Challenges in the personal genomics era

Workshop on Processing of Genomic Information: From Standards to Deployment

19th July, Torino

leonor.frias@madeofgenes.com

Genomics is involved in

9 OUT OF 10

leading causes of death*

in developed countries

* 1. Heart disease
2. Cancer
3. Respiratory diseases
4. ~~Accidents~~
5. Cerebrovascular diseases
6. Alzheimer´s disease
7. Diabetes
8. Influenza and pneumonia
9. Kidney disease
10. Suicide (depression)

The information coded in our DNA can be used to identify predispositions, help the diagnosis or drive the treatment of the main death causes in developed countries.

Source: US Center for Disease Control and Prevention (https://www.cdc.gov/)

MADEOFGENES

Disease prevention

20% of Type 2 **Diabetes** can be prevented

93% of **Cystic Fibrosis** cases can be prevented
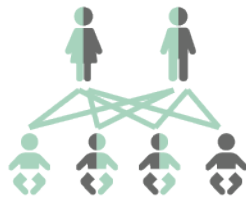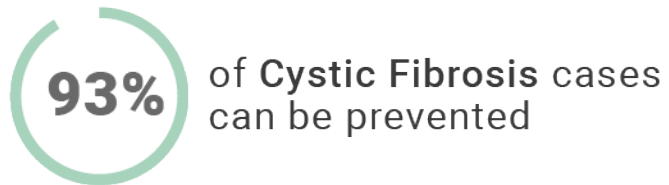
Almost all **rare inherited diseases** can be prevented

Precision medicine

70% reduction of emergencies related to **drug dosage** in patients of cardiovascular accidents

42% of **cancer patients** can benefit for targeted treatment

MADEOFGENES

What happens in a world where everyone gets their DNA sequenced?

Bioinformatics - Forecast

Bioinformatics Market by Sector, Product & Application - Global
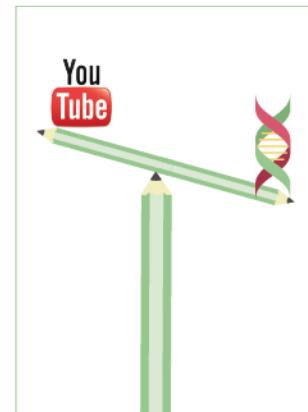Forecast to 2021; Markets & Markets

2016

2021

$6.2B   21% $16.2B

CAGR

"DNA mapping at birth is the norm to avoid disease risk" is identified by Thompson Reuters as one of the 10 trends that will change the world in 2025.

(The world in 2025. Thompson Reuters)

In addition the medical challenge, we are facing a data management problem: if the current growth rates are sustained, in 2025 the world's aggregated -omics data will be larger than the data generated by YouTube.

(Stephens et al. 2015. Plos Biol.)

Is this new paradigm of healthcare **self-sustainable**?

Who **owns** the data and who has **access** to it?

How to interpret **huge amounts of data**?

How to engage the citizenship in **knowledge generation**?

MADEOFGENES

MADEOFGENES

Transactional genomics framework

Made of Genes® provides a unique **framework** for **obtaining**, **storing** and **analyzing**, **-omics data**, either for **healthcare**, clinical **research** or **biomedical** analysis.

