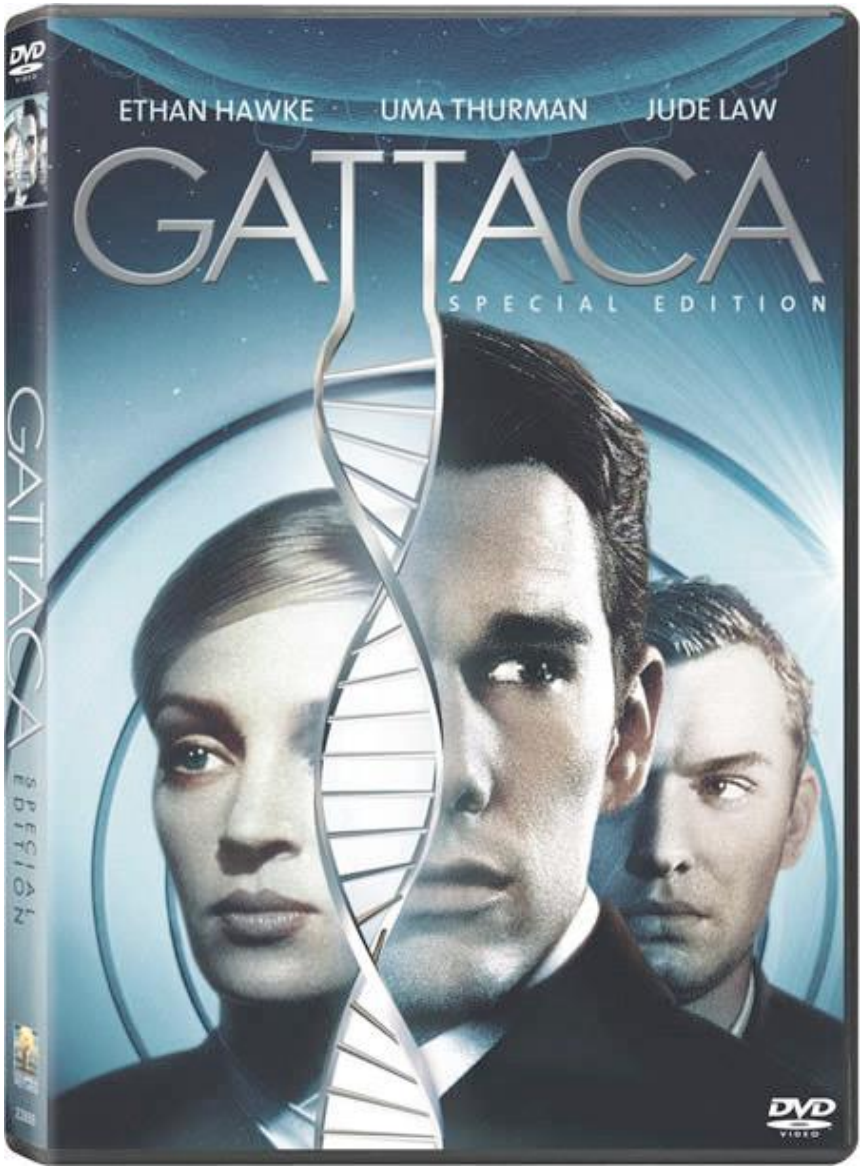
Need to prepare to get everybody sequenced in the future : from womb to tomb MPEG the movie of your life

Prof. Ioannis Xenarios
Center for Integrative Genomics (UNIL)
Departement of Biochemistry (UNIGE)
SIB Swiss Institute of Bioinformatics / Swiss-Prot&Vital-IT groups

