www.bsc.es

EXCELENCIA SEVERO OCHOA

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Are Next-Generation HPC Systems Ready for Population-level Genomics Data Analytics?

Calvin Bulla, Lluc Alvarez and **Miquel Moretó**

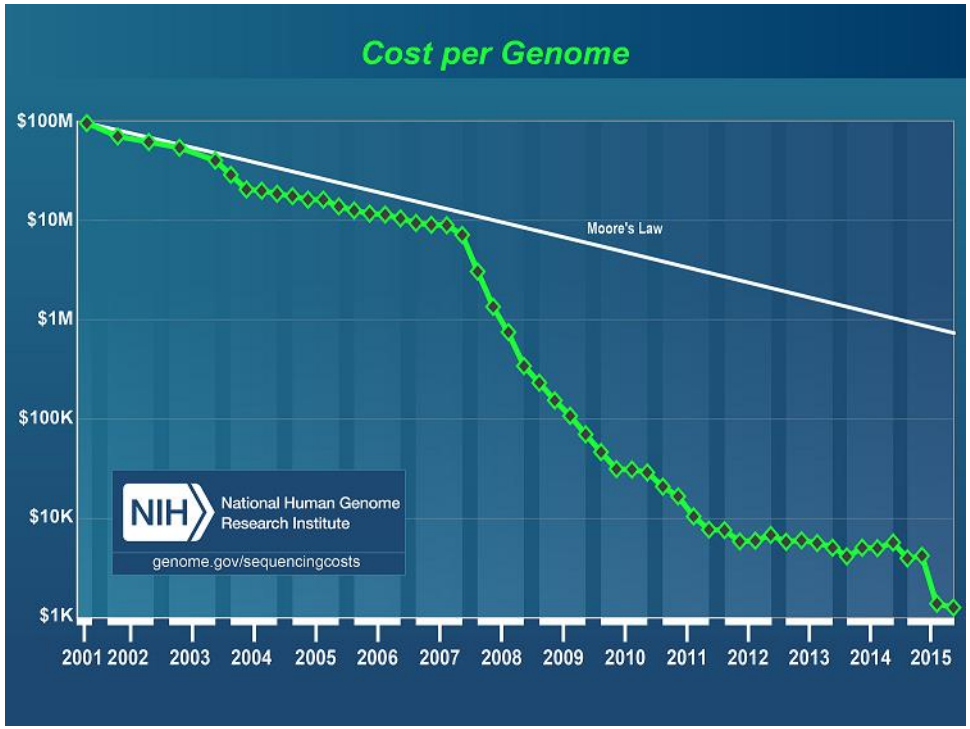
AACBB Workshop, 24/02/2018

HiPEAC

erc

Genome Sequencing Explosion

- Faster-than-Moore's-Law growth!



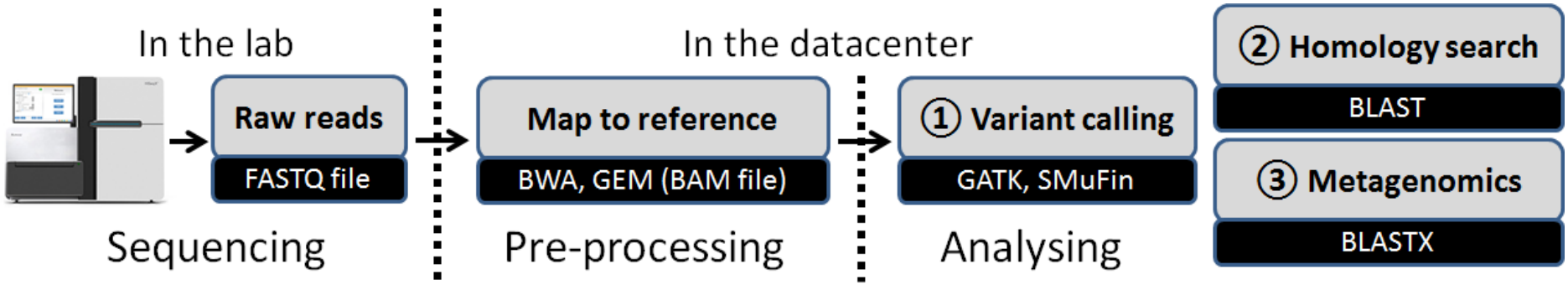
Whole Human Genome (WHS) sequencing cost <1K$

10x increase per year in genomics data

Source (left): National Human Genome Research Institute
Source (right): B. Berger et al., CACM 2016

Genomics Data Analytics

Typical workflow for WHG sequencing analytics



Main challenge: the performance bottleneck in these applications is moving from the sequencing side (as used to be the case in the last decade) towards the computing side.

