Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Accelerating the Wavefront Alignment Algorithm on CPUs, GPUs and FPGAs

Miquel Moreto

Santiago Marco-Sola

4th Workshop on Accelerator Architecture in Computational Biology and Bioinformatics (AACBB), New York City, USA

June 18 2022

I

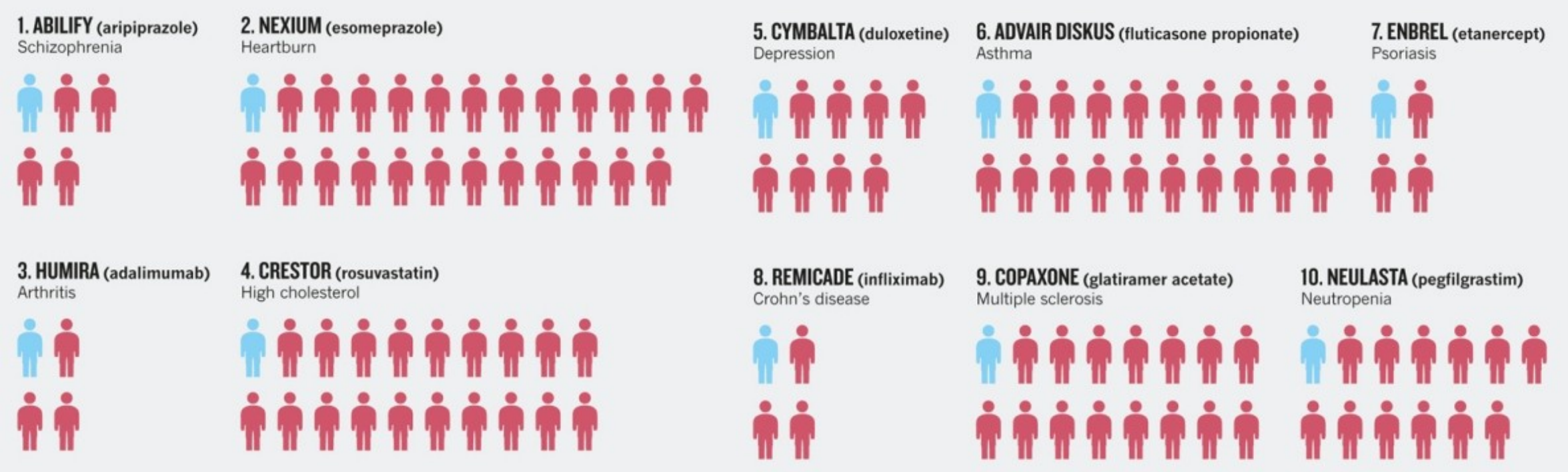
Motivation

Genome Sequencing and Precision Medicine

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Imprecision Medicine

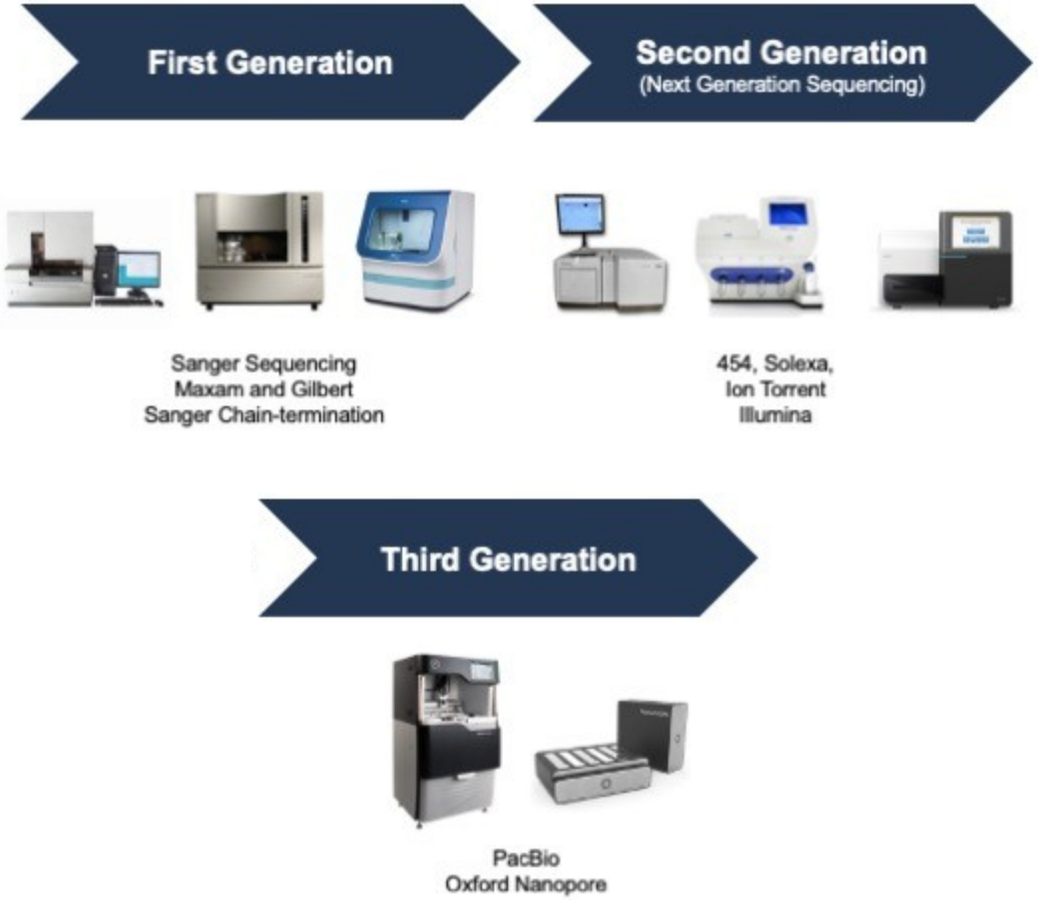
For every person they do help (blue), the ten highest-grossing drugs in the USA fail to improve the conditions of between 3 and 24 people (red).
Schork, Nicholas J. "Personalized medicine: time for one-person trials." Nature, 2015.

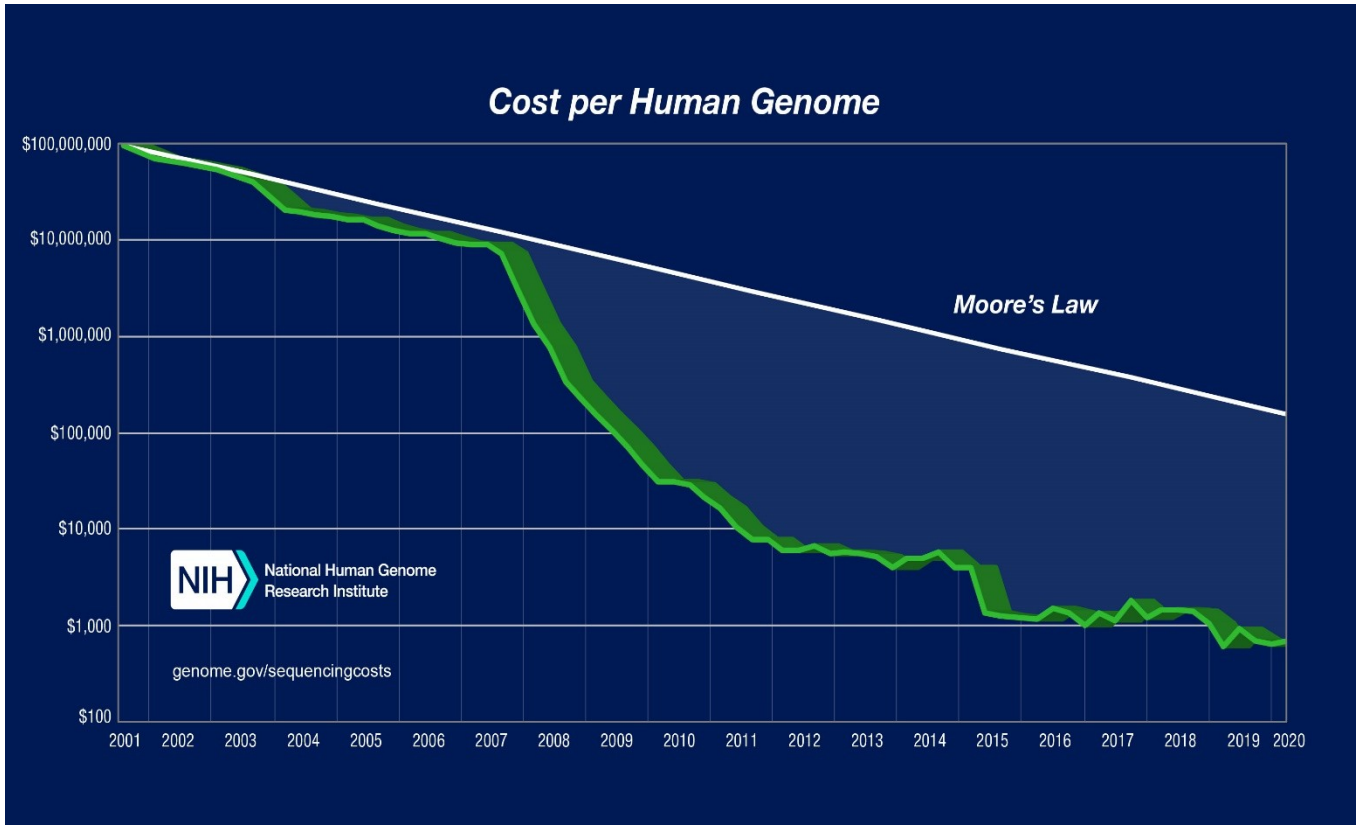


Sequencing Has Become Clinically Affordable

Nowadays, we can sequence a complete individual (i.e., **whole genome**) in less 48h for **less than $1000.**