SIB
Swiss Institute of Bioinformatics

elixir SWITZERLAND



http://www.imdb.com/title/tt0119177/

People at Vital-IT

HPC

Computational Genomics

Computational Systems Biology

Scientific vizualisation

Metagenomics Metabolic Models

Neuro-omics

Evidence-based BioMedecine

OncoGenomics Personalized Genomes


Roberto Fabbretti


Christian Iseli


Nicolas Guex


Robin Liechti


Marco Pagni


Jérôme Dauvillier


Mark Ibberson

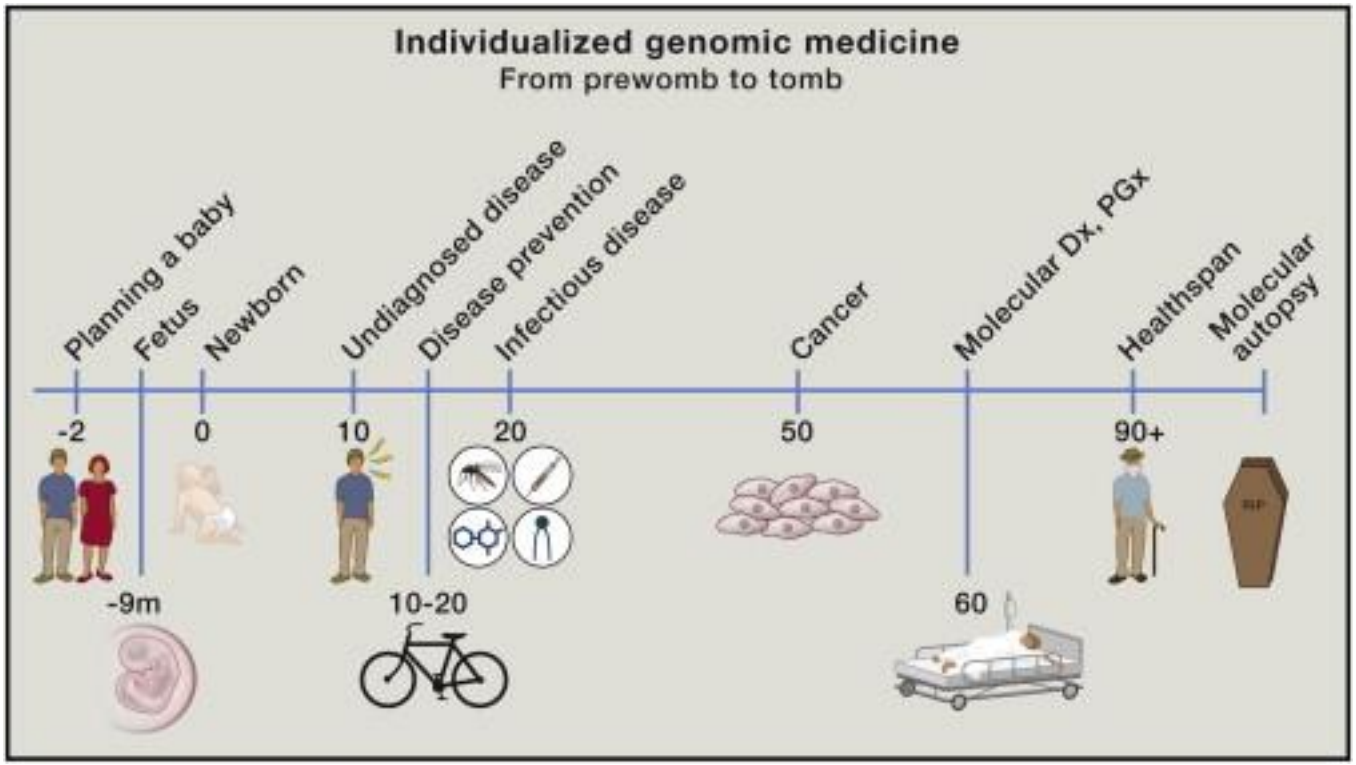

Brian Stevenson

hardware software development data analysis

SIB Swiss Institute of Bioinformatics

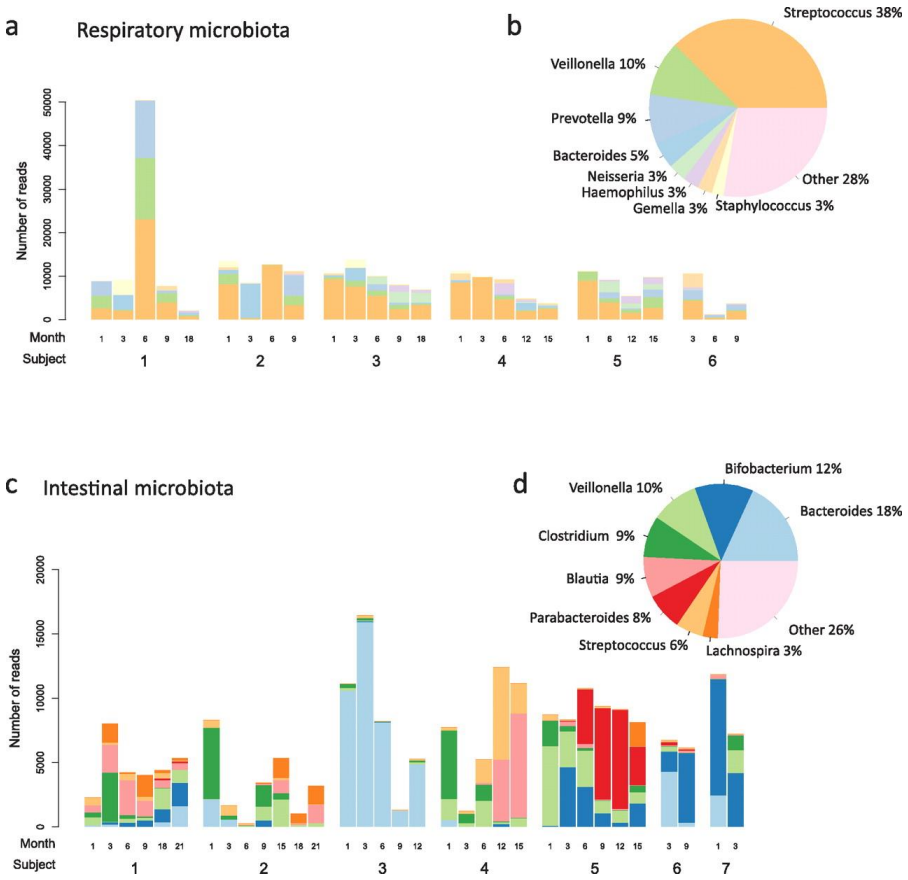
Case study – vision for 2020 (« dream/nightmare »?)