MADEOFGENES

Features

DNA Sequencing
whole genome / exome

Patient-centric encryption
patent: PCT/EP2016/061896

Smart informed consent
electronic signature + access control
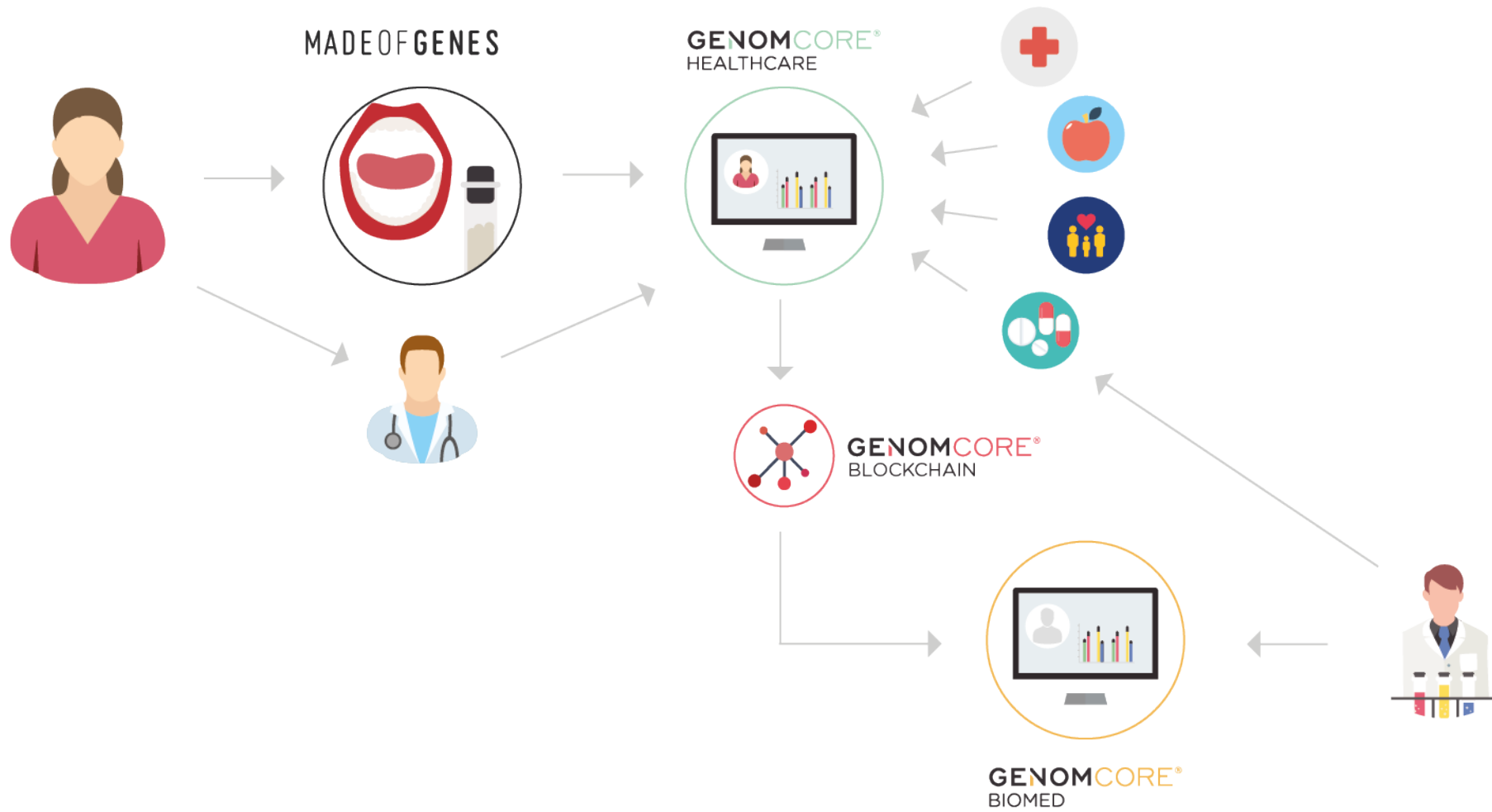

