

Il progetto GDI, un'infrastruttura per lo sviluppo della medicina di precisione

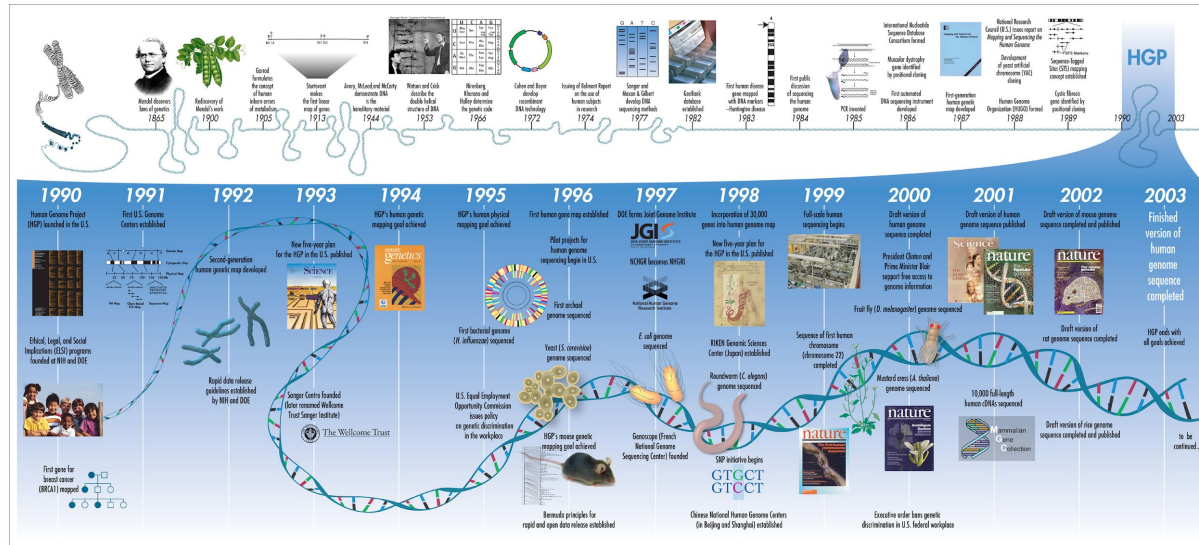
Matteo Chiara

Università degli Studi di Milano

Il progetto genoma umano:



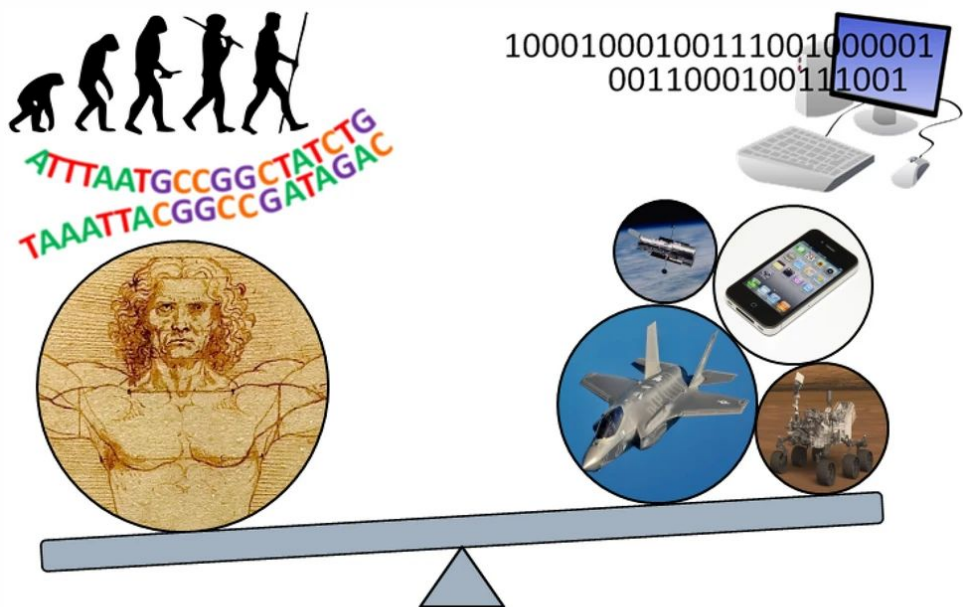
- La prima sequenza completa* di un genoma umano è stata ottenuta nel 2003 (HGP)
- HGP: durata 13 anni. Budget 2.75 miliardi \$\$. Department of Energy and the National Institute of Health.



