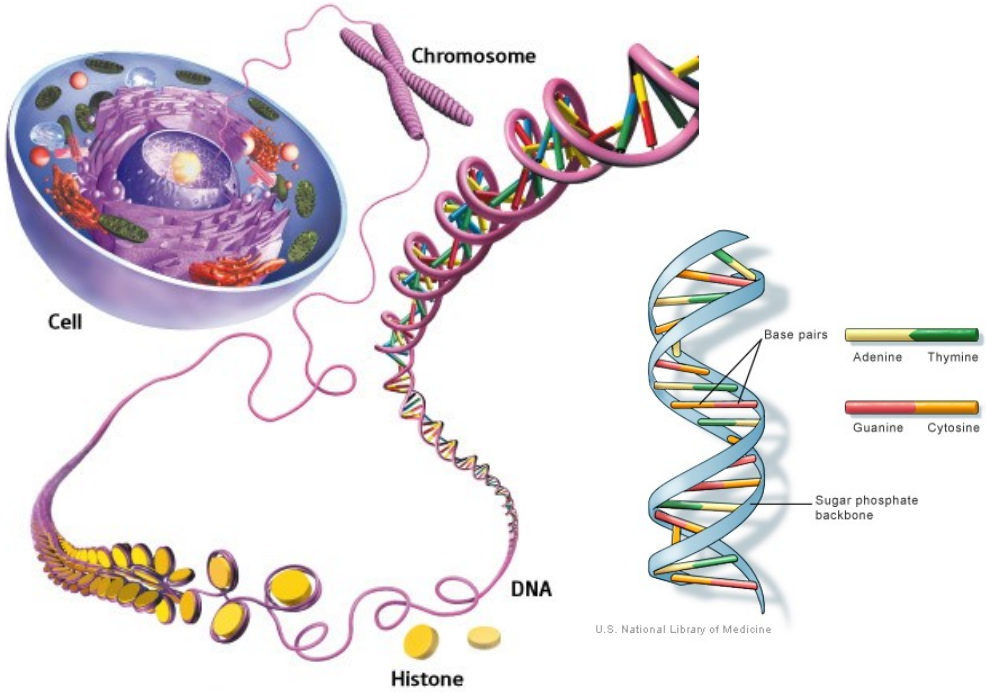
Data management for Dutch NGS Institutes

Next Generation Sequencing (NGS)

How does it work?