Whole Exome for **less than $200.**





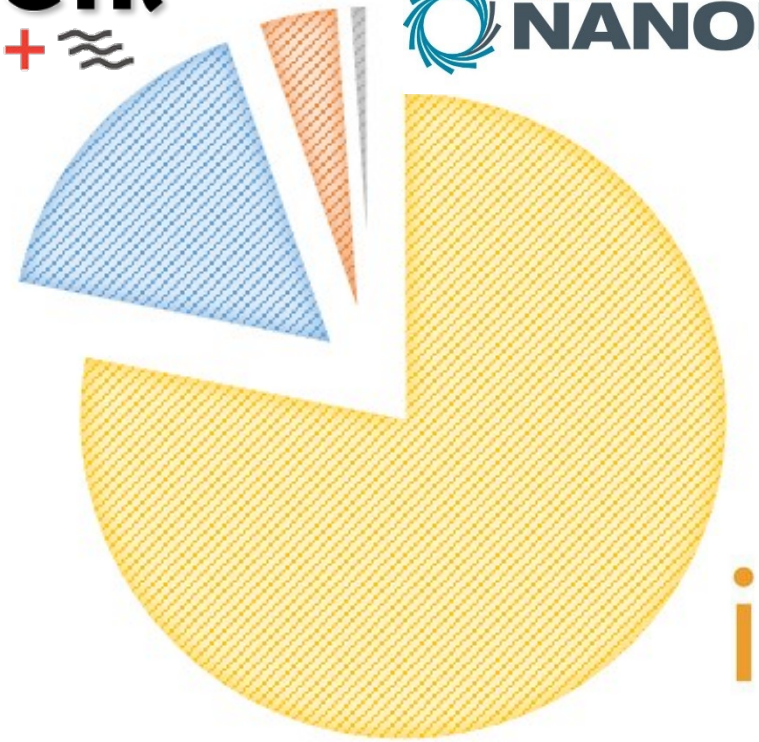
Sequencing Players

ion torrent

PACBIO®

Oxford NANOPORE Technologies

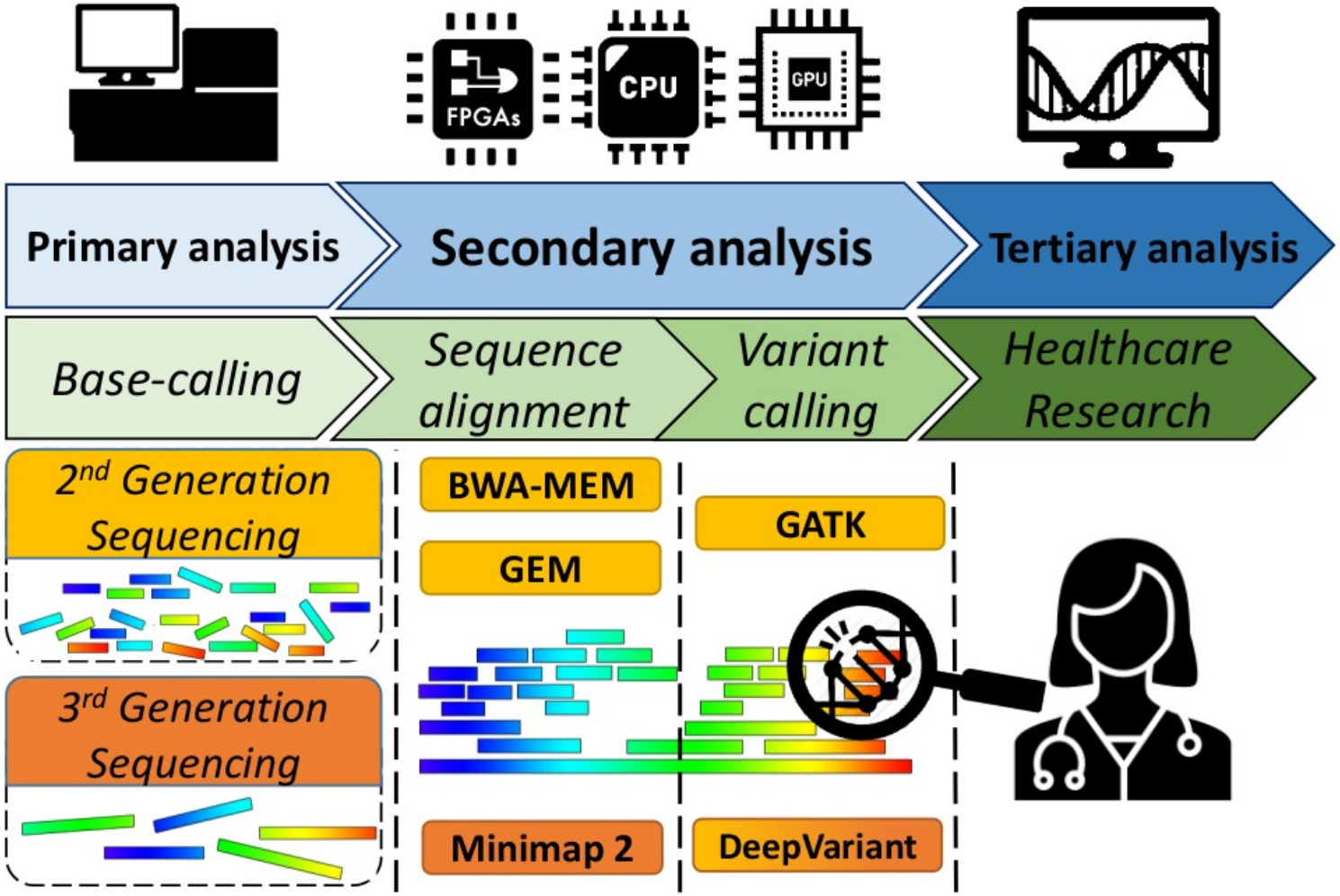
BGI华大



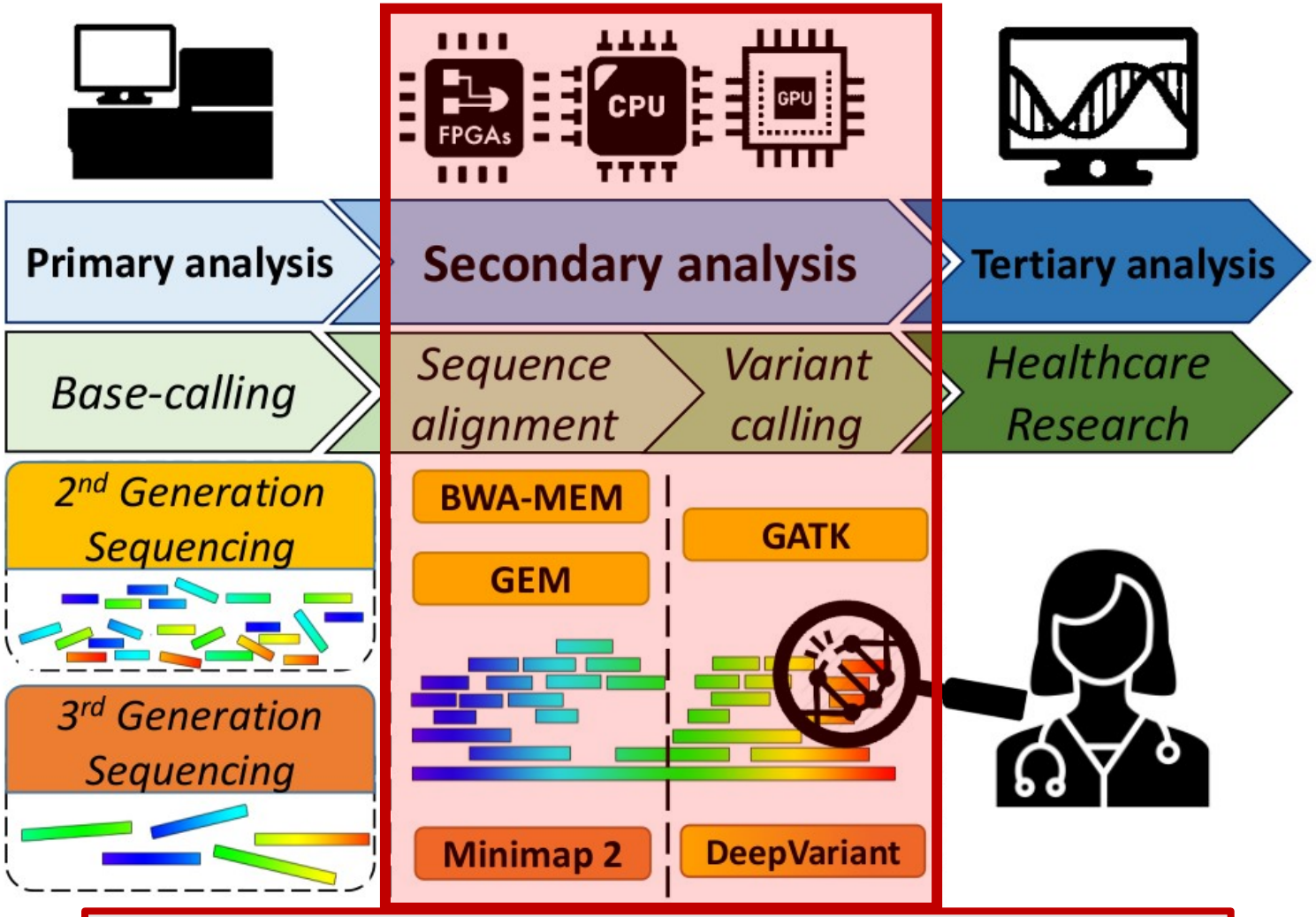
illumina®

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The Bottleneck: Sequencing Data Analysis



The Bottleneck: Sequencing Data Analysis



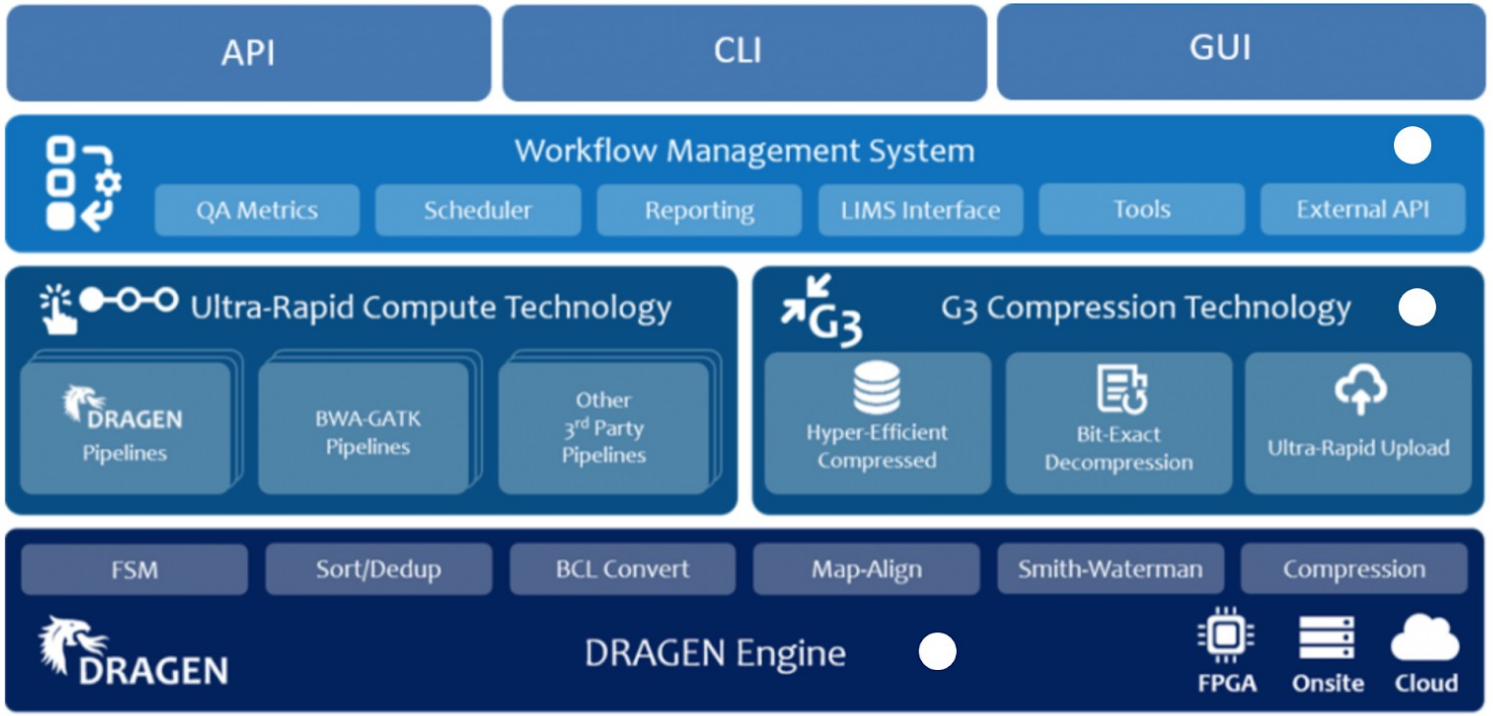
As sequencing becomes inexpensive, computational analyses become the bottleneck.