- Obiettivo: identificare tutti i geni nel genoma umano, trovare le cause delle malattie e relative cure

<http://labiotech.eu/history-of-biotech-25-years-of-the-human-genome-project/>

Genomica e big data



3.3 Billion Base Pairs

44 Million Lines of Code

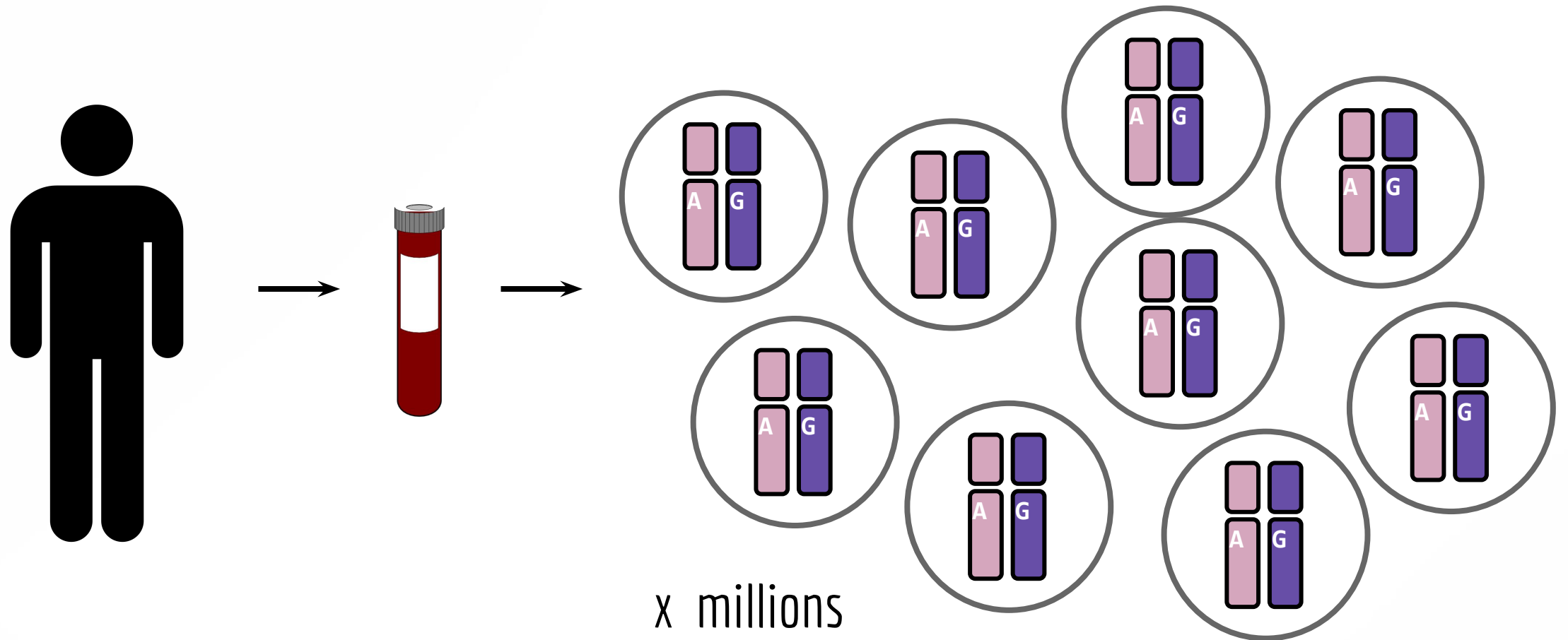
Comparison of complexity between the human genome and several major engineering systems. Image credits: Hubble and Mars Curiosity Rover: NASA/JPL-Caltech; iPhone: Yutaka Tsutano (CC BY 2.0), F-35: MSgt Donald Allen; March of Progress: M. GardeFerdinand (CC BY-SA 3.0).

- La genomica è entrata nell'epoca dei big data!