Illumina HiSeq 2000

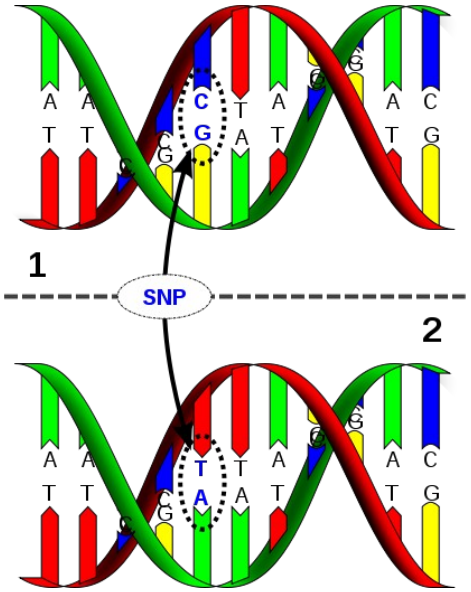
```
ACTGAA ... TTTGCGC
GCATCC ... AATTGCG
...
TCGAAG ... GCGCATG
```

millions

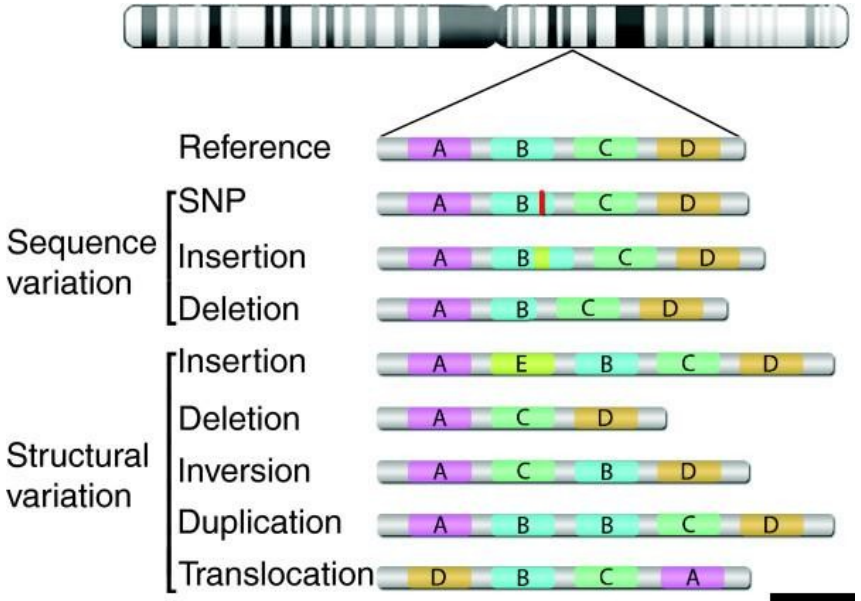
30 – 400 bases

Next Generation Sequencing (NGS)

What are we looking for?



