# Parallel Computation & Genomic

Dominique LAVENIER
Irisa / Inria – Rennes
GenScale team leader

# Agenda

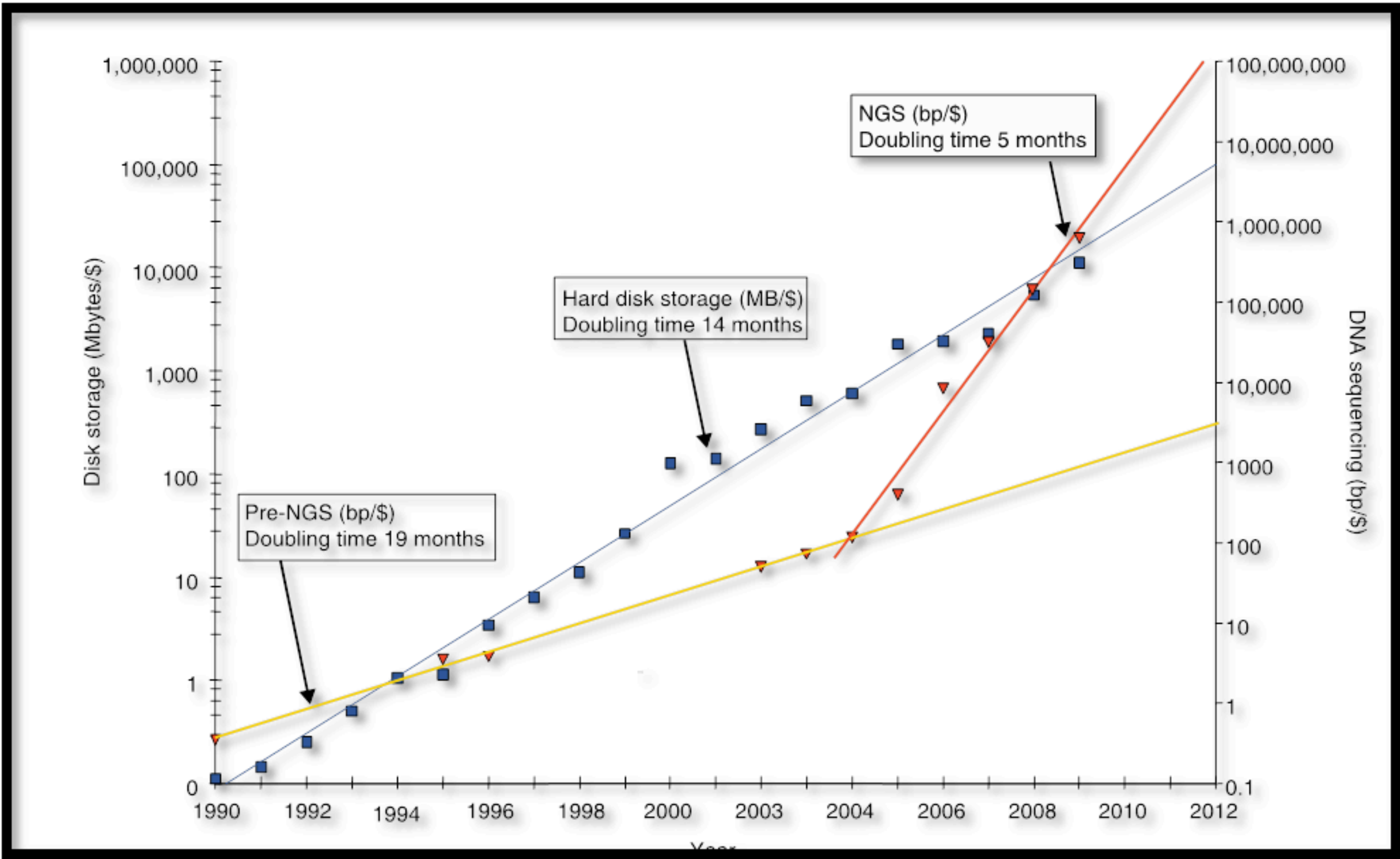
- Genomic data
- Applications
- Bioinformatics treatments
- Parallel implementation