- Everybody will be sequenced at least once **in his/her** lifetime
  (from womb to tomb – Eric Topol Cell 2014)



Swiss Institute of Bioinformatics

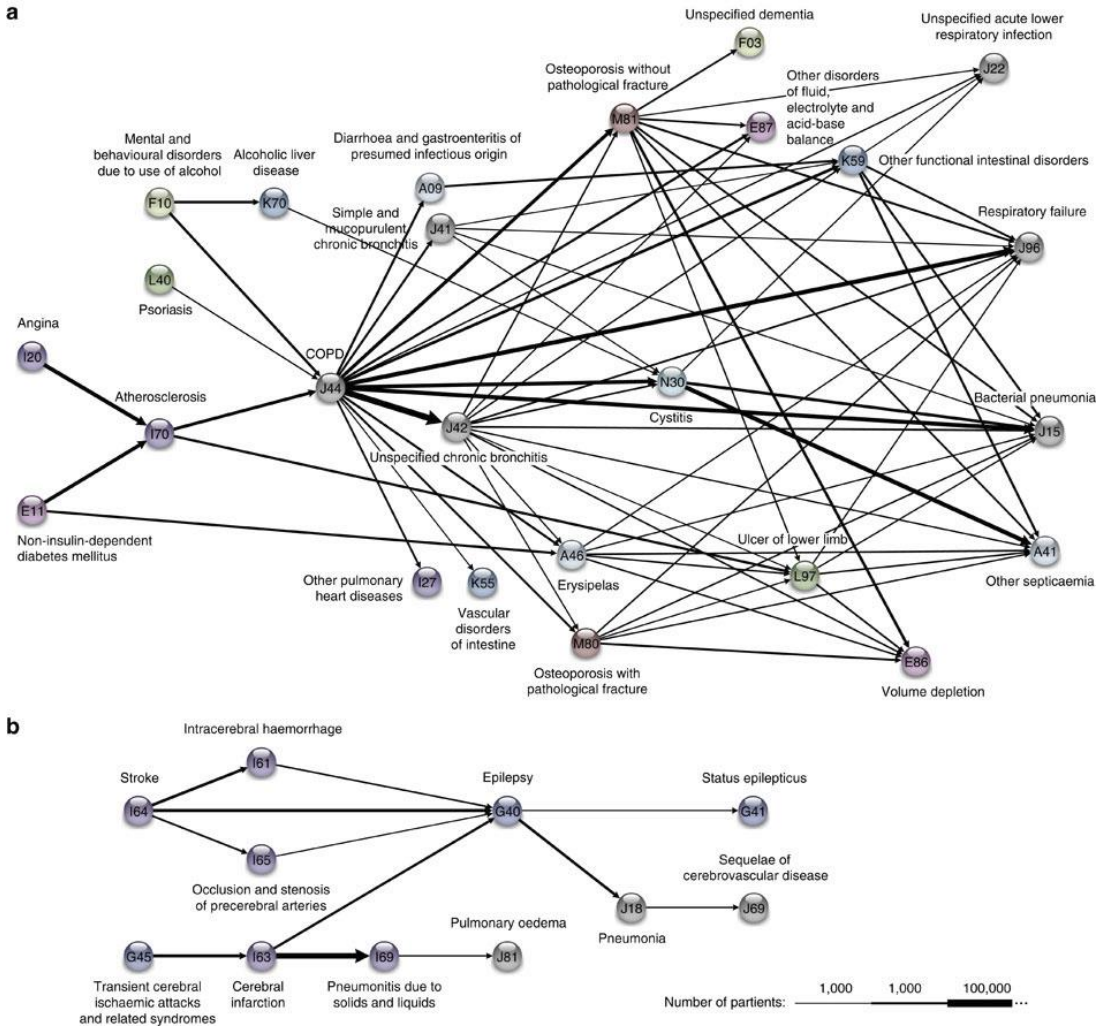
Our genome
+
all the bugs that lives « with/on/into » us