Custom hardware accelerators (Illumina, FPGAs)

illumina®

DRAGEN | gatk

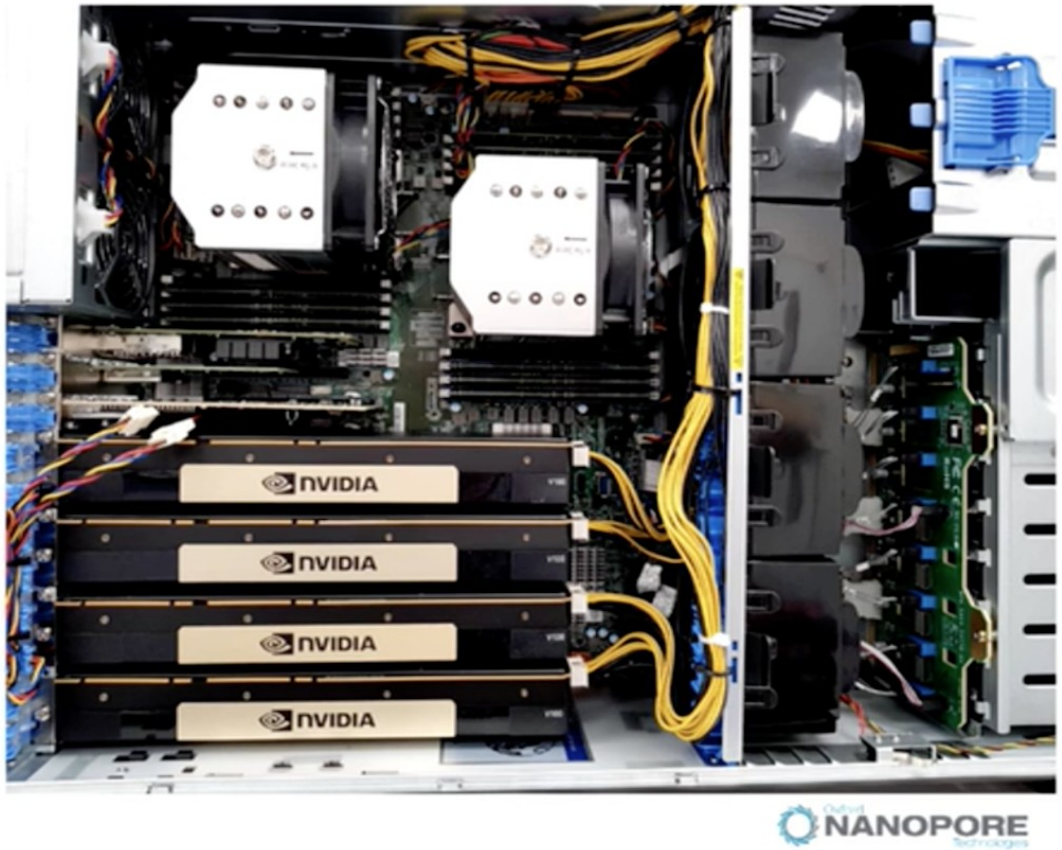
Illumina's **DRAGEN Bio-IT platform (FPGA-based)** can process NGS data for an entire human genome at 30X coverage in about 25 minutes on premise vs. 15 hours on a traditional CPU-based system (**36X**)



Custom hardware accelerators (Nanopore, GPUs)

Oxford NANOPORE Technologies

Promethion system incorporates **4 Nvidia A100 GPU** sin a dedicated system based on Intel Xeons.



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

II

The core building-block

Pairwise Alignment and Wavefront Alignment Algorithm (WFA)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

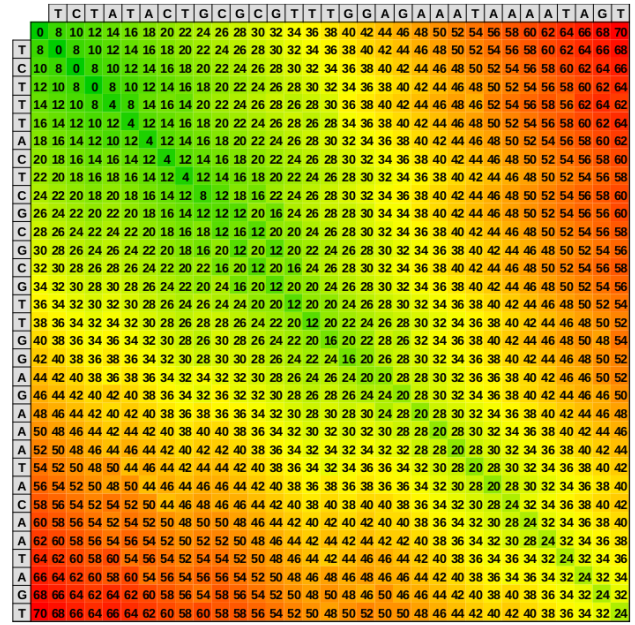
Pairwise Alignment: A fundamental problem

In sequence analysis, **pairwise alignment** compares two sequences in order to find similarities and differences.

> *"How to **convert one sequence into the other** applying a series of alignment operations (i.e., match, mismatch, insertion, and deletion) that minimize a given cost function or distance function"*

Classical pairwise alignment implementations (software and hardware) are based on **dynamic programming algorithms** (i.e., Needleman-Wunsch, Smith Waterman).

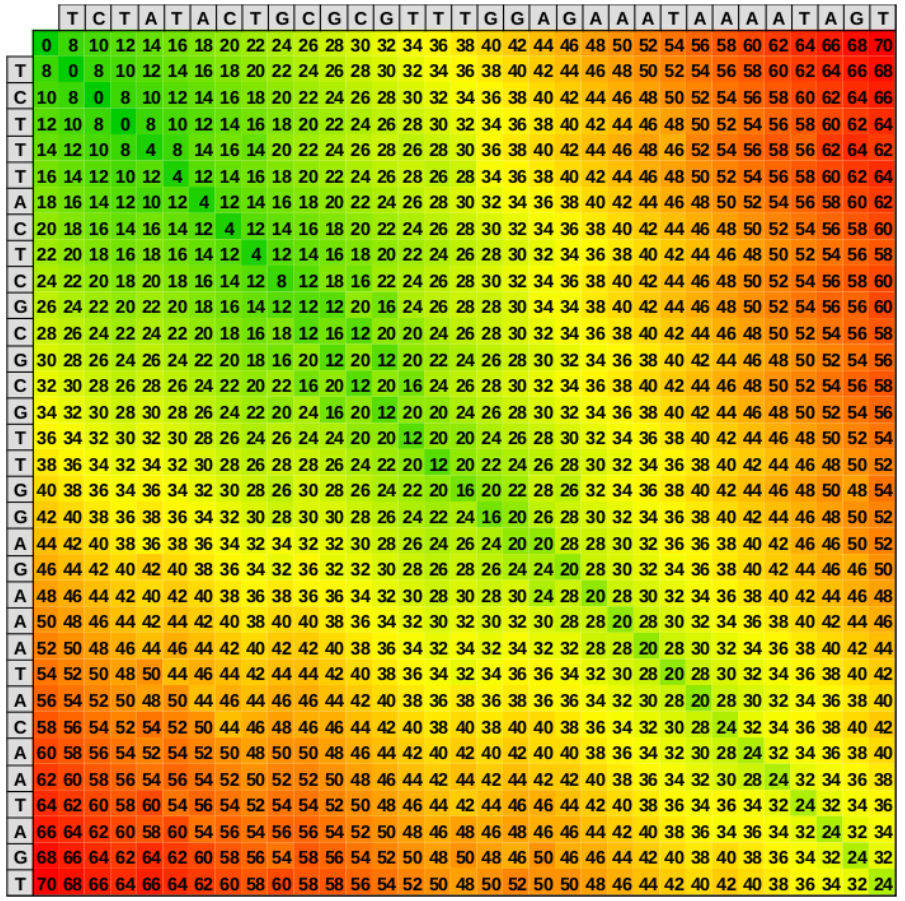
- NW or SW algorithms run in quadratic time **O(n²)** and memory **O(n²)**

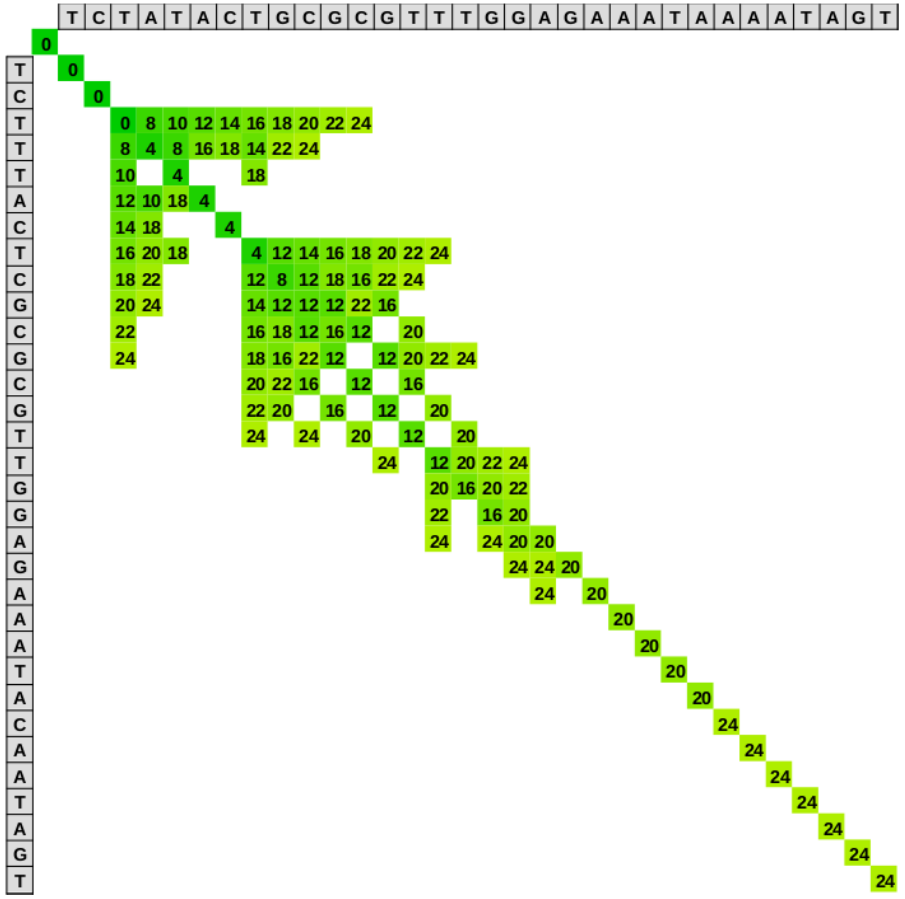


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The Wavefront Alignment Algorithm (WFA)

In a nutshell, the **WFA** algorithm produces the **sequence alignment** **computing fewer cells**, using a simple computation pattern.