# Genomic data

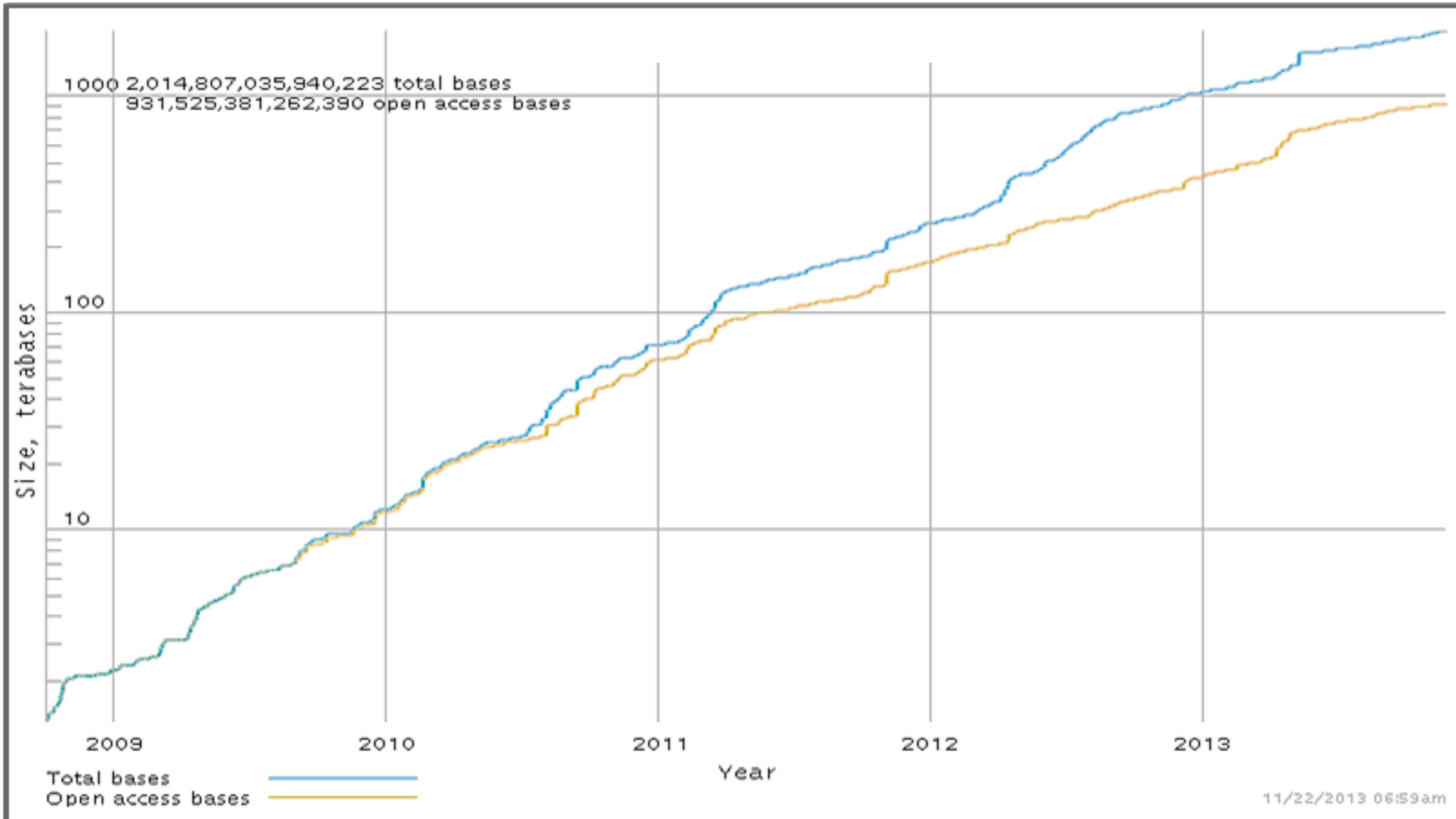
- DNA, RNA, (protein) sequences
- High throughput sequencing machine