Madan JC et al.
Serial Analysis of the Gut and Respiratory Microbiome in Cystic Fibrosis in Infancy: Interaction between Intestinal and Respiratory Tracts and Impact of Nutritional Exposures, Mbio July/August 2015, volume 6 issue 4

Swiss Institute of Bioinformatics

We are following disease trajectories

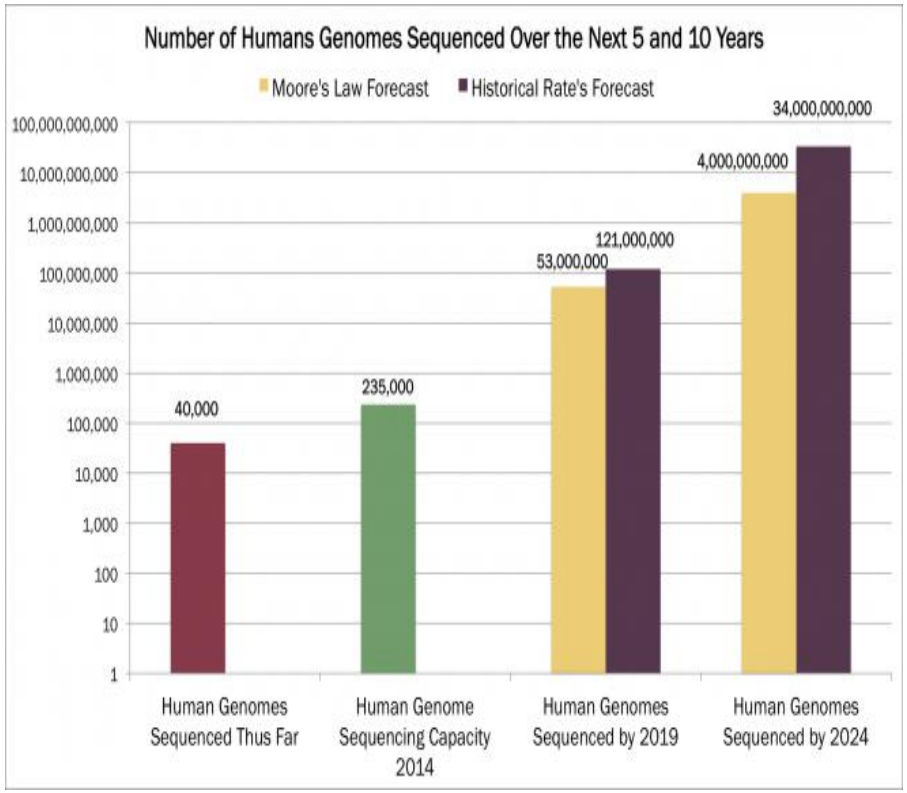


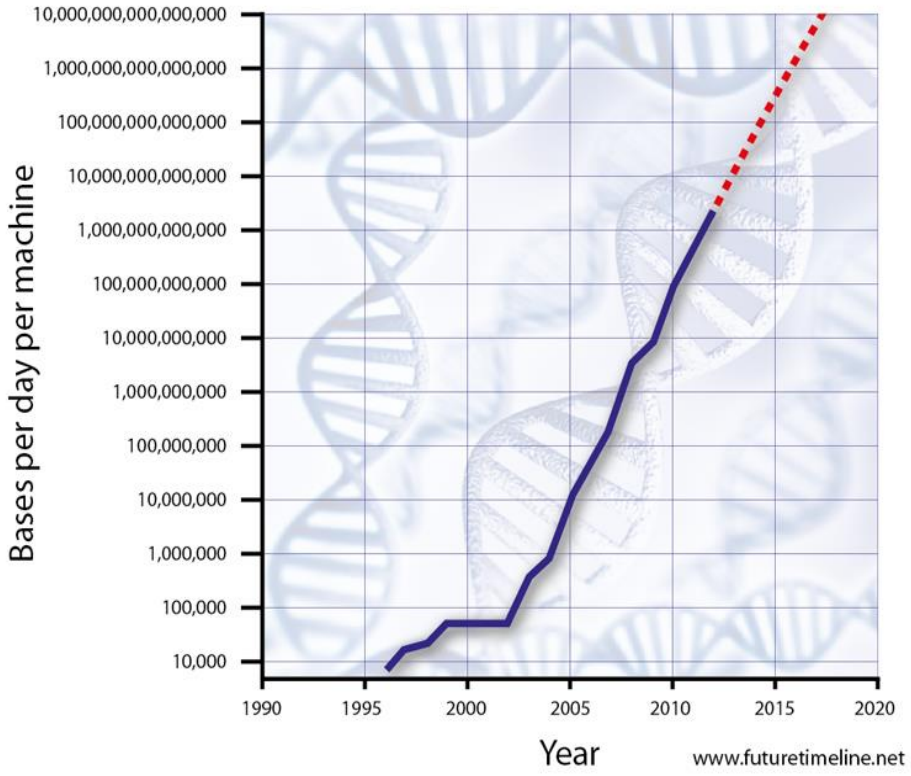
Anders B, et al, Nature Communications 5,Article number:4022 doi:10.1038/ncomms5022