Analysis markeplace
provided by 3rd parties

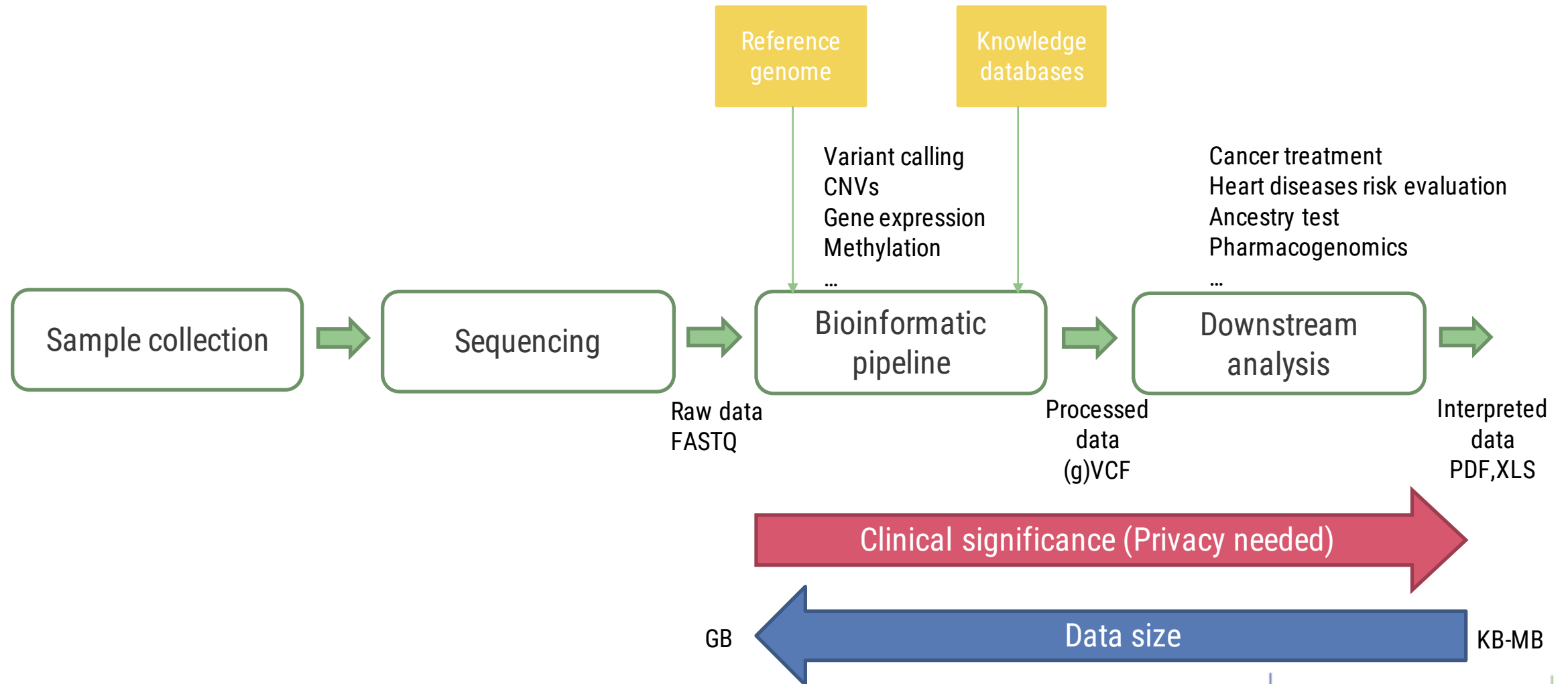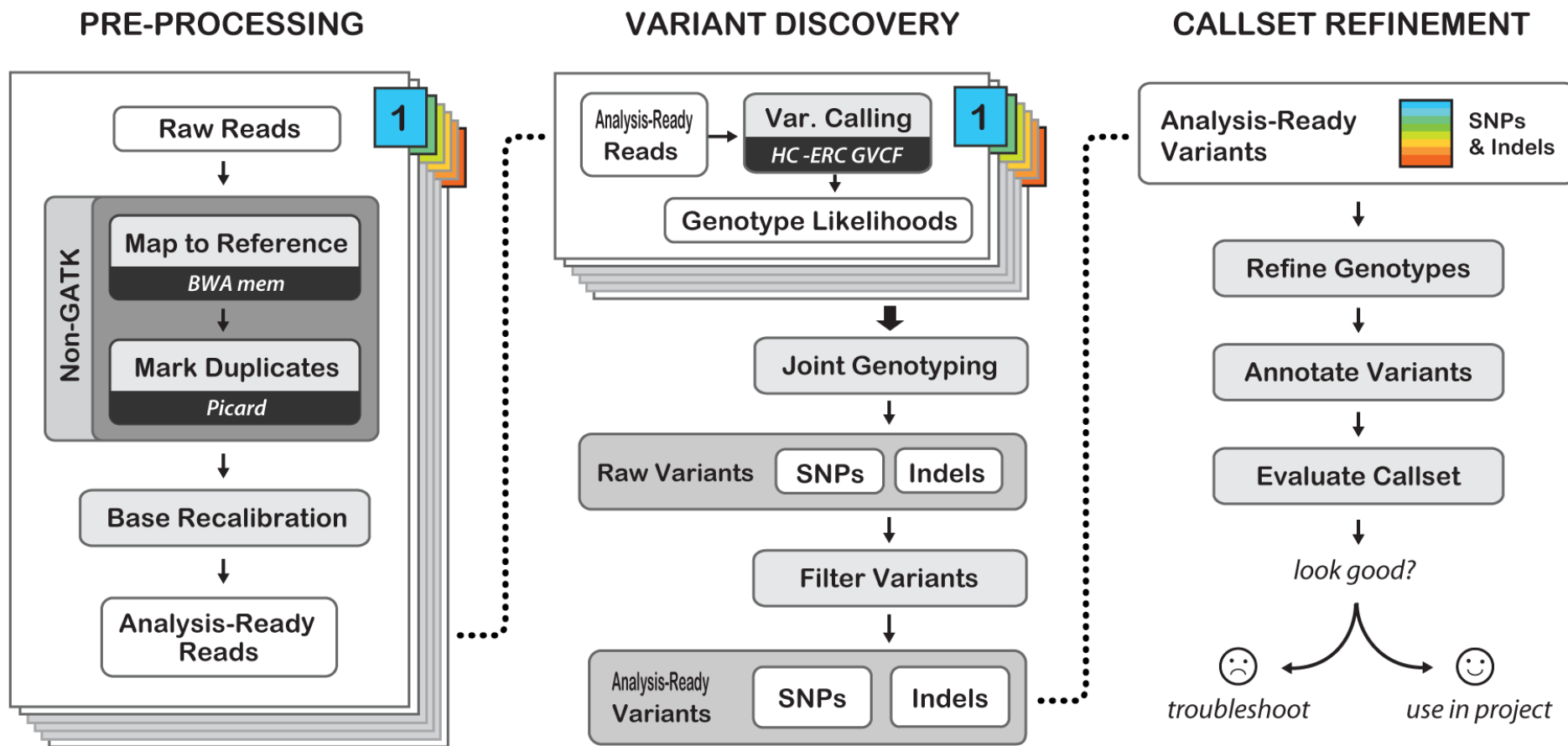**Compliant IT architecture**
cloud / on premises / hybrid