# Historical trends in storage prices versus DNA sequencing costs

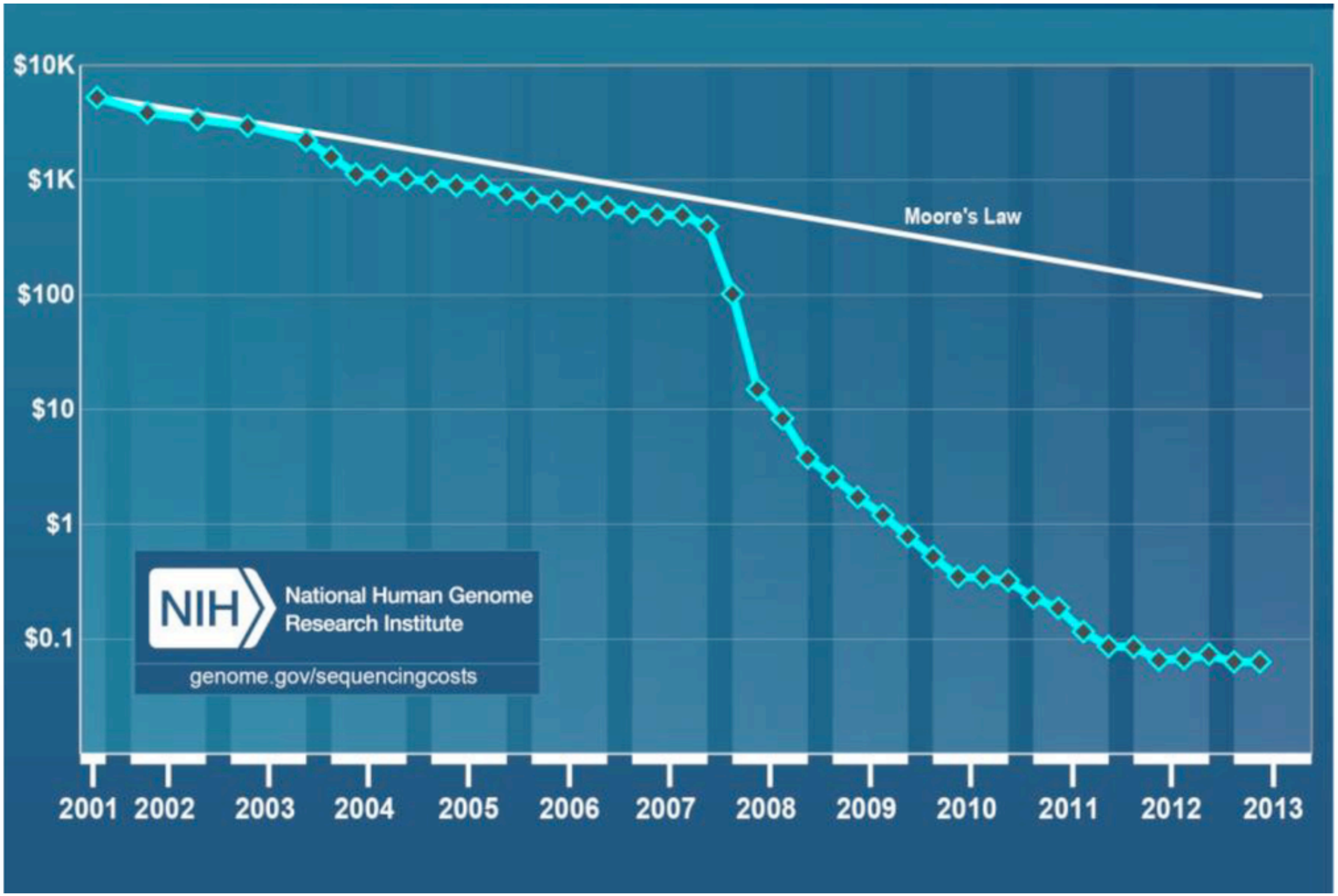
*(from Stein, L.D., Genome Biology 2010, 11:207)*