Swiss Institute of Bioinformatics

The large IT costs (storage, bandwidth) are today a major obstacle to the explosion of genomic medicine.

The huge amount of expected data in the near future will allow sequencing entire populations, sharing data across institutions.





Swiss Institute of Bioinformatics

Alzheimer Disease Neuro Imaging : ADNI

Alzheimer's Disease Neuroimaging Initiative

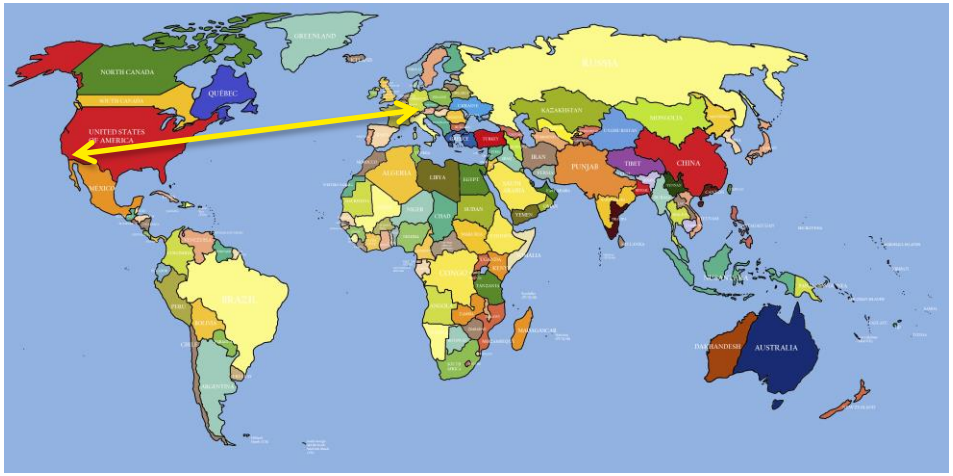
ADNI	809 199 AD 359 MCI 251 CN	Whole Genome Sequencing data 2.5 M SNPs	3 trillion pairs	Cerebral ventricle volume

Shipped Sept. 2014

85 Terabytes

Sequence reads aligned on reference genome (hg18)

Variant call



Swiss Institute of Bioinformatics

Platform for Advanced Scientific Computing (PASC)

PoSeNoGap : Portable Scalable Concurrency for Genomic Data Processing.

The main aim of the project is to develop for the Swiss Platform for Advanced Scientific Computing a new computation node composed of **heterogeneous** hardware, a new compression format for genomic data and a software infrastructure that enables emerging applications such as Genome Analysis to be able to process extremely large volumes of genome data in an efficient and timely way for scientific and diagnostic purposes (clinical application).