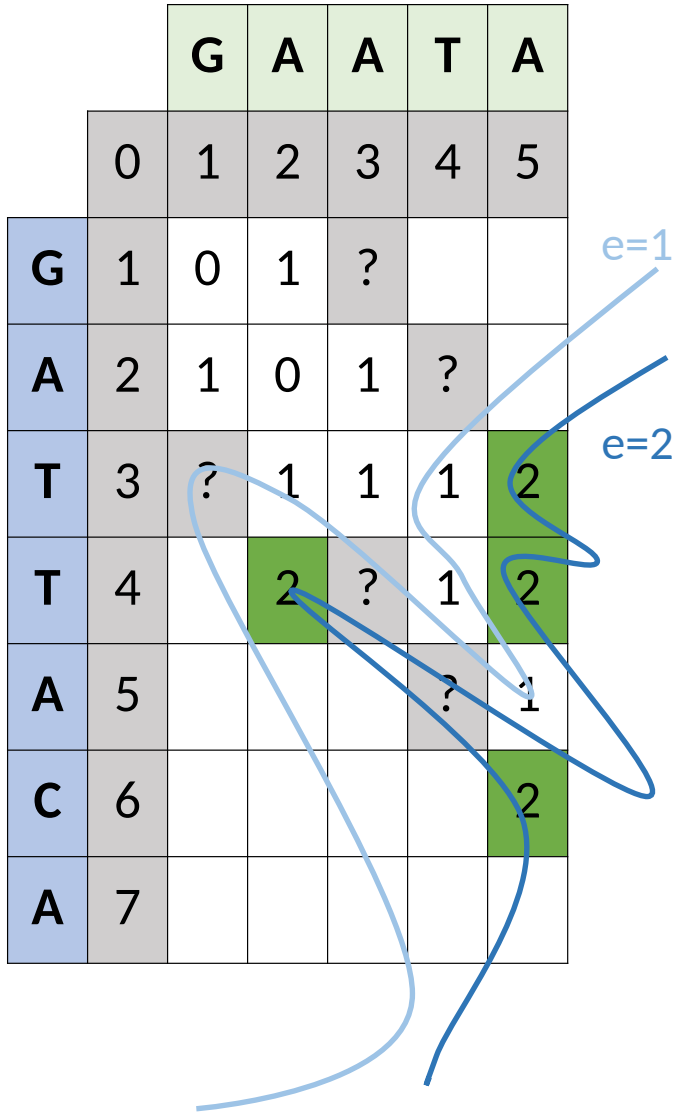


The Wavefront Alignment Algorithm (WFA)

The **Wavefront Alignment algorithm (WFA)** runs in **O(ns)** time and **O(s²)** memory.

Main insights:

- **Compute cells in order of increasing score.**
  - Avoid the computation of suboptimal cells
- **Center penalties at MatchScore=0.**
  - Take advantage that matches along the diagonal don't increase the score **(extend diagonals for free)**
- Note that **scores are monotonically increasing** along the diagonal.
  - Compute only the most advanced cell in each diagonal **(farthest reaching cell)** with a given score



WFA: A Graphical Overview

WFA: The algorithmic recipe

- Change the layout to diagonal transitions.
- Adapt DP equations for diagonal-transitions.

$$\widetilde{I}_{s,k} = \max\{\widetilde{M}_{s-o-e,k-1} + 1, \widetilde{I}_{s-e,k-1} + 1\}$$

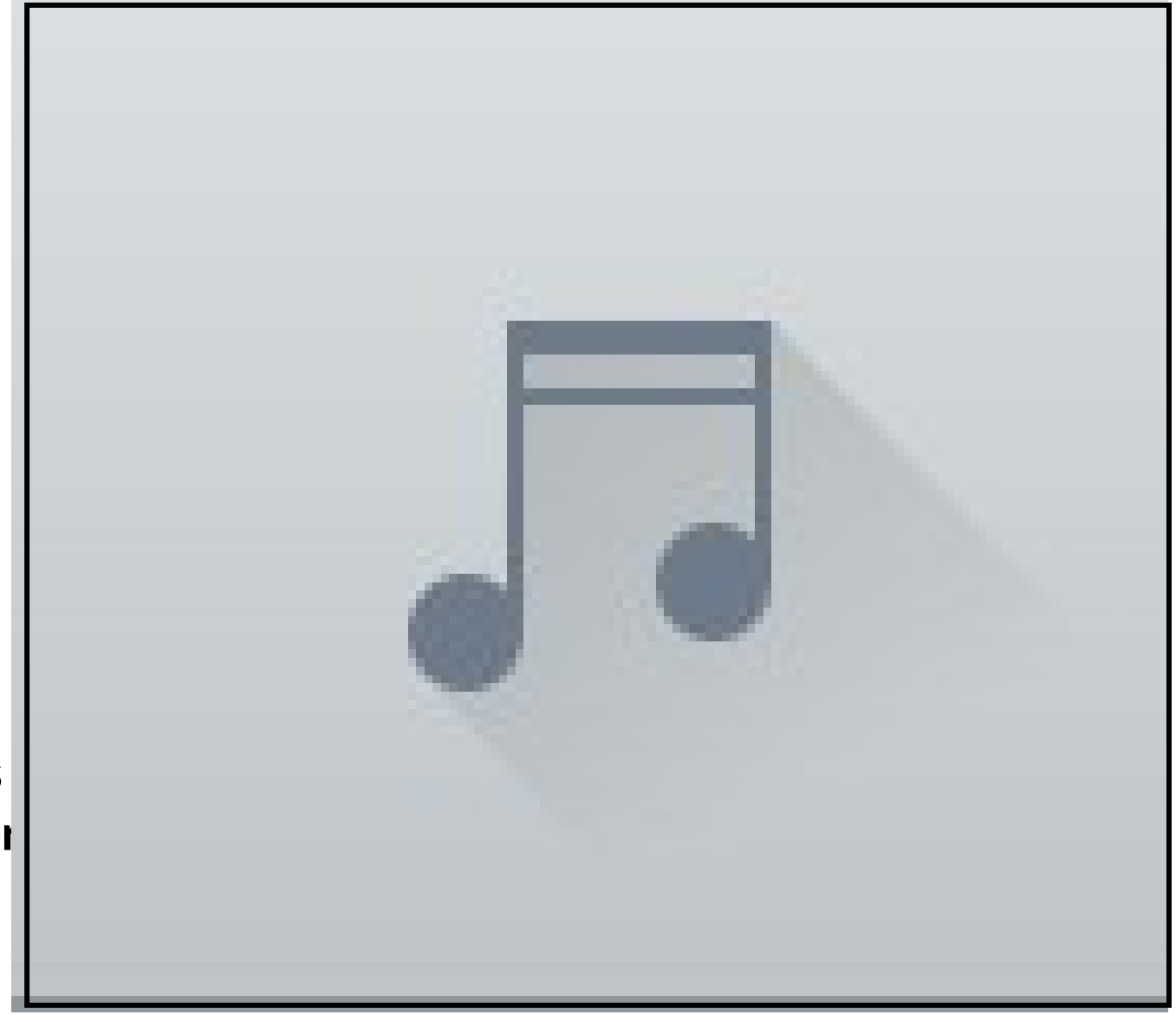
$$\widetilde{D}_{s,k} = \max\{\widetilde{M}_{s-o-e,k+1}, \widetilde{D}_{s-e,k+1}\}$$

$$\widetilde{X}_{s,k} = \max\{\widetilde{M}_{s-x,k} + 1, \widetilde{I}_{s,k}, \widetilde{D}_{s,k}\}$$

$$\widetilde{M}_{s,k} = \widetilde{X}_{s,k} + LCP(q_{\widetilde{X}_{s,k}-k\ldots n-1}, t_{\widetilde{X}_{s,k}\ldots m-1})$$

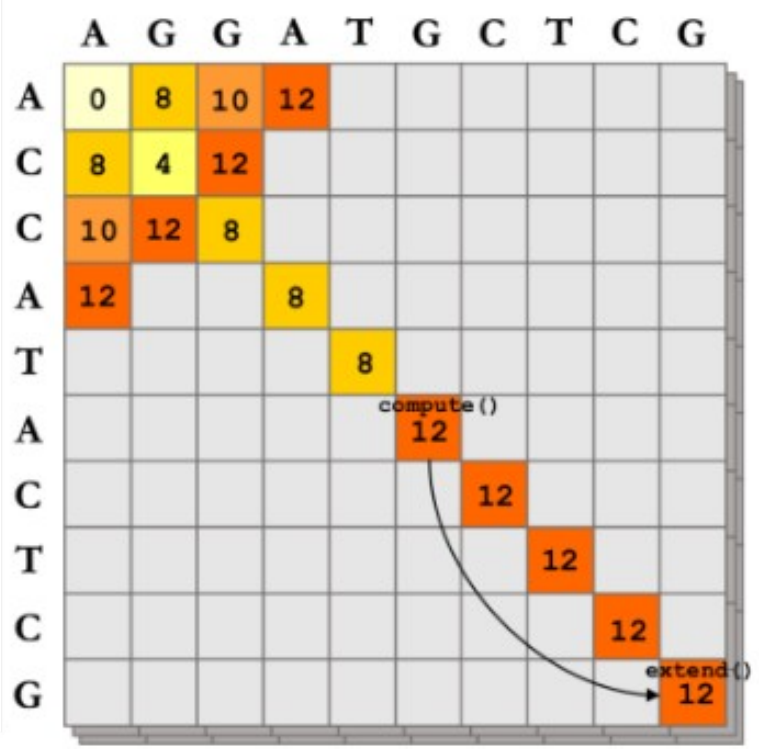
For s **in** {0 … OptimalScore} **do**

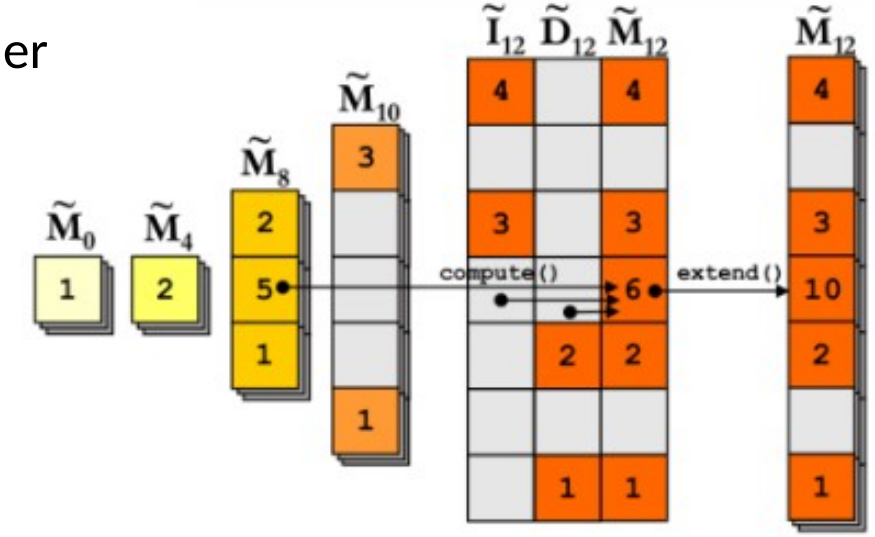
1. **Extend matches** over each diagonal (taking advantage that MatchScore=0.)
2. Compute (s+1) wavefront using recurrences (computes only the **most advanced cell** of **scor s+1** for **each diagonal**.