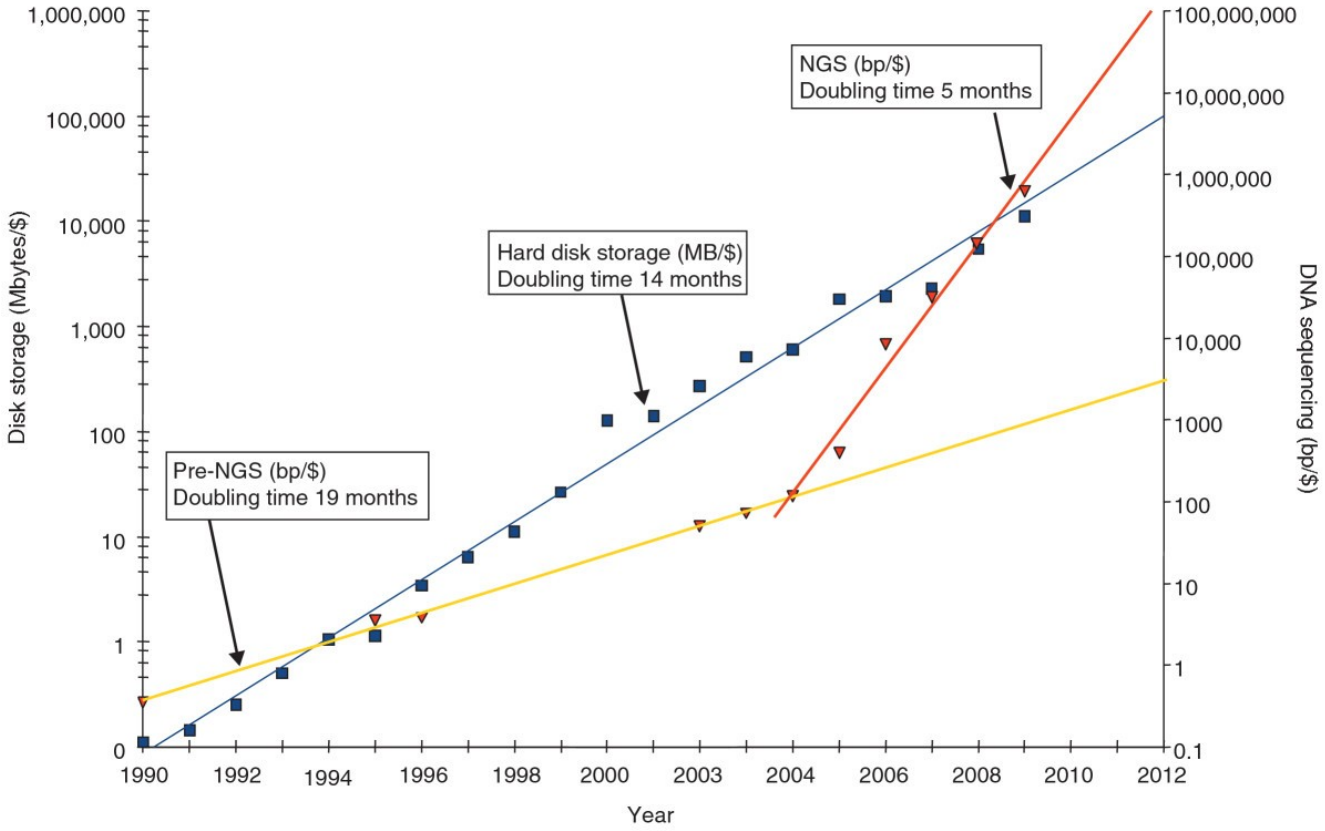
www.ahmedabdelhamid.com



Genetic determinants of phenotypic diversity in humans, Rahim *et al.*, Genome Biology

NGS data explosion

NGS data grows faster than storage



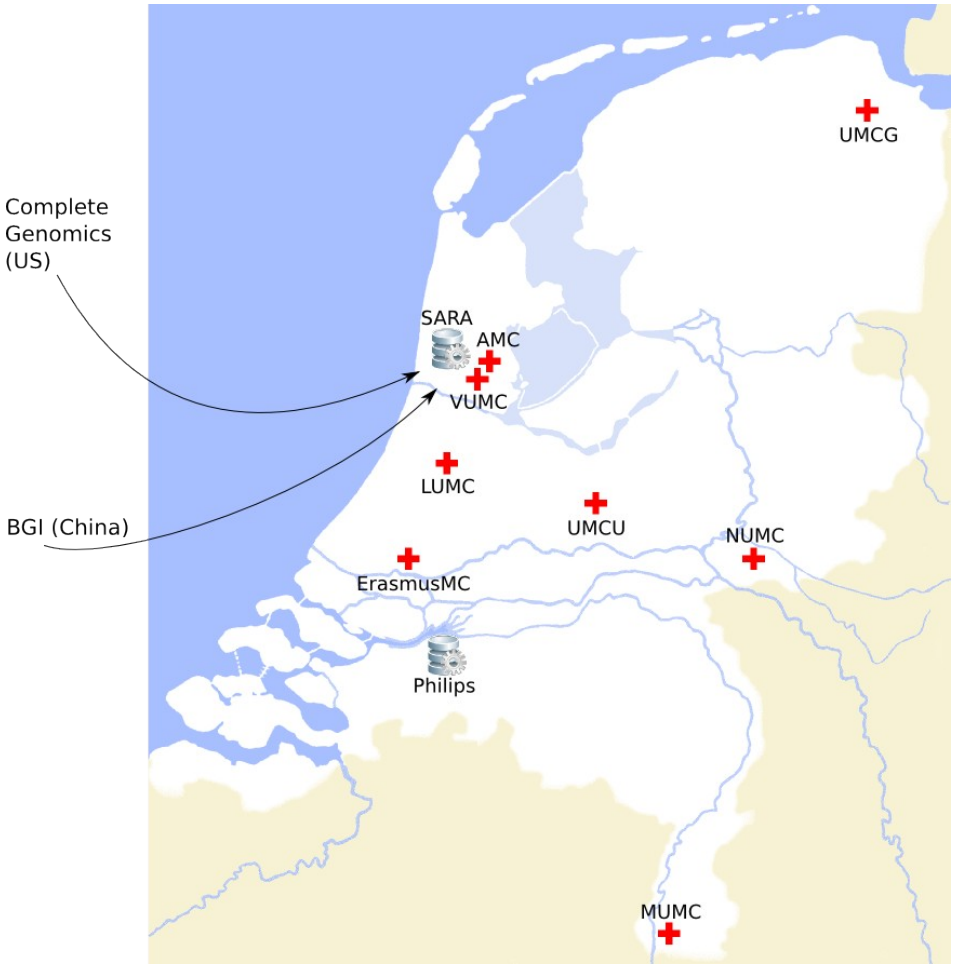
The case for cloud computing in genome informatics, Lincoln Stein, Genome Biology, 2010

Data generation

Data locality

Data is generated at many different locations in the Netherlands, complicating data management and compute resource utilization.

Research partners abroad such as Complete Genomics & the Beijing Genomics Institute also provide sequences and deliver the data in bulk.