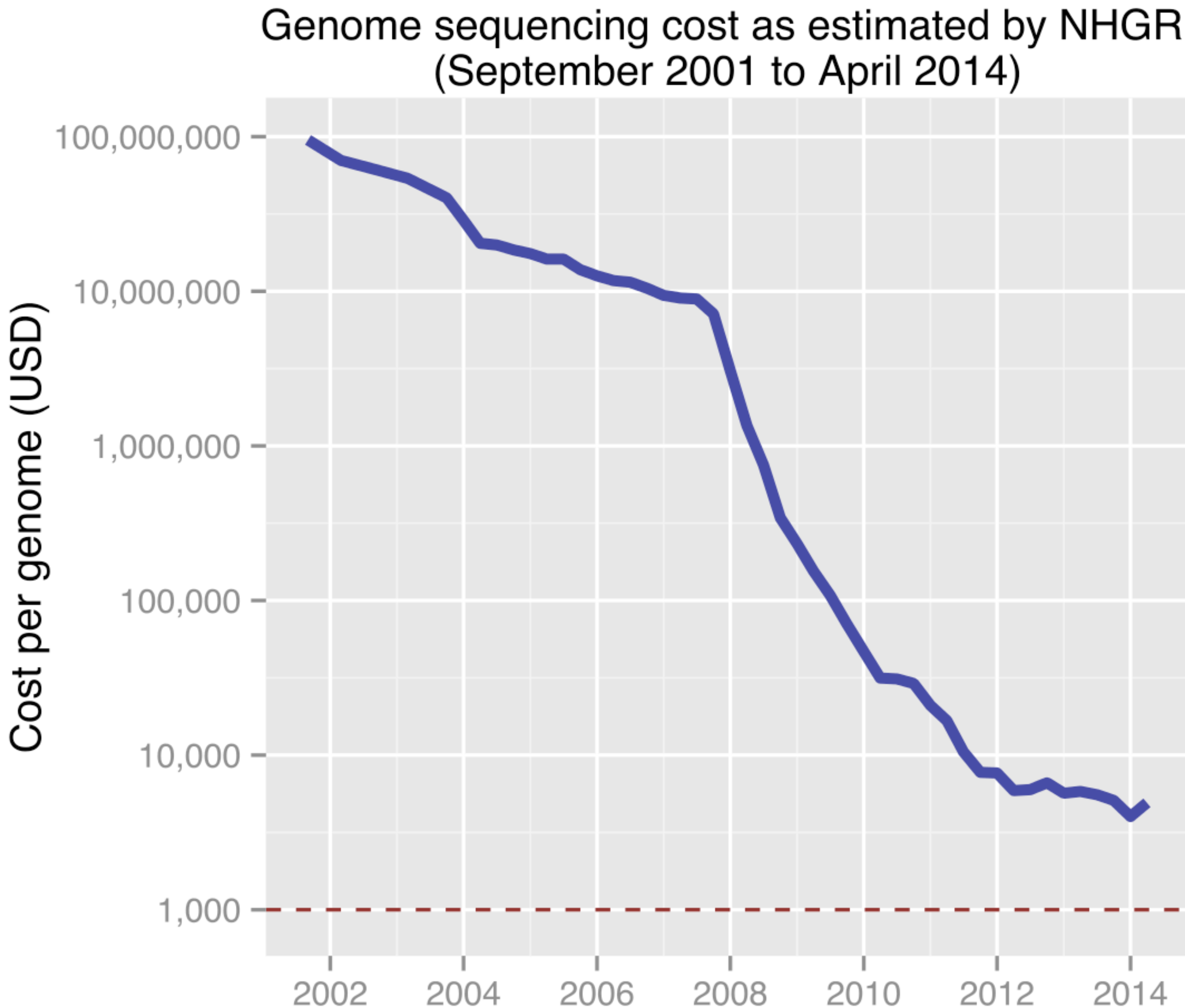


# SRA database growth



# Cost per raw megabase of DNA sequence




Genome sequencing cost as estimated by NHGRI
(September 2001 to April 2014)
Cost per genome (USD)
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
2002
2004
2006
2008
2010
2012
2014

# (bio)Technological breakthrough

- 10 years ago
  - Nearly sequential sequencing
  - A few DNA fragments sequenced simultaneously (~10-100)
  - DNA fragment size: 500 – 1000 bp
  - Low error rate
- Today
  - **Massive parallel sequencing**
  - Billions of DNA fragments sequenced simultaneously
  - DNA fragment size: 36bp – 150bp – 300bp
  - Very low error rate
  - 1 run → 0.1 to 1 TBytes
- Tomorrow
  - $10^6$-$10^8$ long/very long DNA fragments: 10 → 100 Kbp
  - very chip sequencing
  - High error rate

**DATA**
**X$10^6$-$10^7$**

# Applications (1)

- Biomedical
  - Drug design
  - Genomic disease
  - Personalized medicine
  - Diagnostic
    - example : cancer
      - Target sequencing (exome)
      - Detection of mutations in a set of predefined genes
      - Goal : match drug and gene mutation
  - ...

# Applications (2)

- Agronomy, Environment
  - Animal selection
  - Plant improvement
  - Diversity studies
  - …

# Metagenomic

> Seq 1
ATTGAGAGGACCATTG
> Seq 2
TGGACAGGAGGAGATA
> Seq 3
GCCATATGGACCCAGG
> Seq 4
TGGAAATATAGGGATA
> Seq 5
AATAGACCATTATTTC

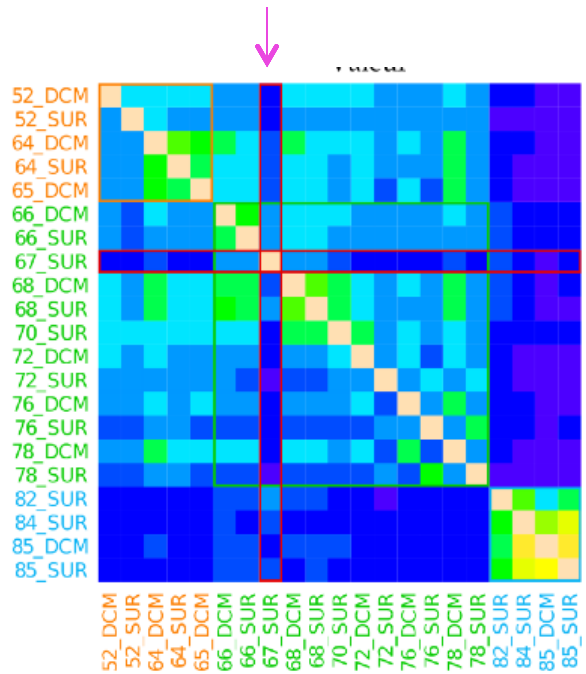
Simultaneous sequencing of all organisms of the same environment