WFA highlights

- WFA is an **exact algorithm** to compute the alignment between two sequences in **O(ns) time** and **O(s²) space**.
  - Takes advantage of homologous regions between the sequences, scaling with the read-length.
- **Alphabet independent,** no preprocessing needed, no fixed band, no assumptions.
- Convenient for **vectorization** and **parallelization**.
  - More effective SIMD as it encodes offsets instead of scores (e.g., 8-bits integers for sequences < 256 bp).
  - **Automatically vectorized** using SIMD instructions by the compiler (e.g., x86-AVX, Arm NEON/SVE, RISC-V V extension)
- **Global** and **ends-free** alignment support.
- **Compatible** with classical **heuristics**:
  - Banded, Adaptive, X-drop, Z-drop, …
- **Compatible** with main-stream distance/**score functions**:
  - Indel (LCS), edit, gap-linear, gap-affine, gap-affine piecewise, concave penalty functions, and others.

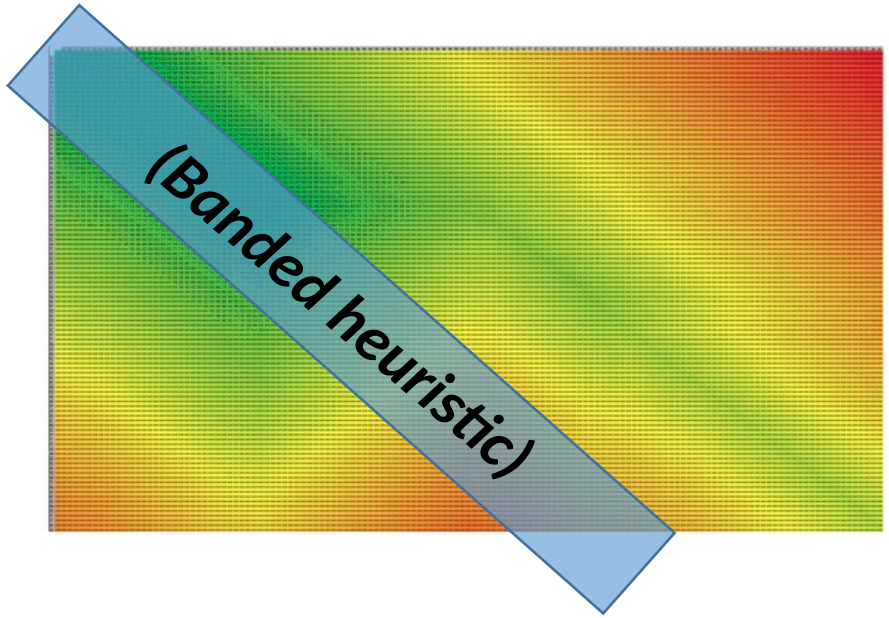




WFA is Exact and Precision in Genomics is Key

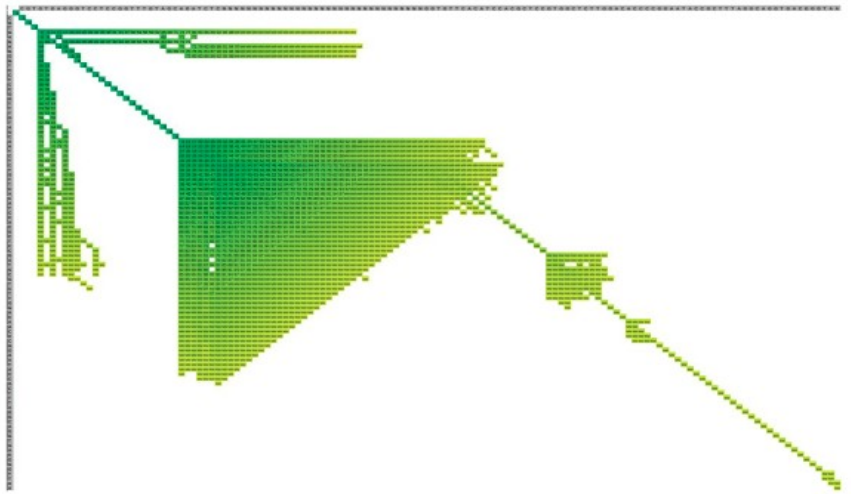
Dynamic Programming (DP)

HEURISTICS = SUBOPTIMAL ALIGNMENT



Wavefront Gap-Affine (WFA)

EXACT = OPTIMAL ALIGNMENT

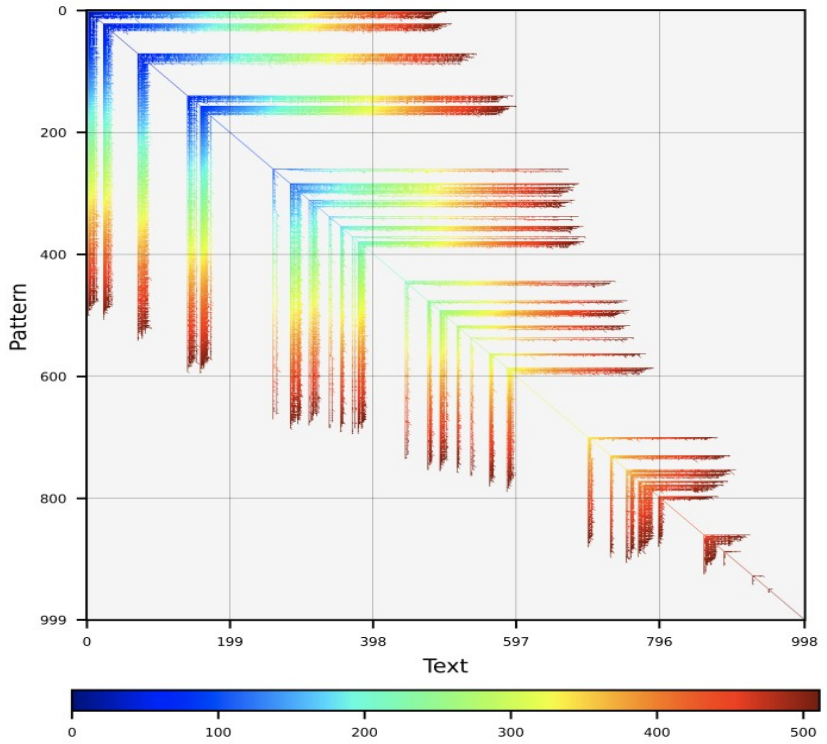


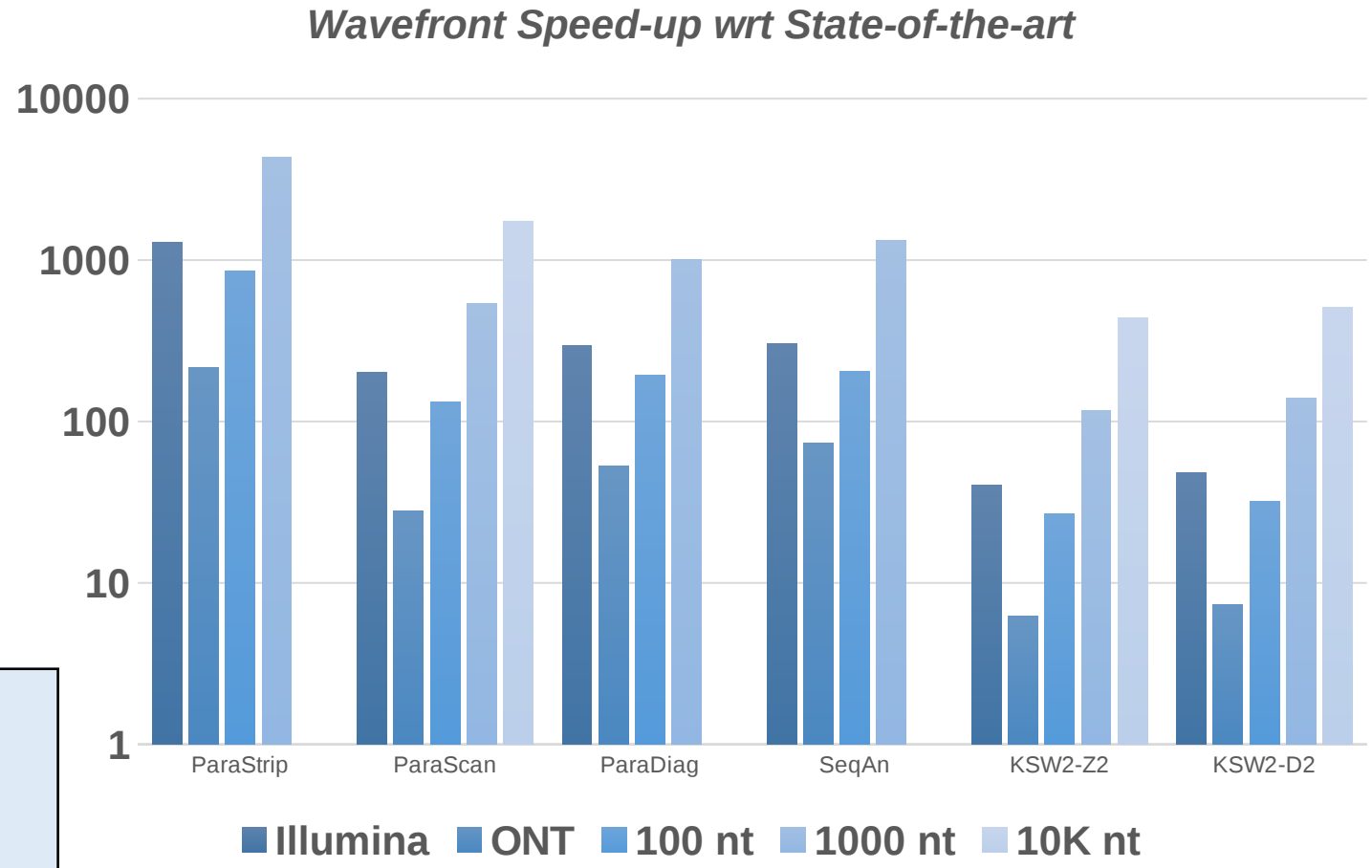
Heuristic methods can miss the optimal alignment
WFA is exact and can navigate through long insertions/deletions

WFA Compared To The State-of-the-art

20-300x faster than other methods aligning Illumina sequences.

10-100x faster than other methods aligning Oxford Nanopore sequences.