Determinare la sequenza di un genoma oggi costa poche centinaia di euro

- Big data?
 - Un umano = 10^{12} cellule
 - Una cellula > 6 Gb di DNA
 - Un umano ~ 6×10^{12} Gb di dati!

Oggi possiamo trovare tutte le differenze in un genoma di un individuo, sequenziando il genoma di migliaia di cellule



Perchè si studia il genoma?

Whole genome sequencing could save NHS millions of pounds, study suggests

Genomics England and NHS England findings highlight benefits of using WGS to help detect rare diseases



“Conoscendo la sequenza del genoma umano, possiamo oggi identificare le cause di una malattia livello molecolare”

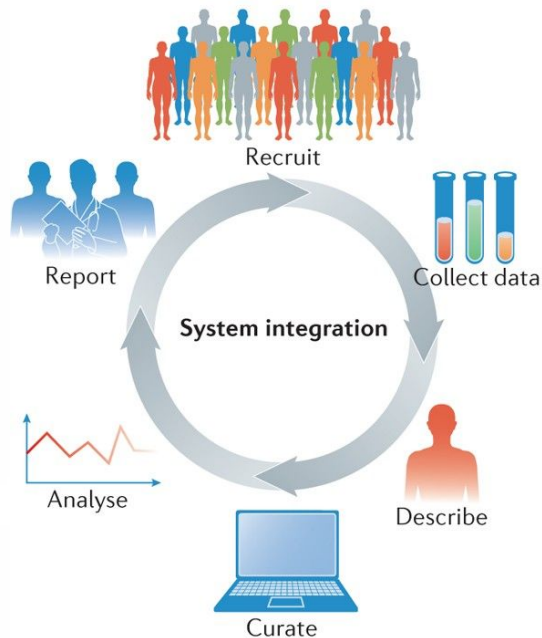
- Siamo composti di miliardi di cellule;
- Le informazioni per “specificare” come una cellula si comporta sono codificate dal genoma
- Difetti nel genoma -> Difetti nelle cellule -> Patologie

Genomica, big data e medicina

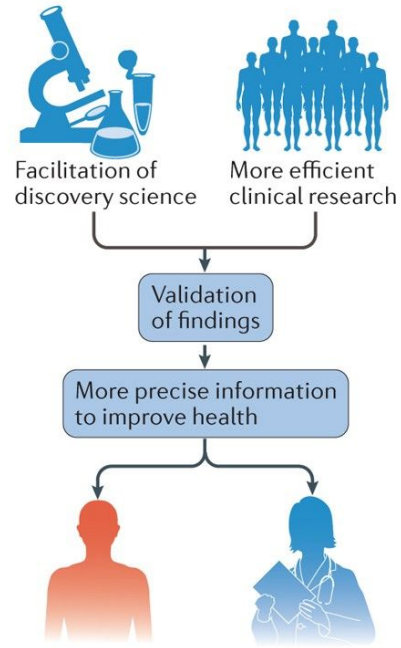


Personalised medicine, also known as precision medicine, tailors treatment and healthcare practices to the individual characteristics of each patient. This approach considers differences in people's genes, environments, and lifestyles to develop more effective and targeted therapies and preventative strategies. It represents a shift away from "one size fits all" treatments towards more individualised care.

a Precision medicine system



b Precision medicine goals



Nature Reviews | Cardiology

- Indirizzare diagnosi e terapia conoscendo il genoma

- Applicazioni:

- Diagnostica: - malattie genetiche rare
- Fattori di rischio - per tratti complessi
- Stratificazione pazienti e definizioni terapie efficaci

Quanto siamo diversi?