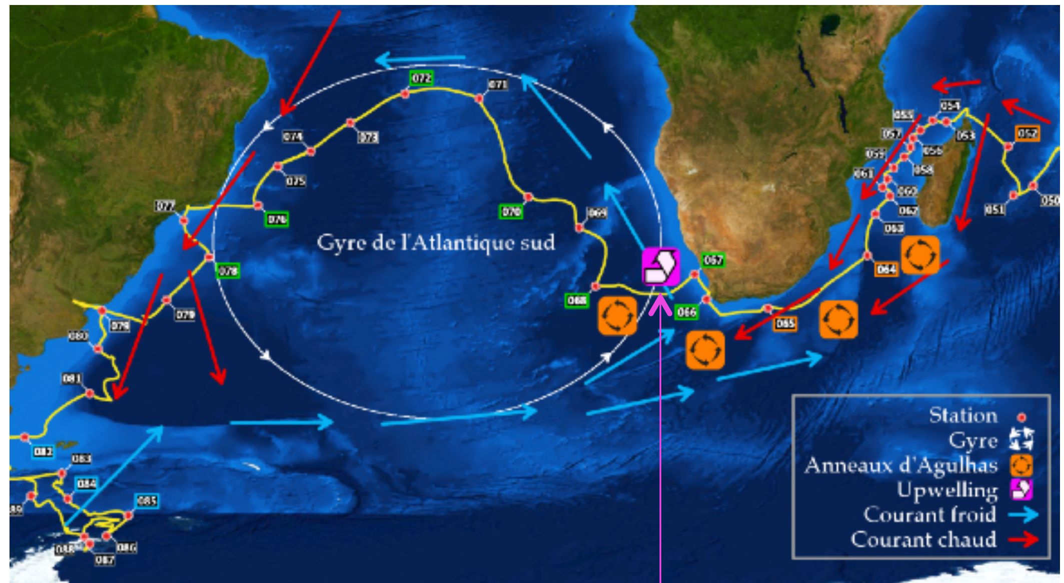
# TARA Oceans Project

Study of ocean streams

Analysis of 21 samples

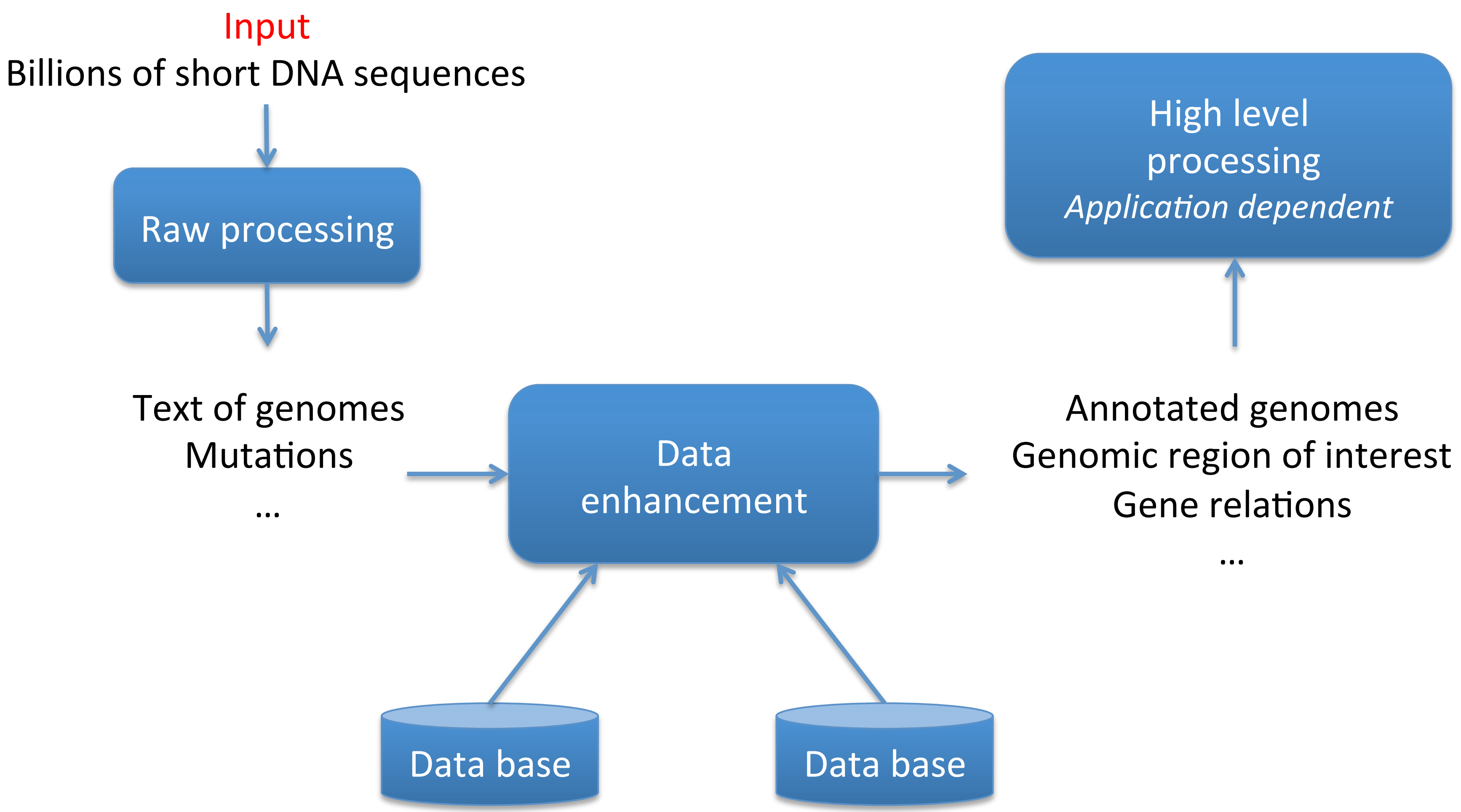
1 sample = $10^8$ reads



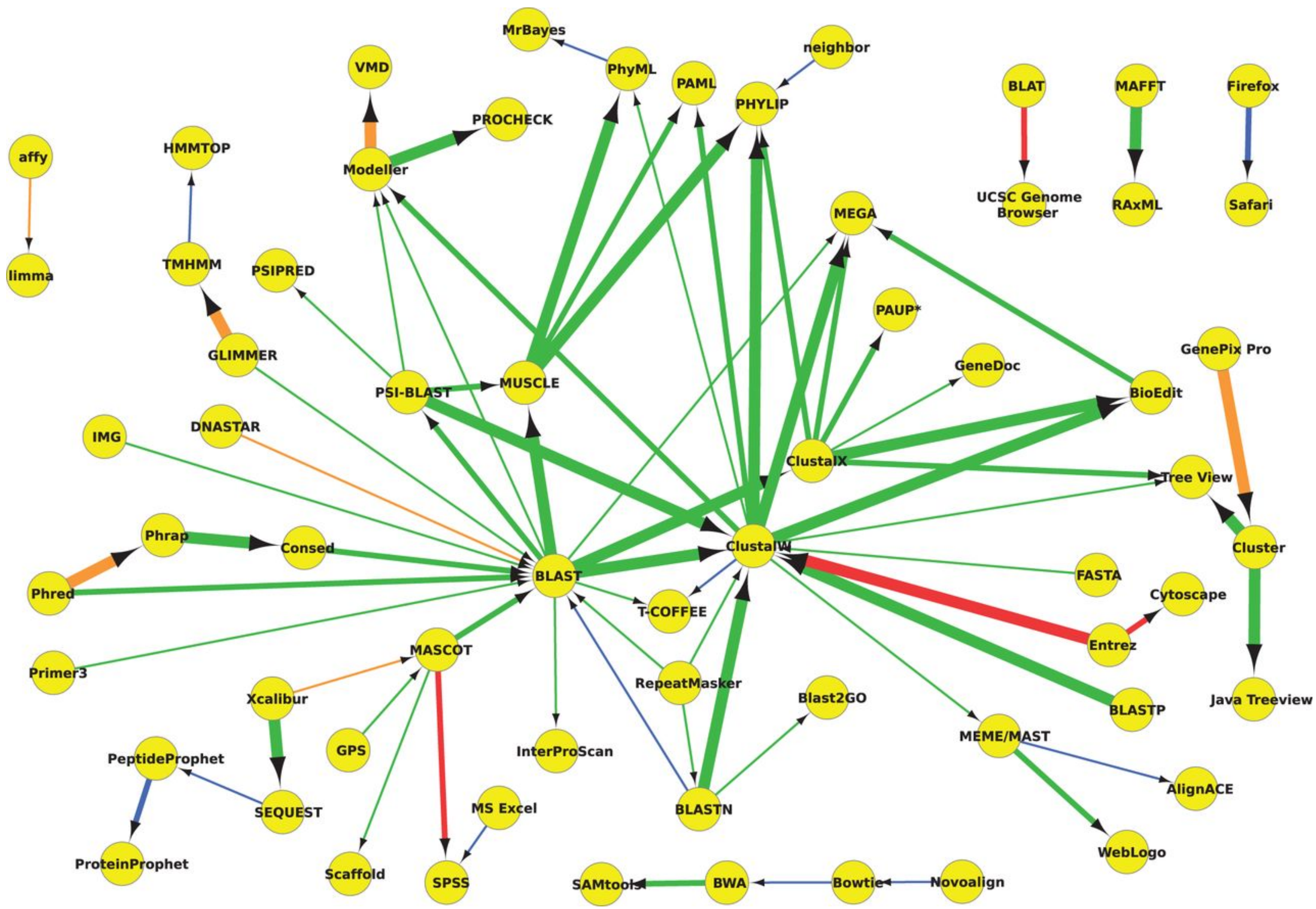




**Upwelling**
Cold and deep water
streams rise to the surface

# Bioinformatics treatments



## Usage network for software name resource pairs, mentioned within the methods section only.



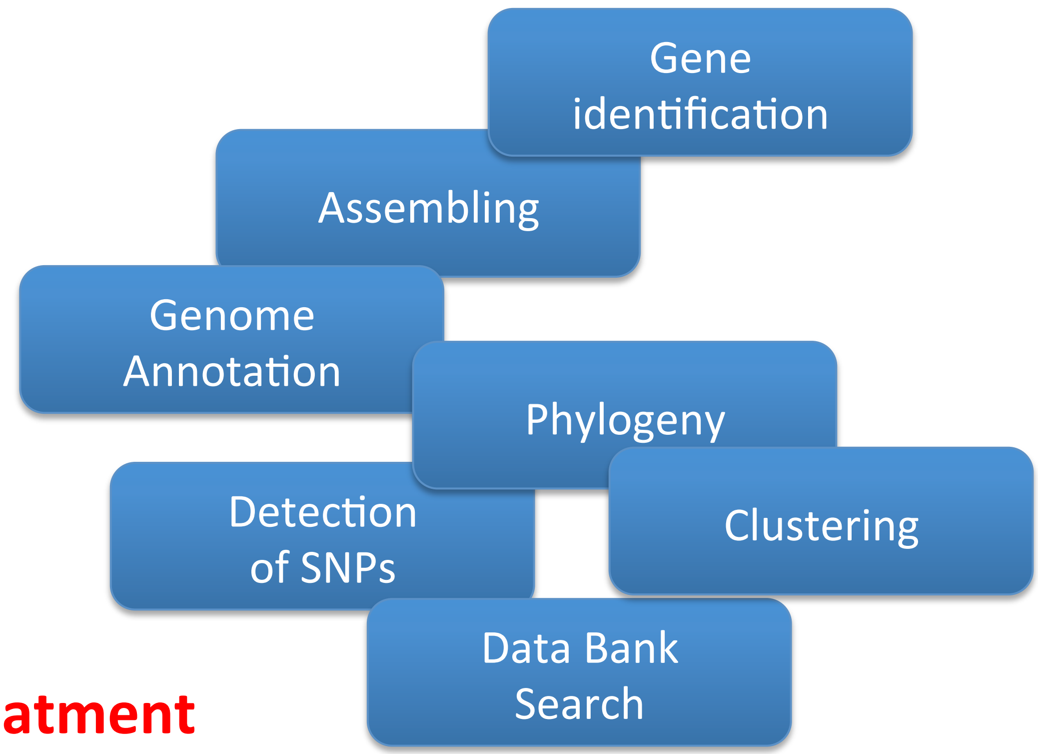