Marco-Sola, Santiago, Juan Carlos Moure, Miquel Moreto, and Antonio Espinosa. "Fast gap-affine pairwise alignment using the wavefront algorithm." Bioinformatics 37, no. 4 (2021): 456-463.

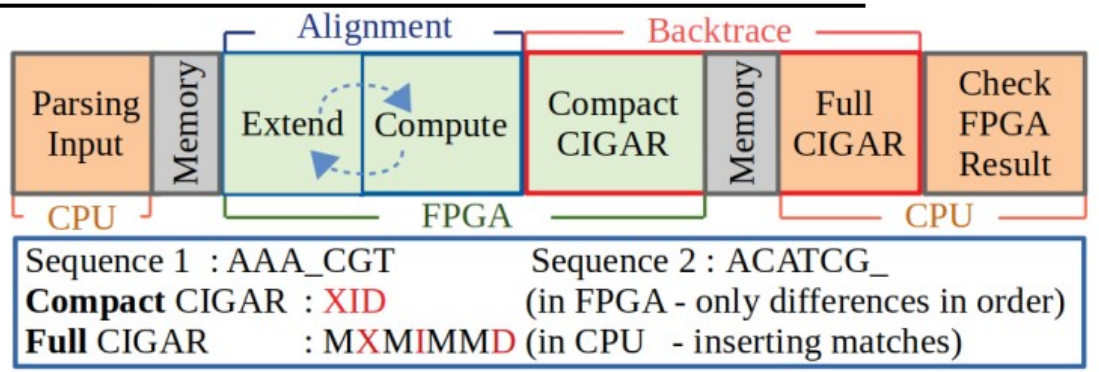
Hardware Accelerators

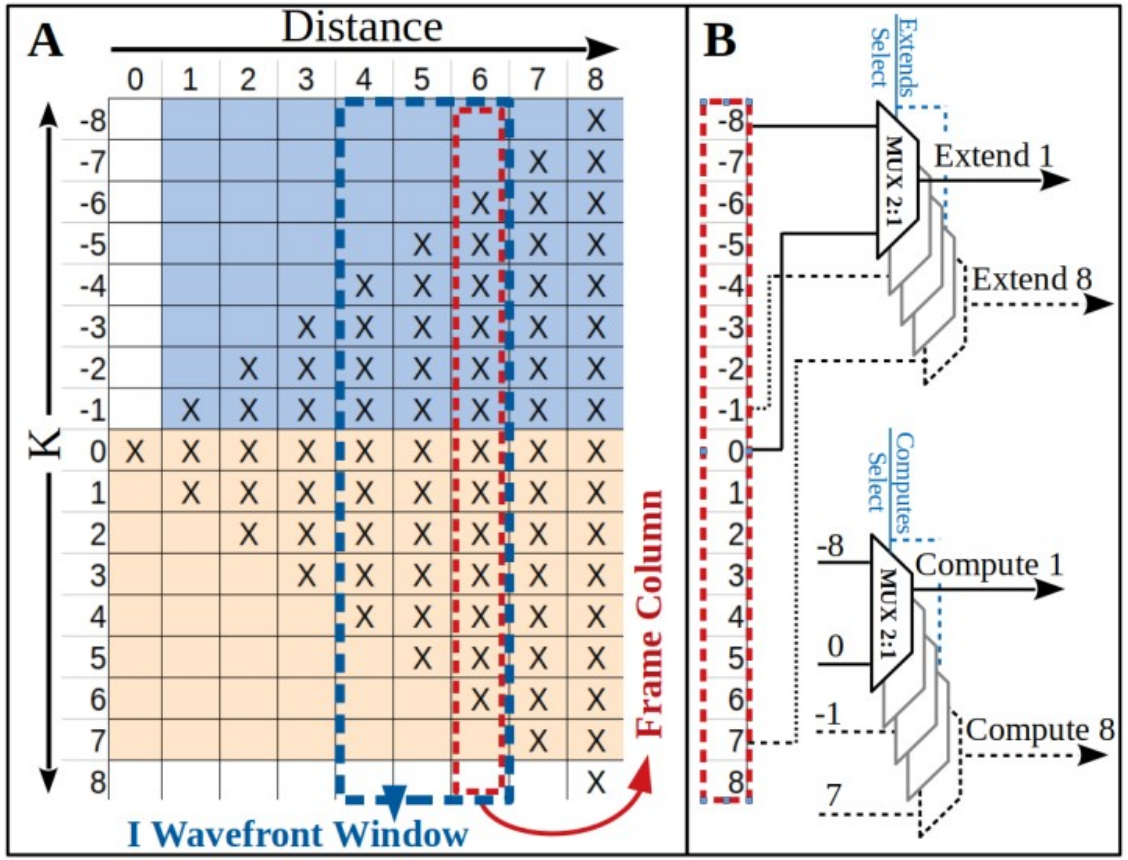
Accelerating the Wavefront Alignment Algorithm

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Short Sequence WFA-Alignment on FPGA

- **Target:** Short Illumina-like sequences
- **Accelerator:** FPGAs
- **Rational:**
  - FPGA's high design flexibility
  - Hardware/Software co-design
- **Results:**
  - Speedups of up to 13.5×
  - Consuming up to 14.6× less energy

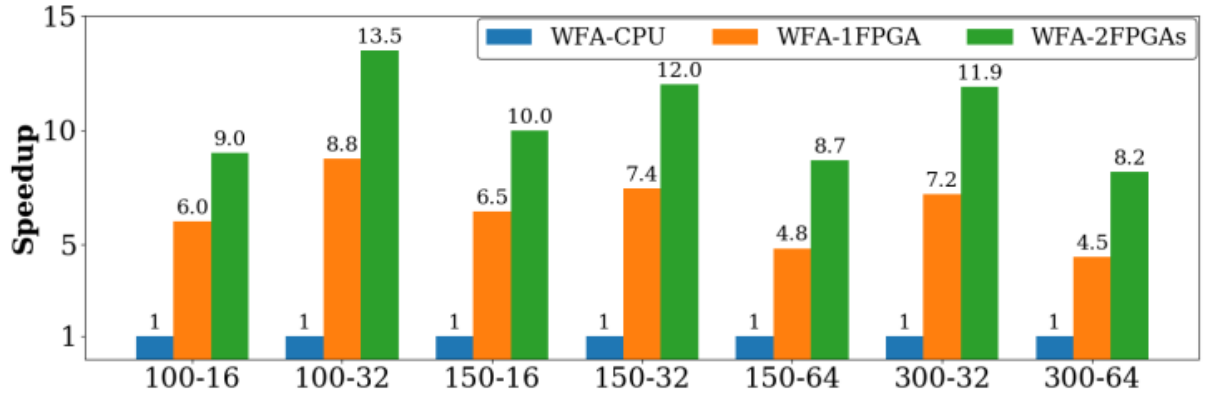


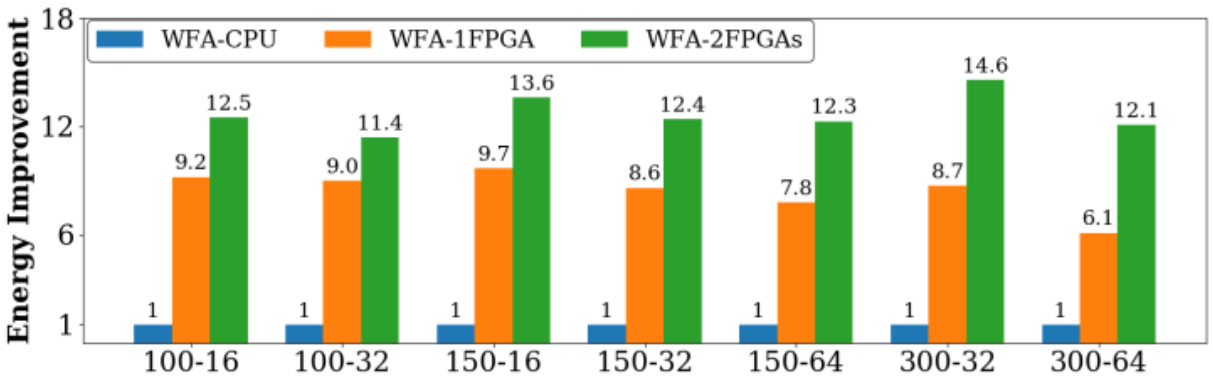


Haghi, Abbas, Santiago Marco-Sola, Lluc Alvarez, Dionysios Diamantopoulos, Christoph Hagleitner, and Miquel Moreto. "An FPGA Accelerator of the Wavefront Algorithm for Genomics Pairwise Alignment." In 2021 31st International Conference on Field-Programmable Logic and Applications (FPL). IEEE, 2021.

Short Sequence WFA-Alignment on FPGA

On a **POWER9** 2xADM-PCIE-9H7 FPGA-boards (CAPI-enabled), our design is **13.5× faster than the CPU-WFA**, consuming **14.6× less energy**. **Delivers 16x more GCUPs** than any other FPGA-based alignment design.