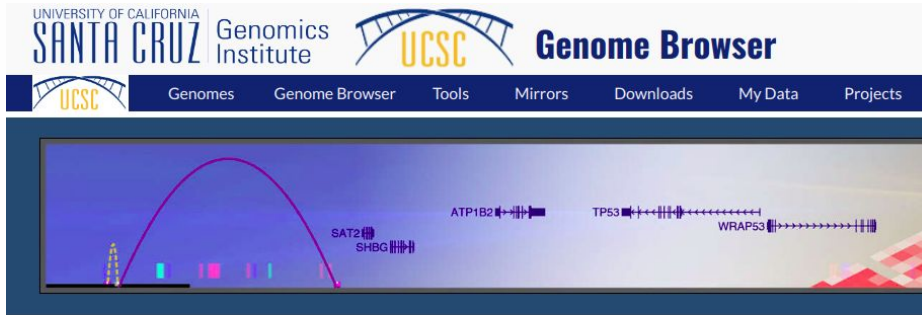
- In media ogni individuo ha 4M di SNVs (differenze nel genoma)
- Tra queste, un numero molto piccolo causa/è associato malattie
- L'obiettivo della "genomica"
 - è capire come e perchè

-99.8% identical DNA









99% identical DNA



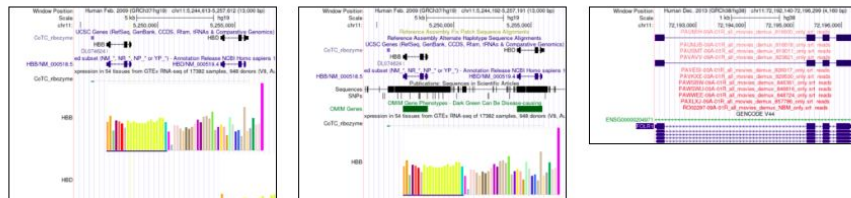
Come si studiano i genomi?



Tools

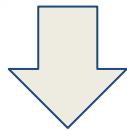
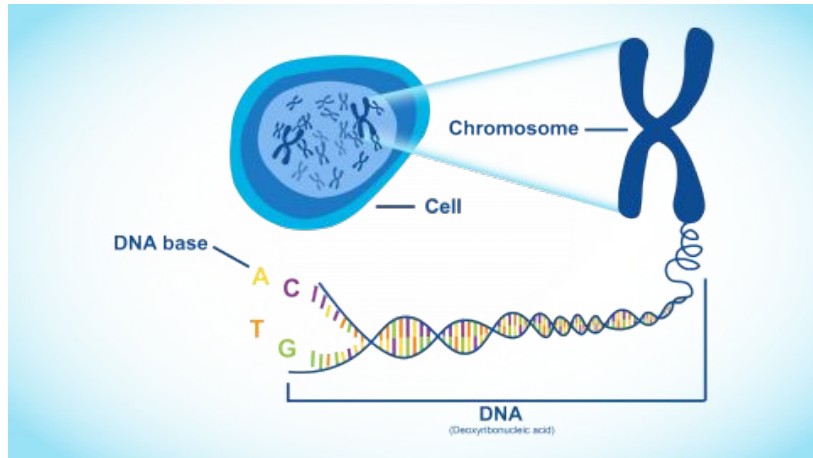
-  **Genome Browser** - Interactively visualize genomic data
-  **BLAT** - Rapidly align sequences to the genome
-  **In-Silico PCR** - Rapidly align PCR primer pairs to the genome
-  **Table Browser** - Download and filter data from the Genome Browser
-  **LiftOver** - Convert genome coordinates between assemblies
-  **REST API** - Returns data requested in JSON format
-  **Variant Annotation Integrator** - Annotate genomic variants
-  **More tools...**

Sharing data



- La biologia è una scienza empirica, si basa su osservazione e analisi
- La genomica, -la scienza che studia il genoma- allo stesso modo, osserva e confronta le sequenze del genoma
- **CONCLUSIONE:** per capire come funziona il genoma, dobbiamo sequenziare tanti genomi!
 - DIFFERENZA → FUNZIONE?

Cosa serve per studiare i genomi?