Data generation

The data re-analysis problem

First analysis done locally, problems occur at a later stage:

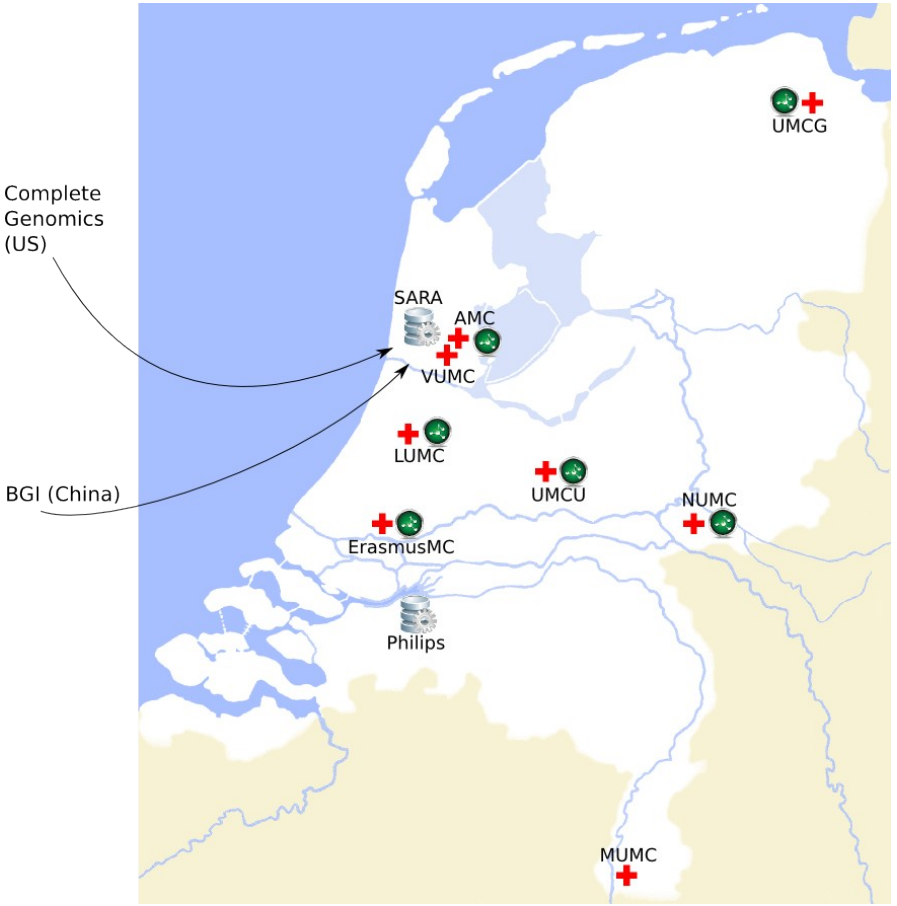
Realignment
- New reference genome
- Better algorithms

Reevaluation
- Additional data

Utilizing grid resources to share data

Dutch Life Science Grid

- Many of the UMCs already have a cluster connected to the grid
- All clusters have their own SRM
- External connections and local network are the bottlenecks
- Existing connections are not secure



NGS data management

Current challenges

Sharing data with partner projects

- Overcome transfer hops due to incompatible protocols (*e.g.* aspera)
- Centralized data ingestion facilities (Amazon like)
- Easier SRM access

Compute Challenges

- Requesting the appropriate compute resource multi-core, available disk space and memory

Dynamic lightpaths

A future project

Connect existing Dutch infrastructure and international partners to compute & storage sites

1 Gbit/s dynamic lightpaths

- Data from each local site comes in burst
- Data from international partners comes in streams

Discussion

- Bioinformaticians are taking their first steps in data intensive (grid) computing
- Transfer of that data to centralized facilities is vital for their success
- Connecting medical centers using dynamic lightpaths seems a viable solution

Thanks

TU Delft
- Marcel Reinders

SARA
- Coen Schrijvers
- Tom Visser
- Sander Boele

Surfnet
- Nicole Gregoire
- Michiel de Vos
- Gerben van Malenstein

Current data delivery

Ingestion is a challenge