Duck G et al. Bioinformatics 2014;30:i601-i608



Bioinformatics

# Sequence comparison

- Declined in many ways:
  - Pairwise alignment
  - Multiple alignment
  - HMM search
  - Mapping
  - Detection of motif
  - …




**Basic Bioinformatic Treatment**

# Sequence comparison

- Alignment computation

```
A T T G C T G T C A A C G T T G G T A C A
| | |   | | |   |     | | | |   | | | | |   → SCORE
A T T A C T G A C - - C G T T A G T A C A
```

- Highly parallel process
  - N sequences vs M sequences

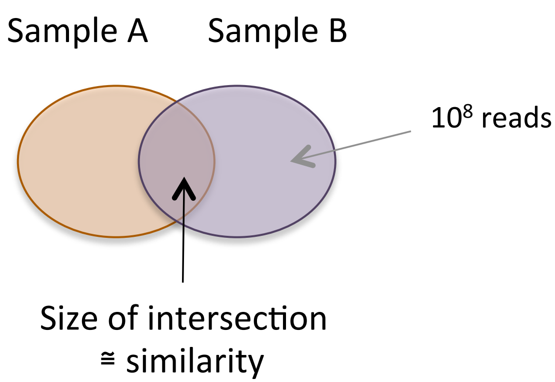
    ➔ NxM elementary comparisons

    ➔ independent processes

> Doesn't require floating point computation power

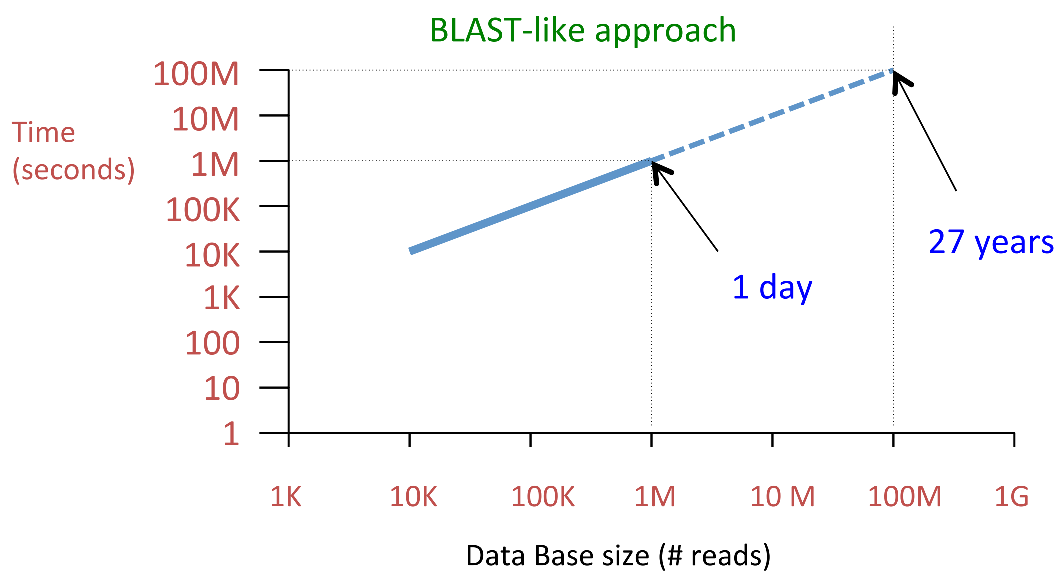
- Limits
  - Number of elementary comparisons to process