Barcelona Supercomputing Center (BSC)

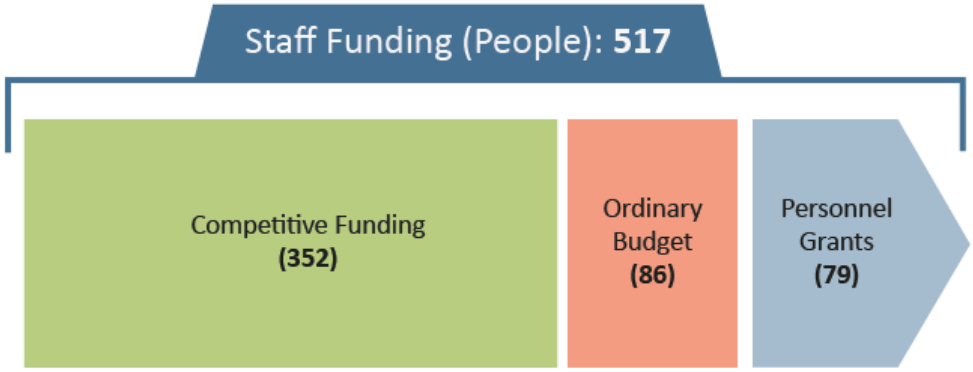
BSC objectives:

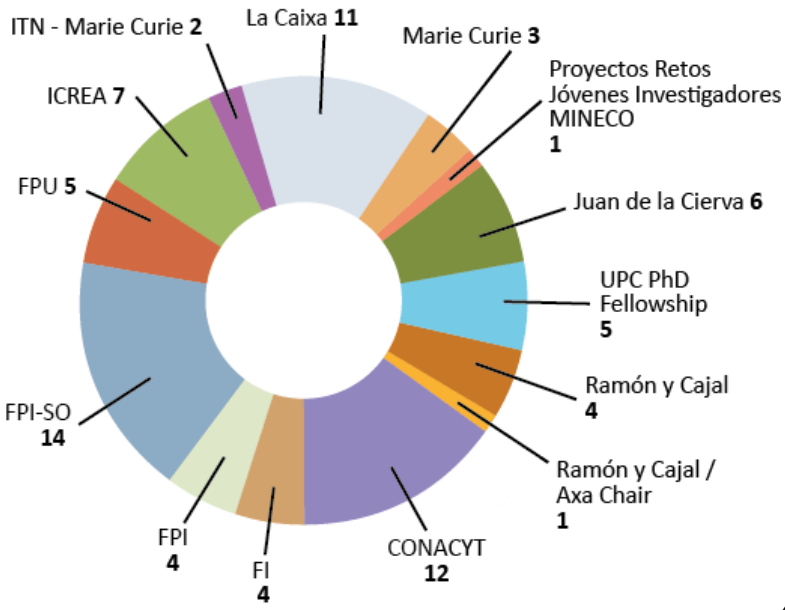
- Supercomputing services to Spanish and EU researchers
- R&D in Computer, Life, Earth and Engineering Sciences
- PhD programme, technology transfer, public engagement

BSC is a consortium that includes:

Spanish Government	60%
Catalan Government	30%
Univ. Politècnica de Catalunya (UPC)	10%

447 people from **44** countries *31^{th} of December 2015*





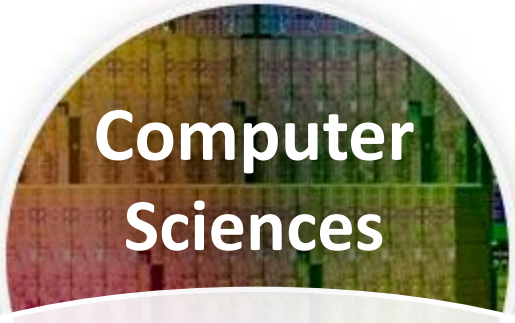
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The MareNostrum 4 Supercomputer

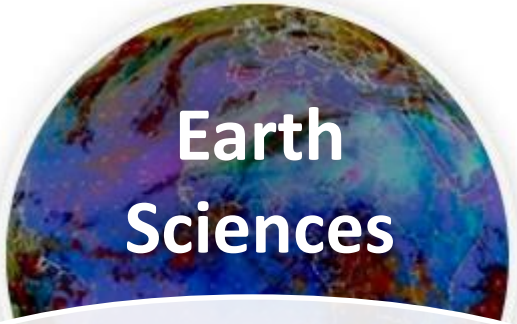
Over 10^{16} Floating Point Operations per second