Biomedical Blockchain
crowd-research for empowered patients

Model overview

Characterization of the genome of an individual

PRE-PROCESSING

Raw Reads — 1

Non-GATK

Map to Reference — *BWA mem*

Mark Duplicates — *Picard*

Base Recalibration

Analysis-Ready Reads

VARIANT DISCOVERY

Analysis-Ready Reads → Var. Calling — *HC -ERC GVCF* — 1

Genotype Likelihoods

Joint Genotyping

Raw Variants | SNPs | Indels

Filter Variants

Analysis-Ready Variants | SNPs | Indels

CALLSET REFINEMENT

Analysis-Ready Variants — SNPs & Indels

Refine Genotypes

Annotate Variants

Evaluate Callset

look good?

troubleshoot   use in project

Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

Long-term storage of *raw data* per user (and per sequencing experiment)

. Efficiency / Compression (ratio rather than speed)

. Redundancy / Backups

. Privacy / Encryption

Efficient (real-time) position-based queries to *processed data* per user

Different roles have different access permissions to the data and may have a time constraint (views)

Flexible metadata description and relationship with no genomic data

Roles and uses of genomic data in the Made of Genes case

Role	Raw data		Processed data		Interpreted data	
	Actions	Affectation	Actions	Affectation	Actions	Affectation
Owner	Read, Share, Purpose-share	Part, All	Read, Share, Purpose-share	Part, All	Read	All
Analyst			Read*	Part, All	Generates*	All
HealthCare Profesional					Read	All
Worker	Read*	All	Generates*	All		
Researcher	Read*,**	Part, All	Read*,**, Generates*,**	Part, All	Generates*, **	All

(*) When granted permission by the owner for an specific purpose (and time) and specific experiment

(**) If shared by the owner anonymously or traceably for a given experiment

1. An individual shares the raw/processed data of specific genomic regions and experiment with an analyst, so that he/she performs a bioinformatics/downstream genetic analysis.

2. An analyst accesses the raw/processed data of specific genomic regions from a given individual and experiment

3. An analyst shares with a healthcare professional the outcome of a genetic analysis

4. A healthcare professional accesses the genetic analysis of an individual.

5. An individual shares his raw/processed data with a researcher for a given project. He/she wants to be recontacted back if necessary

6. A researcher asks an individual for more experiments to be used in a research study

Problems with *De facto* standard genomic formats: fastq, sam, bam, bed, vcf,...

Computationally inefficient:

. designed as an ASCII format -> redundant information, bulky

Error-prone metadata:

. Either inexistent as in FASTQ, even GATK Best Practices recommend uBAM instead

. Not straightforward to extract even with "standard" samtools/bamtools: lot of parsing

Privacy / access permissions not considered

...BUT all the tools are designed on the top of them

Computationally efficient:

- Compact, takes advantage of genomic features

- Implements compression

Different types of access: streaming, per position

- Ideally, block-based integrated encryption and compression

Integration with bioinformatic software: critical

- Efficient converters and "tools"

- Parsing modules for main pieces of software

Further wishes for a (standard) format

Consider different compression strategies:

- Long-term storage

- Temporal storage (pipeline processing)

Flexible and easy-to access metadata

- Possibility to add and access user-defined fields

Definition and transparent access control based on roles

- Ideally also being able to define regions inside a file

Pure genomic storage and processing: we use *de facto* standard formats

A)   Long-term storage of the raw data as chromosome chunked BAMs

→ would benefit most from a high quality standard, critical the interoperation with current tools
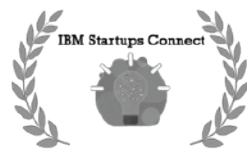
Rest of requirements: usage of external tools

. Data encryption with per owner key using AES256

. Development of a flexible framework to describe any kind of omics data (*Genomcore Biomed*)

. Informed consents and blockchain to control and track access permissions to the genomic data

MADEOFGENES

All genomics. All yours.

Winners Wayra
Global Call 2015

Winners IBM
StartupsConnect 2015

Winners MIT
Innovators under 35

Dubai Future Accelerators
Class of 2017