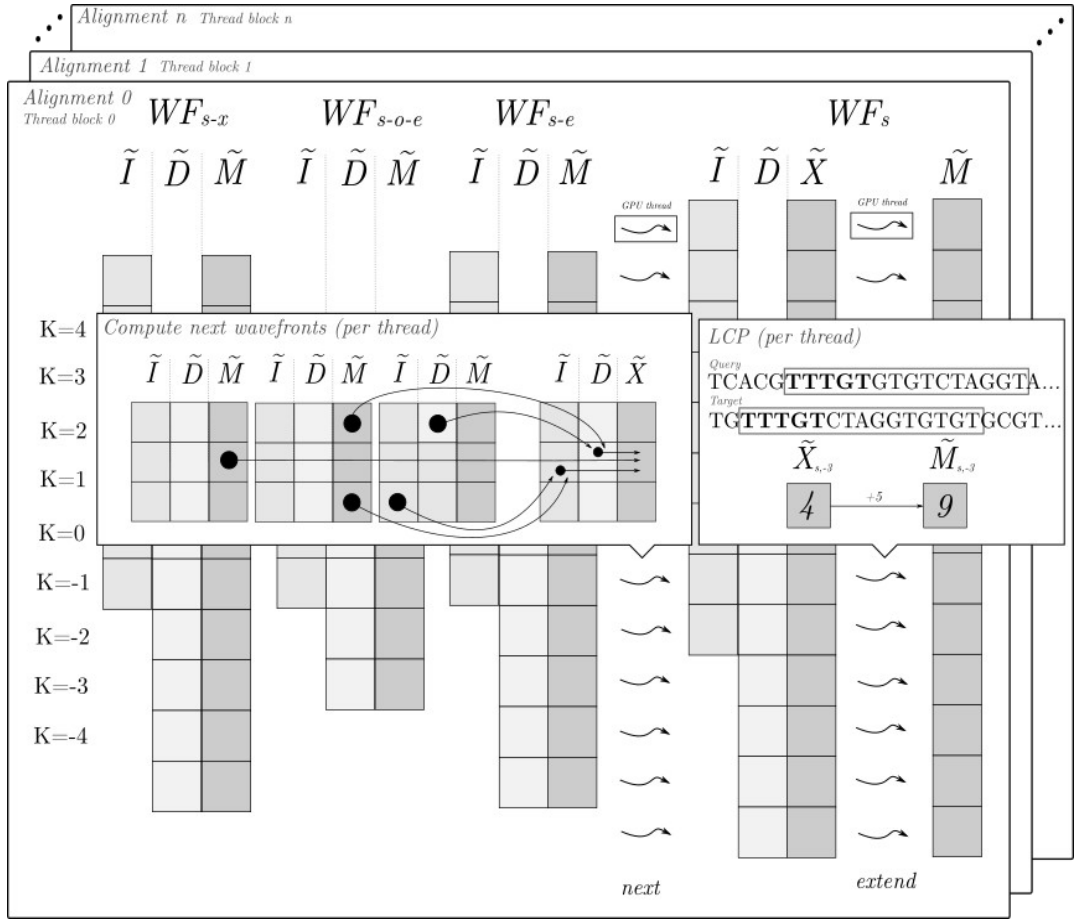




Paper	Year	Device	Freq. (MHz)	GCUPS
Ours	2021	2× Xilinx Virtex U+ XCVU37P	200	2073.7
Ours	2021	1× Xilinx Virtex U+ XCVU37P	200	1251.7
[56]	2019	Xilinx VU9P Ultrascale	200	8.7*
[57]	2018	Altera Stratix V	n/a	58.4
[14]	2018	Xilinx Virtex7 XC7VX485T	200	105.9
[54]	2018	Intel Arria 10 GX	n/a	125.0**
[64]	2013	Altera Stratix V A7	200	24.7
[58]	2011	Xilinx XC5VLX330T	130	129.0***
[63]	2009	Xilinx XC2V6000-4	47.6	8.0
[59]	2007	Altera EPS1S30	82	6.6

Long Sequence WFA-Alignment on GPU

- **Target:** Long sequences like those produced by PacBio or Oxford Nanopore technologies
- **Accelerator:** GPUs
- **Rational:**
  - Big alignments generate large wavefronts that allow performing large parallel computations
  - High computing throughput
  - High memory bandwidth
- **Results:**
  - Up to 176X compared to other GPU implementations
  - Up to 4 orders of magnitude faster than other CPU implementations

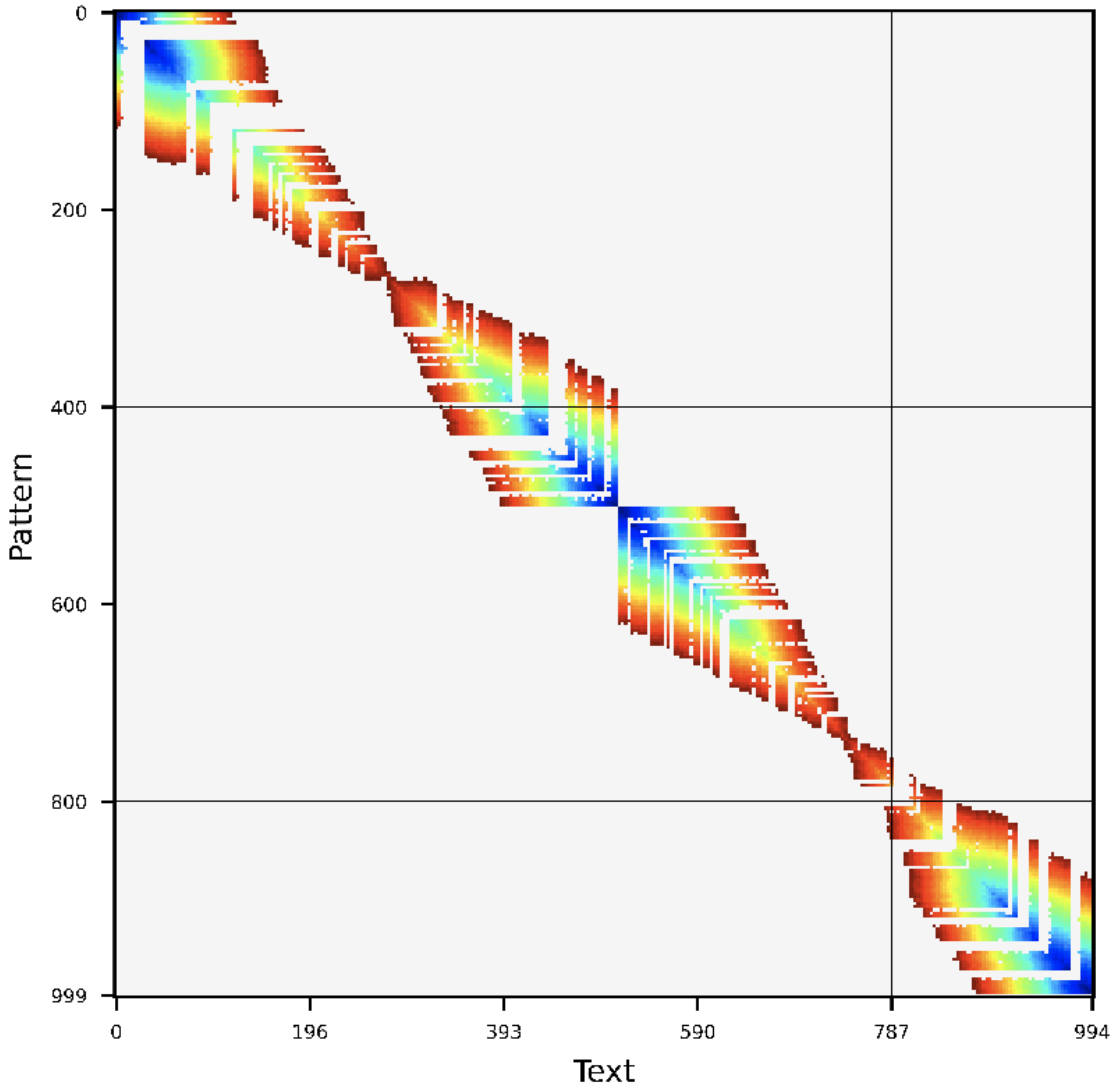


Aguado-Puig, Quim, Santiago Marco-Sola, JuanCarlos Moure, David Castells, Lluc Alvarez, Antonio Espinosa, and Miquel Moreto. "Accelerating Edit-Distance Sequence Alignment on GPU using the Wavefront Algorithm." IEEE Access (2022).

Aguado-Puig, Quim, Santiago Marco-Sola, Juan Carlos Moure, Christos Matzoros, David Castells-Rufas, Antonio Espinosa, and Miquel Moreto. "WFA-GPU: Gap-affine pairwise alignment using GPUs." bioRxiv (2022).

Ultra-Long Sequence WFA-Alignment (BiWFA)

- **Target:** Ultra-long Nanopore sequences, assembly contigs, or whole genomes
- **Problem:** Even a $O(s^2)$-space algorithm fails to scale to MBs-long sequences.
- **BiWFA core idea:**
  - Perform the WFA algorithm simultaneously from both ends (i.e., forward and reverse)
  - Find the optimal breakpoint (where both wavefronts end)
  - Repeat BiWFA on the remaining halves.

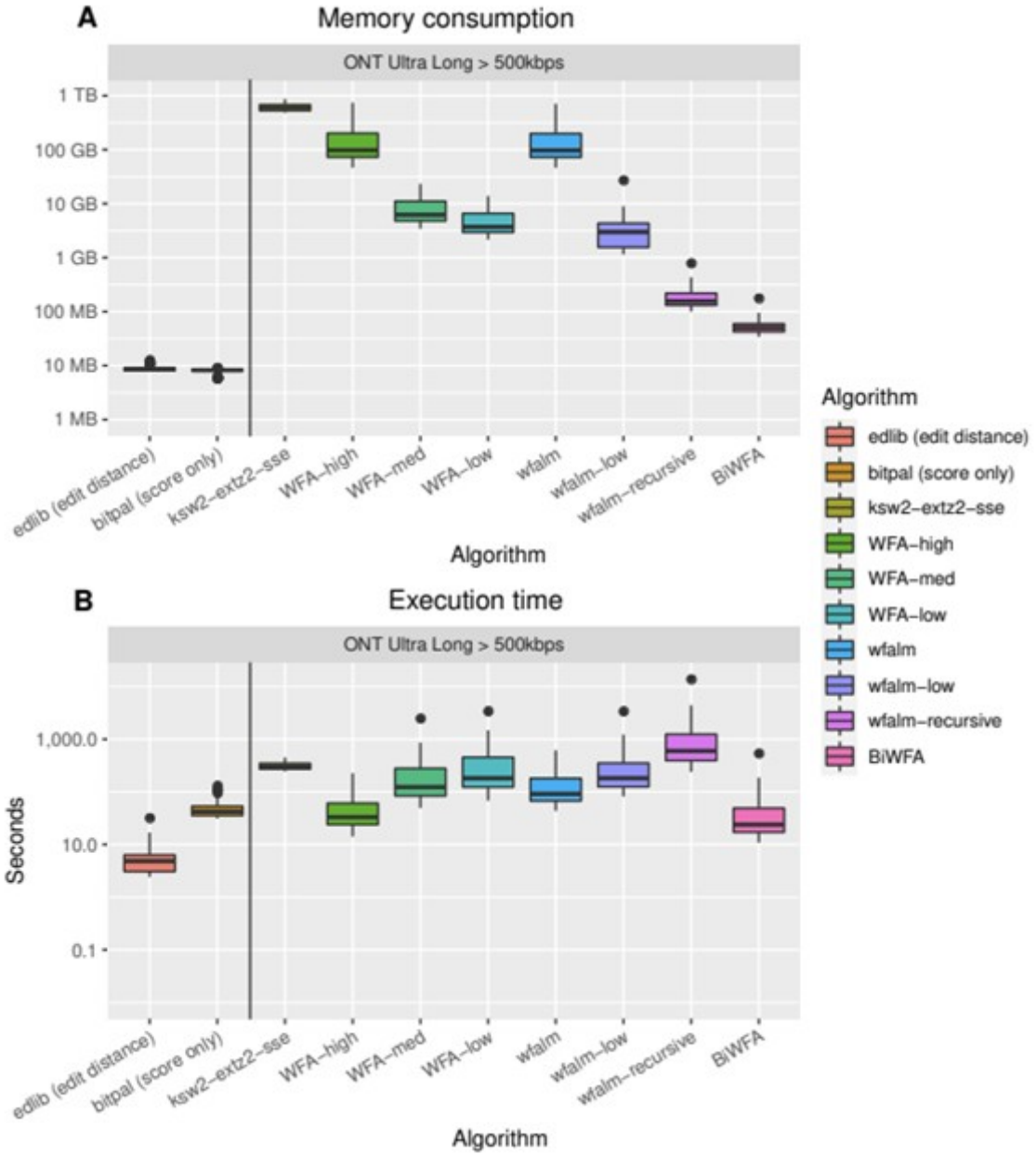


Marco-Sola, Santiago, Jordan M. Eizenga, Andrea Guarracino, Benedict Paten, Erik Garrison, and Miquel Moreto. "Optimal gap-affine alignment in O (s) space." bioRxiv (2022).