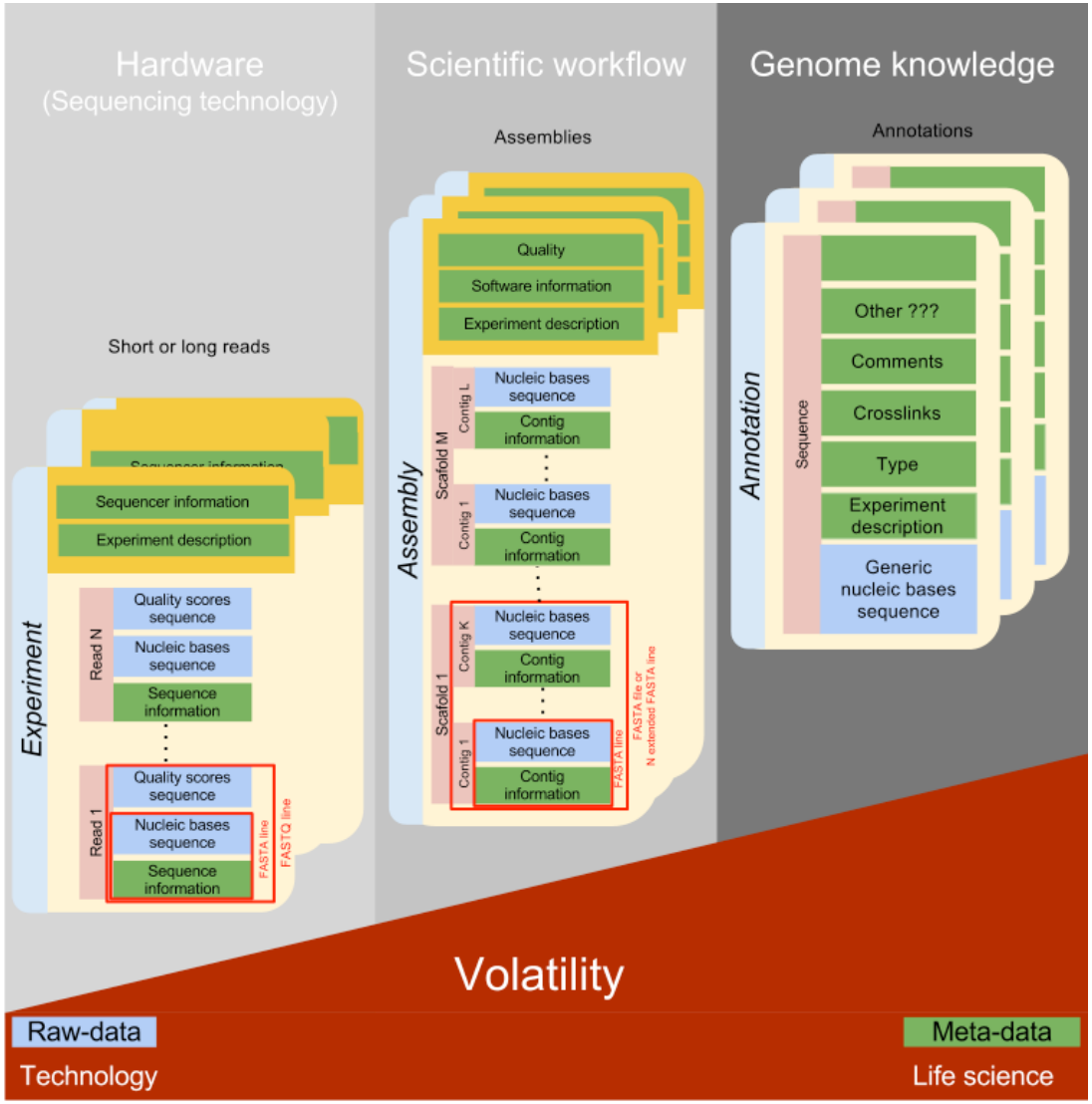
Ioannis Xenarios
SIB- (Vital-IT)

Marco Mattavelli
EPFL - (MPEG)

Yann Thoma
HES (FPGA)

Swiss Institute of Bioinformatics

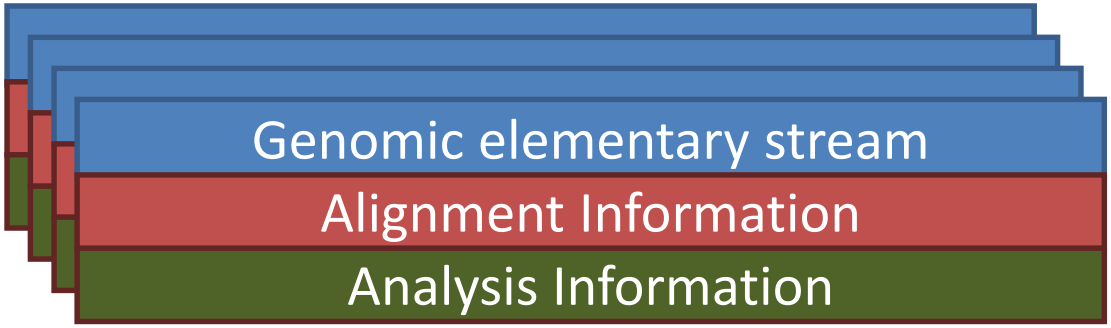
The pillars from technolog(ies) to interpretation(s)