Nearly **150,000 cores**	**331.8 TB** of main memory	**14 PB** of disk storage

Mission of BSC Scientific Departments



To influence the way machines are built, programmed and used: programming models, performance tools, Big Data, computer architecture, energy efficiency



To develop and implement global and regional state-of-the-art models for short-term air quality forecast and long-term climate applications

To understand living organisms by means of theoretical and computational methods (molecular modeling, genomics, proteomics)

To develop scientific and engineering software to efficiently exploit super-computing capabilities (biomedical, geophysics, atmospheric, energy, social and economic simulations)

BSC: A National Lab for Precision Medicine

Development and application of computational solutions for Genome Analysis in Biomedicine

ICGC-PanCancer

Nature 2011,
Nature Gen. 2012
Hum. Mol. Gen, 2012
PLoS Genetics 2012
Gut, 2013
Gastroenterology 2015
Nature Biotech. 2014
Human Mol. Gen. 2014
Nature Genetics 2014
Nature 2015
Nature 2016

Involved in international research consortia for genomics and disease

SMUFIN

Technology Transfer

Alliances with Hospitals and health foundations

BSC in the Health Care system. Pilot phase Prec. Med.

National Supercomputing Platform for Clinical Genomics
Research Lab. for Precision Medicine

Genome Sequence

Management of primary data Storage / Data Base

Genome Analysis Identification of variants

SNVs
SVs
Indels
CNV

Data Analytics Relational DataBase Functional Interpretation

Patient Care

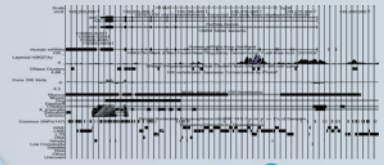
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Virtuous Circle for Precision Medicine