# Example: Comparative Metagenomic

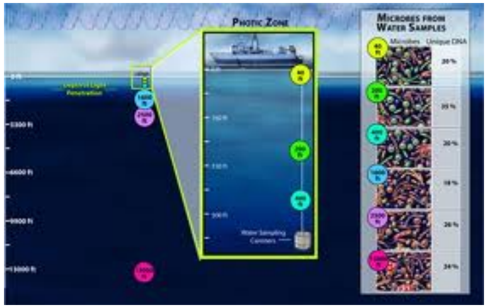



or 1 day with
10K node cluster

1 Metagenomic project
=
$10^2 - 10^3$ samples
➔ $10^6$ elementary comparisons



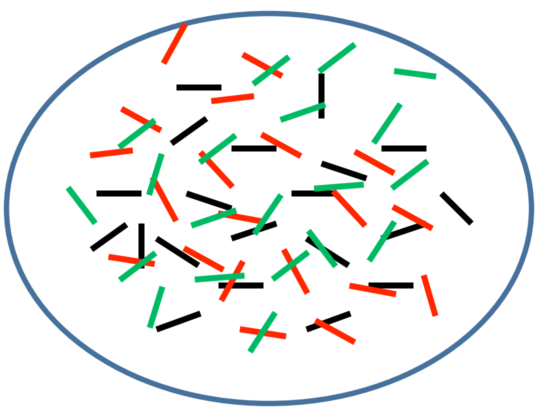
# Tara Oceans Project

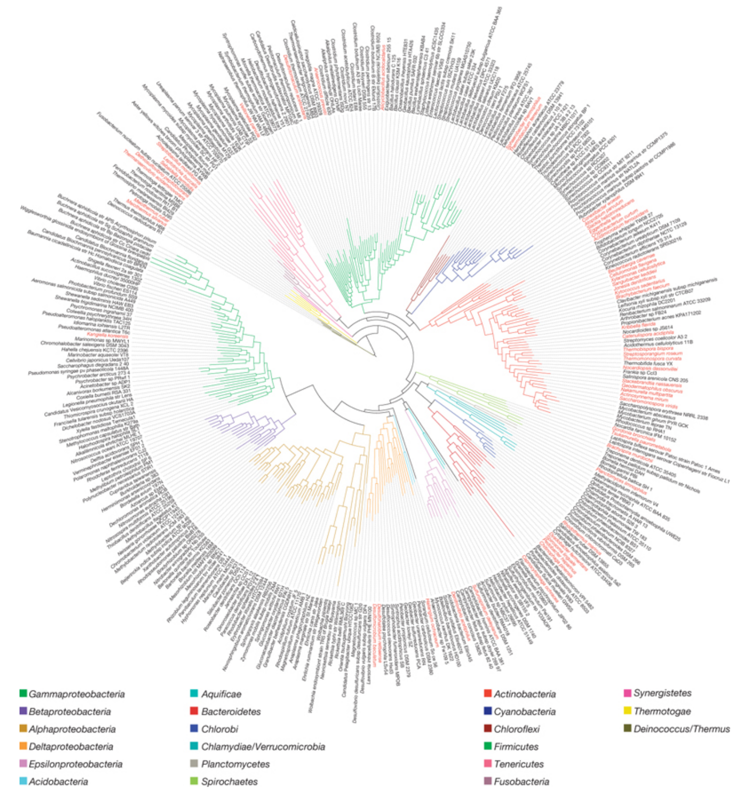


TARA
Oceans
Project

300 spots

$10^3$-$10^4$ species






Process of 3 only stations = 4 000 000 CPU hours

# How to speed-up these computations ?

- Software improvement
  - New sequence comparison algorithms
  - Specialization of applications
  - Data structures
- Dedicated hardware accelerator
  - ASIC, FPGA → parallel architectures


- Parallelism
  - Consider all levels of parallelism
    - SIMD → SSE instruction, GPU
    - Multi-threading → multi-core, many-core
    - Distributed computing → Cluster / cloud

# Custom Hardware Accelerators

- ASIC / FPGA
- Fine grained parallelization (algorithm level)
- Advantage
  - Significant speed-up
  - Low consumption
- Drawback:
  - Market niche → expensive
  - BLAST-like heuristic has not been yet efficiently parallelized at the algorithm level
  - I/O bottleneck (?)

# GPU

- Many bioinformatics algorithm have been implemented on GPU
- Modest speed-up (X2- X5) due to:
  - Data bandwidth, limited memory
  - SIMD programing restriction
  - Floating point capacity not used
  - No regular memory access
- Exception for some treatments
  - Computation requiring statistical analysis
  - Structural bioinformatics

Comparison with optimized multithreaded implementation

8-core processor
use of SSE instructions

I/O << computations

# Multicores

- Efficient implementation by combining SSE instructions and multi-threading
- Algorithms does not scale well with the increase of processors
  - Many irregular accesses to the share memory
    - →bad news for many-core architectures ☹
- Most current bioinformatics software support a multi-threaded implementation

# Clusters

- Data parallelism
- Time consuming bioinformatics processes based on sequence comparison can be easily parallelized
- Limitation :
  - Reorder large set of data
  - Data access to storage devices
    - Network bandwidth is often the bottleneck

# Conclusion

- More and more genomic data
- Bioinformatics treatment features
  - Dominated by data
    - Large volume of data
    - Low computation complexity
      - I/O and memory data access is often the bottleneck
    - No floating point computation
- Parallelism
  - Multi-core (SSE + multi-threading)
    - Scaling to many-core won't be straightforward
  - Cluster
    - Need infrastructure (network, device storage) adapted to handle large data flow