Swiss Institute of Bioinformatics

Tools in genomics

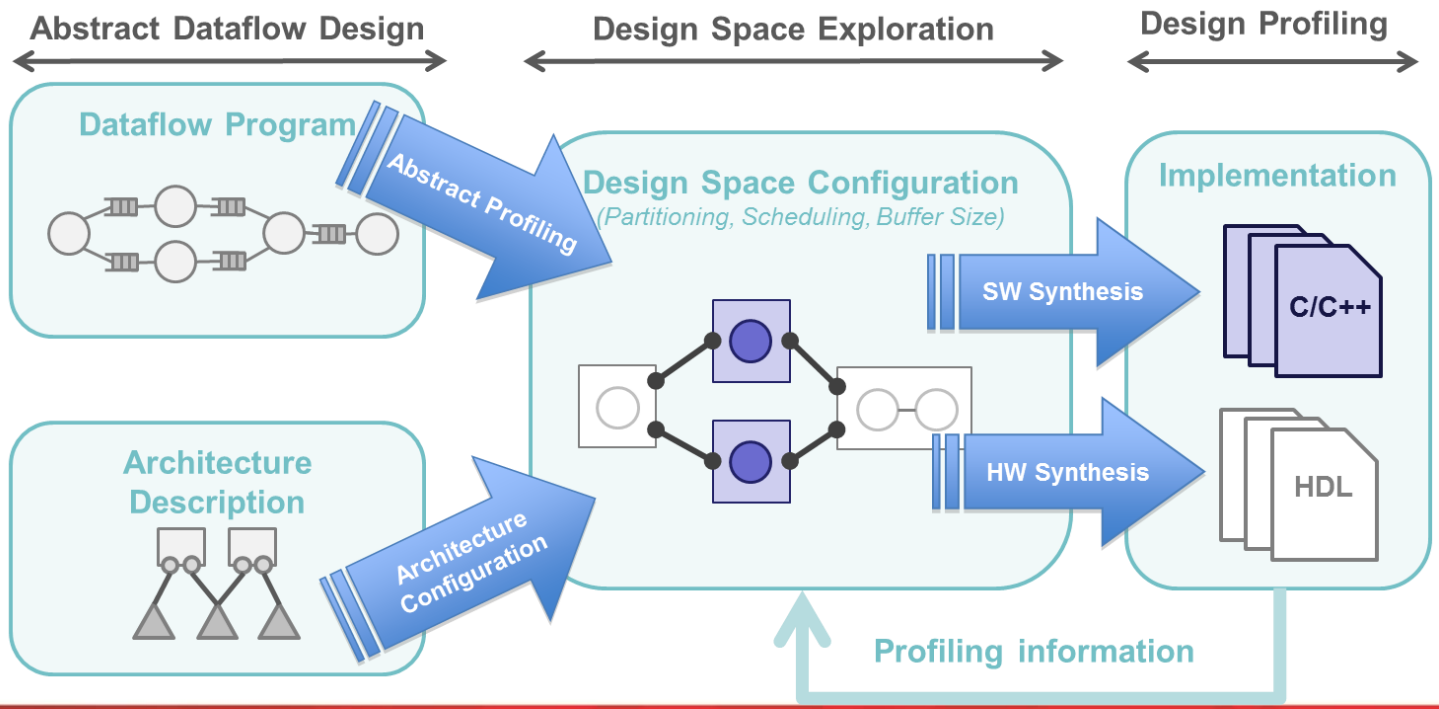
Structured data for selective access and streaming



Raw data (FastQ)

Aligned/Mapped data (BAM)

Downstream analysis (VCF, etc.)