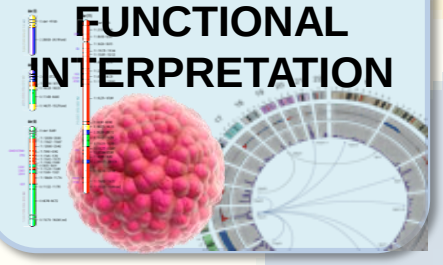
GENOME SEQUENCING

GENOMIC DATA MANAGEMENT

GENOME DATA ANALYSIS



CLINICAL AND FUNCTIONAL INTERPRETATION



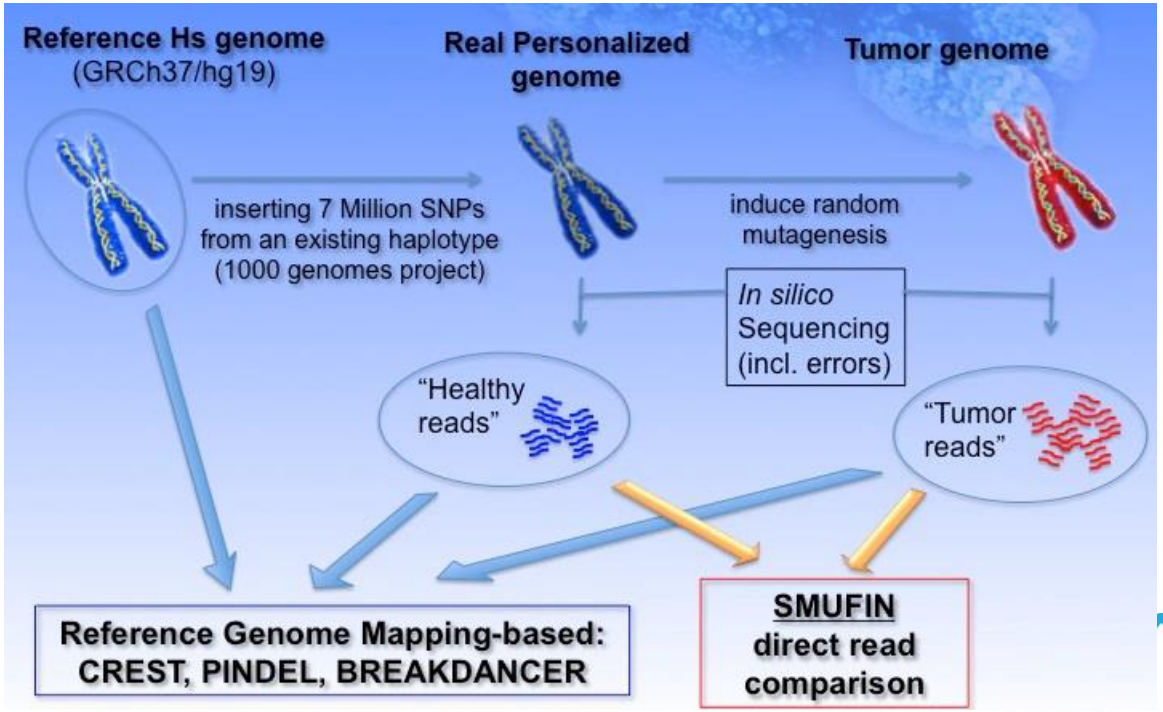
DECISION

HOSPITAL

Patient

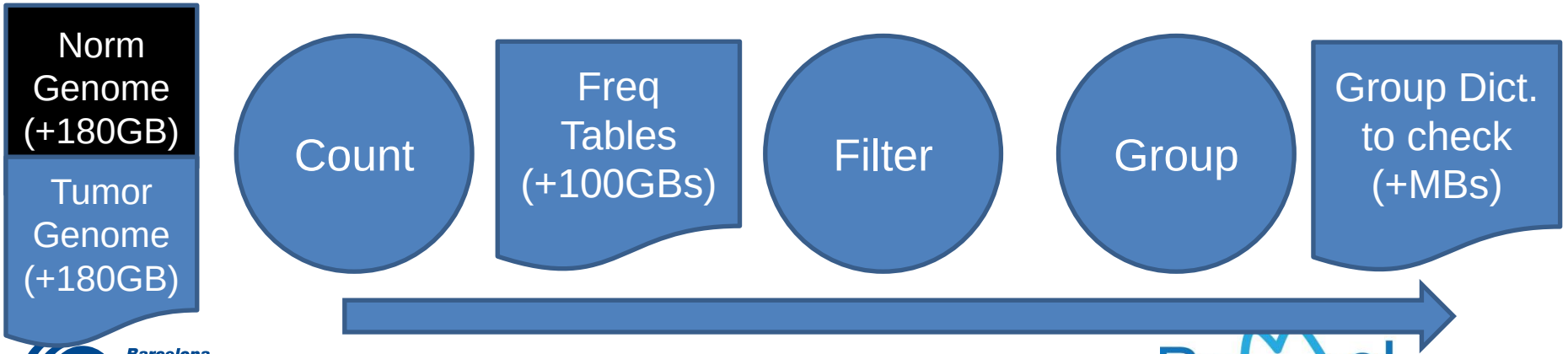
Smufin

- **S**omatic **Mu**tation **Fin**der
  - Identification and analysis of somatic mutations related to different diseases
  - Identify mutations on tumour genomes comparing them against the corresponding normal genome of the same patient



Smufin steps

- Identify tumor-specific reads
  - Build sequence tree using tumor and normal reads
  - Extract unbalanced branches
  - Group into read blocks; expanded by aligning corresponding normal reads
- Define and classify potential tumor variants
  - Small variants: SNVs and SVs within read length
  - Characterization of large structural rearrangements

