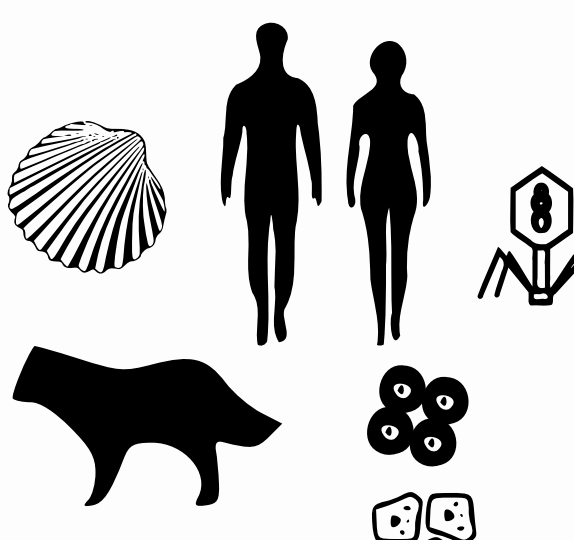
Nature manuscript validations (Accepted)

De novo assembly of cockle genome

Yeast mutation analysis

Cockle phylogenetic tree

Near 200 samples processed multiple species



Delicate Sample preparation: Contamination, short fragments Small amounts of DNA/RNA

Frequent changes in reactivities and library preparation

More than 10 versions of the sequencing, basecalling and pipelines programs

Big Informatic requeriments, more than 50 million files 30 TB of data. Thousands processing hours of in house server and Cesga super-computer

## 4 Challenges

## 3

G & D sequencing facility

**CONTACT :** Genomes & disease lab: Avenida de Barcelona, S/N Santiago de Compostela, 15706, Spain

<https://www.mobilegenomes.tech/>

[info@mobilegenomes.tech](mailto:info@mobilegenomes.tech)

[@mobilegenomes](https://twitter.com/mobilegenomes)