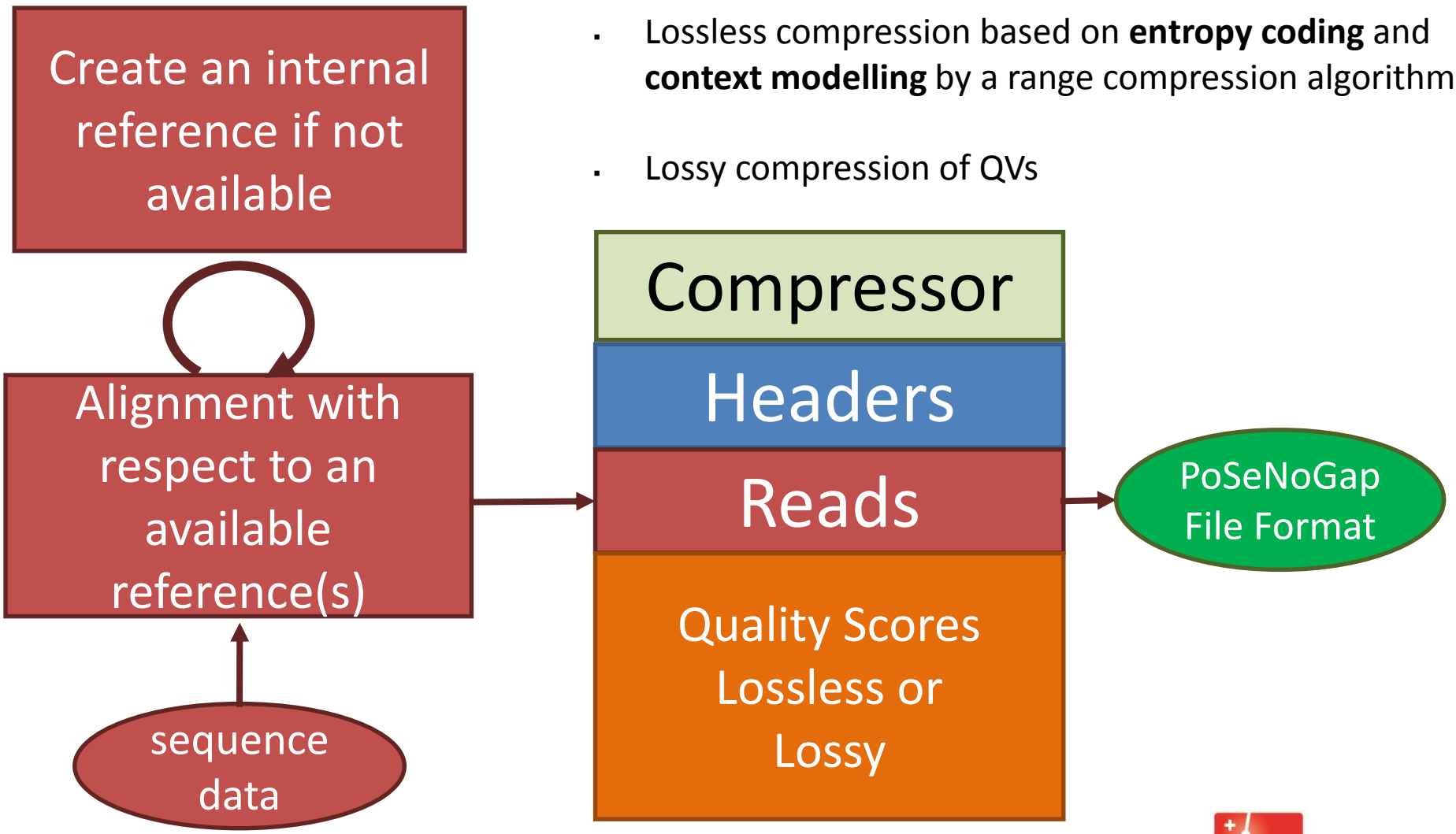


Innovative SW/HW co-design
High-level analysis
Automated code synthesis
Design Space Exploration
Code partitioning
Efficient porting to parallel architectures

SIB Swiss Institute of Bioinformatics

Encoding strategies

- Reference based compression of sequences
- Lossless compression based on **entropy coding** and **context modelling** by a range compression algorithm
- Lossy compression of QVs



Current Challenges

The number of projects that use NGS is exploding.

Sequencers deliver output as fastq files

```
LP6005115-DNA_G11_R1.fq   (150GB)
LP6005115-DNA_G11_R2.fq   (150GB)
                          =======
                           300GB
```

paired-end 2 x 656,935,771 **= 1.31 billion reads**

For the ADNI project alone **> 800 patients.**
300Gb x 800 = 240000 GB **= 240 TB**
(BAM file– 84Tb)

variant calling from a single sample (LP6005115-DNA_G11)
~7 days to complete using bwa/Picard/GATKv3.2 pipeline

Swiss Institute of Bioinformatics

Conclusions

• Individuals **will be sequenced several times** over their entire live, the ability to compare and evaluate along a timeline will become critical

• Compression-based algorithms are **a must** but should not require heavy reprocessing it should include **academic and industry stakeholders**

• The landscape in sequencing are rapidly evolving providing robust mechanisms to represent genetic information is a must with the **known/known, unknown/known and the unknown/unknown**

•MPEG has a long standing expertise in maintaining standard and also could serve as a basis to maintain the **genomic movie of each individual life**

PoseNoGap contributors

Swiss Institute of Bioinformatics

Vital-IT
High Performance Computing Center

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

heig-vd
Haute Ecole d'Ingénierie et de Gestion
du Canton de Vaud

Dmitry Kuznetsov
Nicolas Guex
Christian Iseli
Thierry Schuepbach
Heinz Stockinger
Ivan Topolsky
Daniel Zerzion
Ioannis Xenarios

Claudio Alberti
Marco Mattavelli

Enrico Petraglio
Yann Thoma