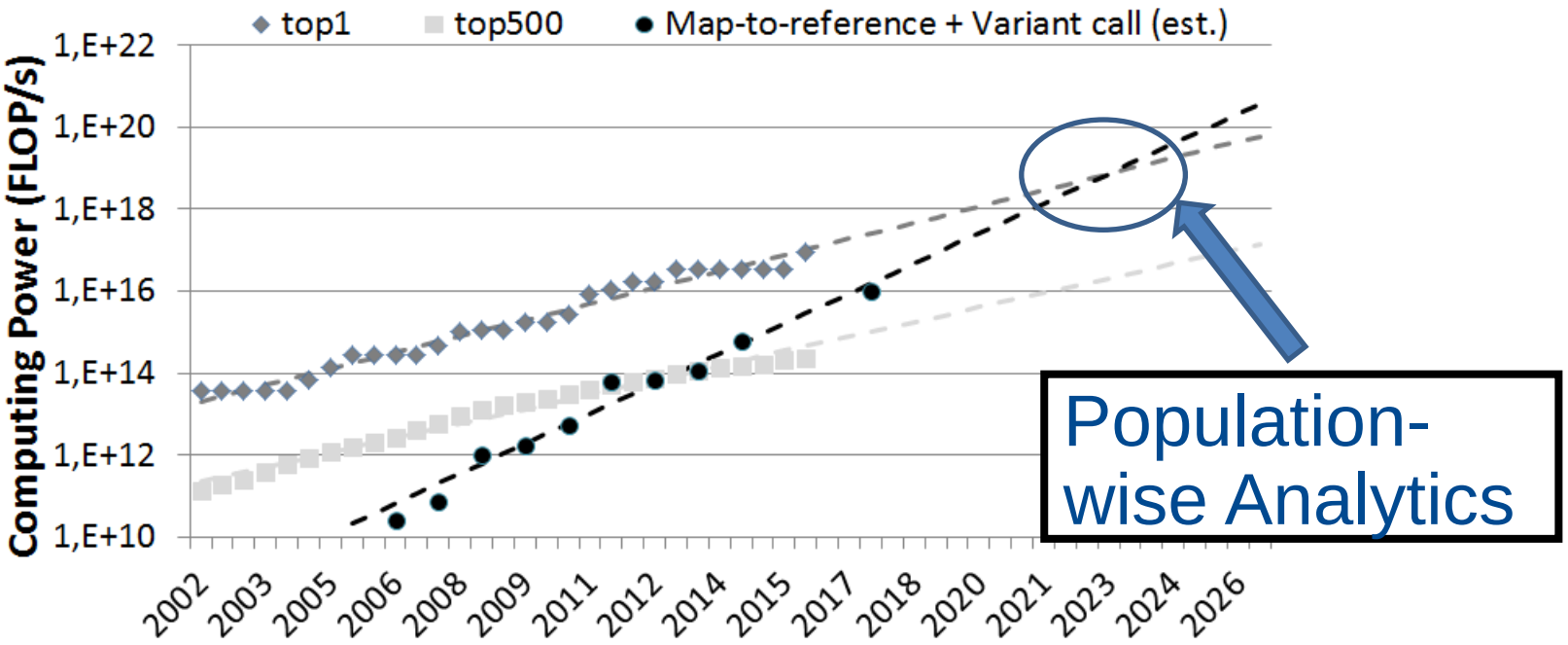


Smufin in numbers

- Inefficient execution on current processors:
  - 6 hours run on 16 Intel Xeon nodes (total of 256 cores)
  - Huge memory and I/O constraints
    - Input: 375 GB gzipped data
    - Reads: 4,288 million strings of length 80
    - Substrings of length 30 (in billions):
      - 218 (potential), 76 (actual), 14 (*interesting*)
    - Over 2TB of main memory requirements
  - Streaming pattern
    - 5-10x more loads than stores
  - Poor LLC locality
    - ~15% hit rate; ~5 MPKI

HPC Requirements of Genomics Data Analytics

Signifincat improvements (several orders of magnitude) are needed to enable population-wise genomics data analytics:
Better algorithms and HPC architectures



Source: www.top500.org and B. Berger et al., CACM'16

HPC Architectures for Genomics

- Data-centric architectures for genomics
  - Near-Memory or Near-Storage Computation
    - Pattern matching small reads on a huge data set in memory
    - Computation on very small integer data types (8 bits or less)
    - Embarrassingly parallel + data set distributed across nodes
    - MICRON's Automata; on-board FPGA; Active storage technology

HPC Architectures for Genomics

- Domain-specific Accelerators
  - GPGPUs to exploit data-level parallelism and high bandwidth
  - Vector processors
    - ISA extensions that fit well genomics workloads (AVX512, SVE, ...)
    - Explore long vectors for energy efficiency
  - Devise new accelerators for genomics workloads
    - Exploit on-chip FPGAs and build custom accelerators

Conclusions

- Genome sequencing is becoming faster and cheaper following an exponential growth
  - Population-wise sequencing will be a reality in the next 5-10 years
- Data analytics based on sequenced human genomes require a significant computation power and suffer inefficient execution (memory and I/O-bound)
  - Only relying on Moore's Law won't provide enough compute power to perform genomic data analytics at a population level
- Novel algorithms, HPC architectures and accelerators will be required to achieve such challenge

Thanks to…

- Computational Genomics research group at BSC
  - David Torrents (group leader)
  - Romina Royo
- Data-Centric Computing research group at BSC
  - David Carrera (group leader)
  - Jordà Polo

www.bsc.es

EXCELENCIA SEVERO OCHOA

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Are Next-Generation HPC Systems Ready for Population-level Genomics Data Analytics?

Calvin Bulla, Lluc Alvarez and **Miquel Moretó**

AACBB Workshop, 24/02/2018

HiPEAC

erc