Ultra-Long Sequence WFA-Alignment (BiWFA)

- **Aligning Nanopore >500 Kbps long** sequences, KSW2 (Minimap2) was the most efficient solution (requiring ~1TB of memory).
- Only **WFA-based methods** were able to lower the memory footprint (~100GB).

- **BiWFA** is the first gap-affine algorithm capable of computing optimal alignments in **O(s) memory** while retaining the WFA's complexity of **O(ns) time**.
- In practice, it **never requires more than 183 MB** to align long and noisy sequences up to 1 Mbp long, while **maintaining competitive execution times**.

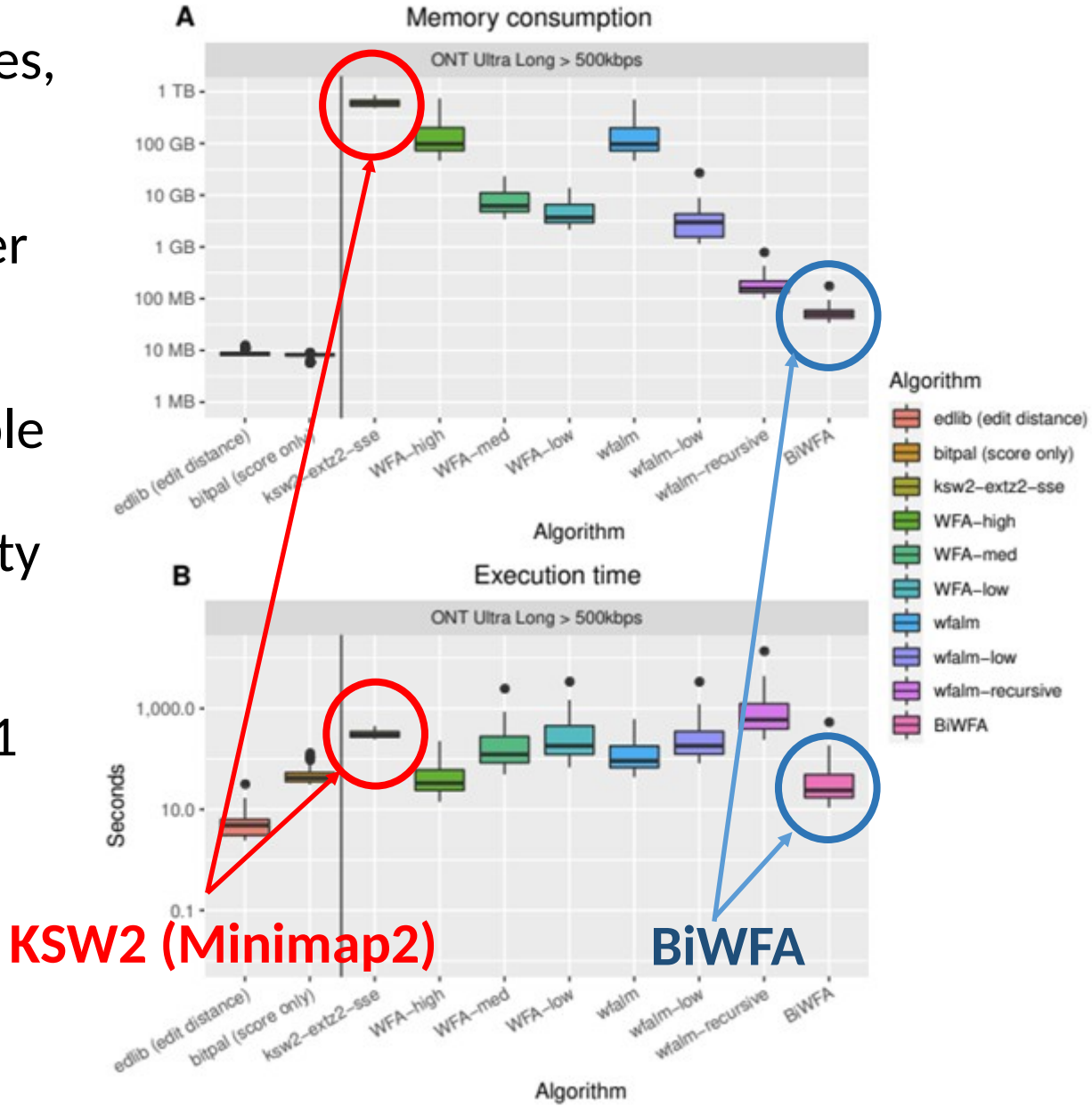


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Ultra-Long Sequence WFA-Alignment (BiWFA)

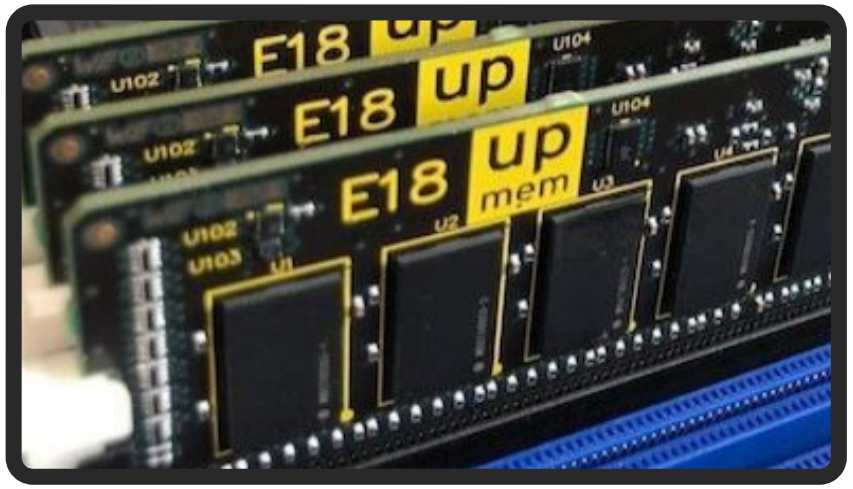
- **Aligning Nanopore >500 Kbps long** sequences, KSW2 (Minimap2) was the most efficient solution (requiring ~1TB of memory).
- Only **WFA-based methods** were able to lower the memory footprint (~100GB).

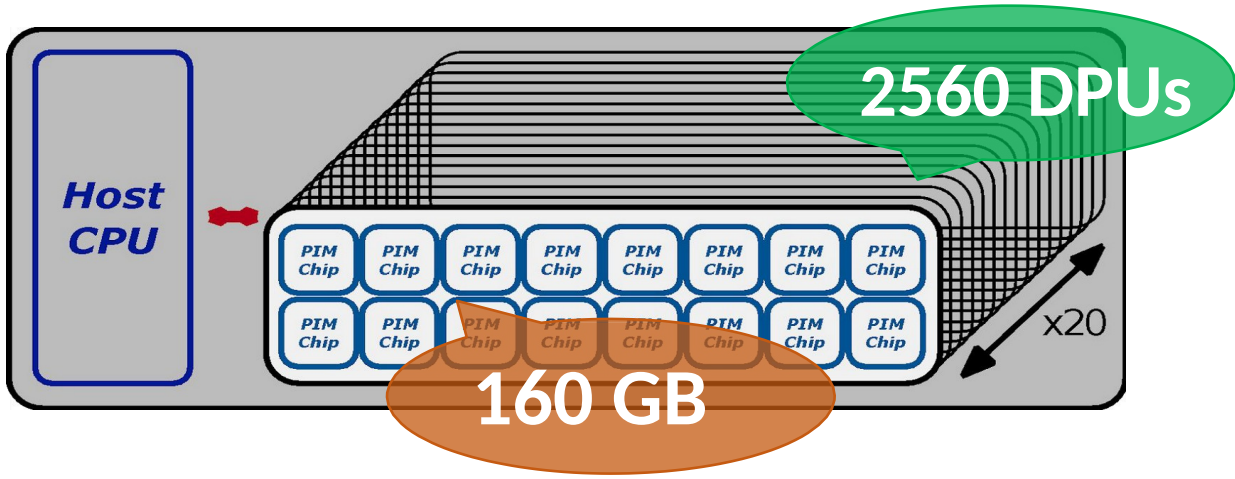
- **BiWFA** is the first gap-affine algorithm capable of computing optimal alignments in **O(s) memory** while retaining the WFA's complexity of **O(ns) time**.
- In practice, it **never requires more than 183 MB** to align long and noisy sequences up to 1 Mbp long, while **maintaining competitive execution times**.



Ultra-Long Sequence WFA-Alignment In-Memory

- **Accelerator:** In-Memory Processing
- **Rational:**
  - Still, a 100MB memory footprint is bounded by memory penalties.
  - Reduce memory movement penalties computing wavefronts in-memory.
- **WIP:**
  - Each DPU receives a set of pairs to compare.
  - Each Tasklet computes a subset of the DPU's pairs.
  - WFA structures are located in DRAM, which allows to compare longer sequences and improves performance.
- **Early results:** 19x-42x speedup vs single CPU thread, computing only the score.





IV

The Impact

WFA in the community: enabling faster and scalable tools

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

WFA Beyond Publications

WFA has gained traction in the community.
Many research groups are **adopting the WFA algorithm**: re-implementing, extending, and improving it.

- Impact in **Life Science Research**
  - **AnchorWave:** Plant Genome aligner (Cornell University).
  - **WFMash:** DNA sequence read mapper based on the WFA (University of Tennessee)
  - **VG mapper:** Sequence mapper to a variation graph or pangenome (UCSC Genomics Institute)
  - **MiniGraph:** Sequence mapping to sequence-graph (Harvard Medical School)
- Impact in **Computer Science Research**
  - **MiniWFA**: WFA reimplementation (Harvard Medical School)
  - **Wfalm**: WFA reimplementation for low-memory (UCSC Genomics Institute)
  - Alt. **FPGA** Implementations using HLS. (Politecnico di Milano)
  - Alt. **GPU** Implementations (Politecnico di Milano)
  - **In-Memory WFA** Implementation (American University of Beirut and ETH Zurich)

Conclusions and Next Steps

- Pairwise alignment is a fundamental building block in many genome analysis applications
- The WaveFront Alignment (WFA) algorithm is an **algorithmic breakthrough** that reduces the complexity in time and memory of traditional approaches **with precision**
- Acceleration by several orders of magnitude using CPUs, GPUs and FPGAs
- Many popular toolkits are starting to incorporate the WFA algorithm
- Currently working on:
  - Integration of the WFA algorithm in full mappers: minimap2, BWA-MEM2, GEM
  - Fabrication of an ASIC in GF 22nm that accelerates the WFA algorithm:
    - Lagarto RISC-V processor with custom vector instructions
    - On-chip accelerator integrated with AXI
    - RTL freeze next month

Lenovo 联想 arm DRAC ePProcessor

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The eProcessor project has received funding from the European High-Performance Computing Joint Undertaking Joint Undertaking (JU) under grant agreement No 956702. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Sweden, Greece, Italy, France, and Germany.

Designing RISC-V-based Accelerators for next generation Computers

Thank you!!!

The DRAC project with -file number 001-P-001723- has been 50% co-financed with € 2,000,000.00 by the European Union Regional Development Fund within the framework of the ERDF Operational Program of Catalonia 2014-2020, with the support of Generalitat of Catalonia.