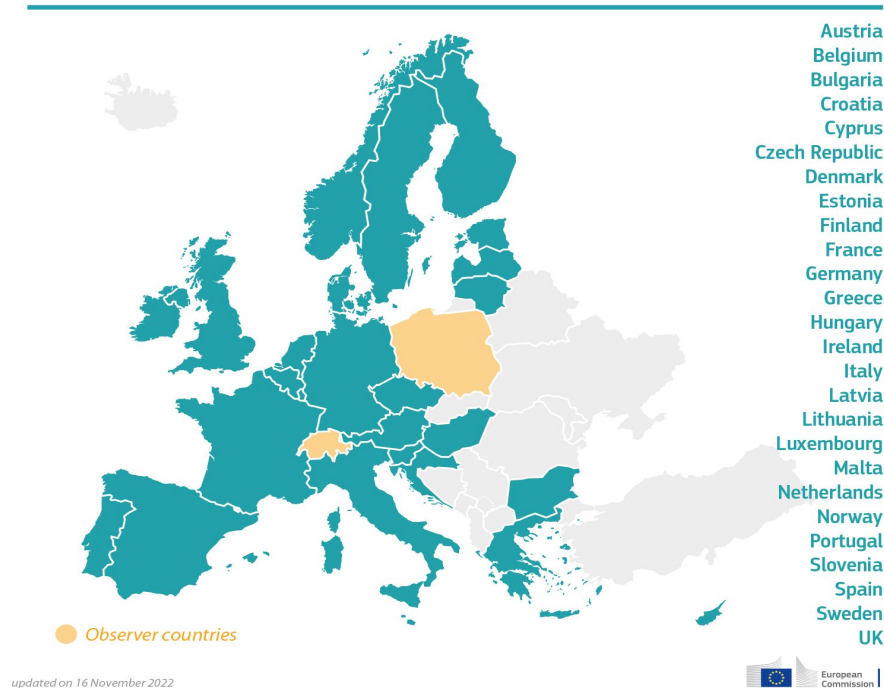


- Conoscere la sequenza. Per farlo ~150 Gb di dati
- Confrontare la sequenza con altri genomi. Per farlo risorse HPC
- Confrontare la sequenza con altri genomi. Per farlo devo accedere ad altri genomi



Cosa serve per la medicina di precisione?

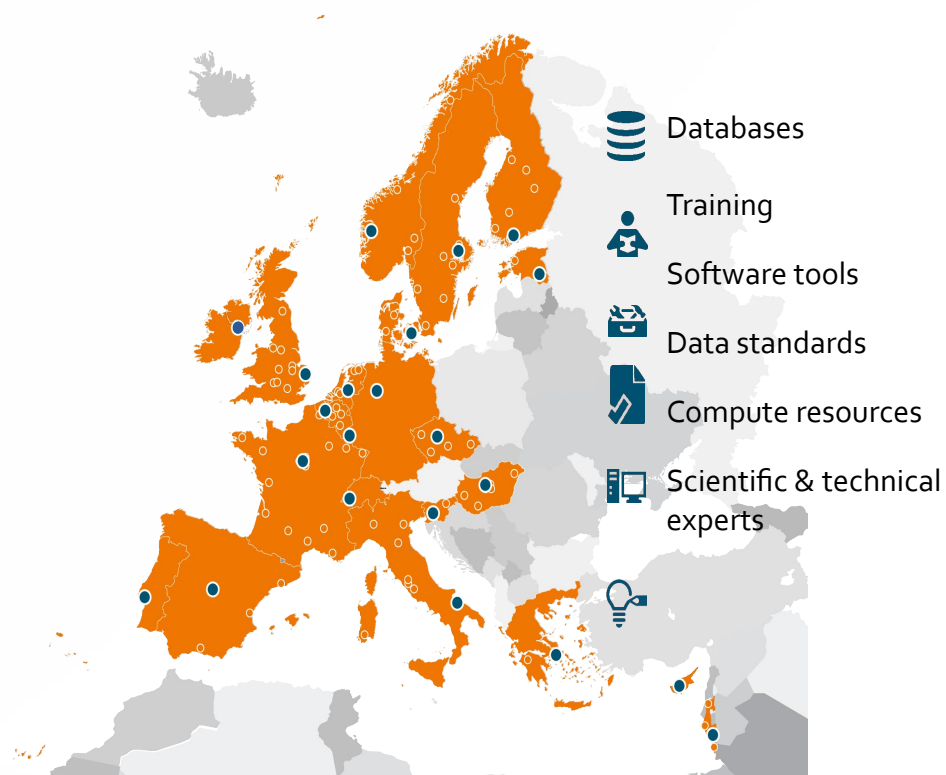
Countries that have signed
the 1+MG Declaration since 2018



1+MG facilitates countries to realise a practice of personalised medicine and health, based upon a shared 'trust framework' and the infrastructure to safely access and integrate high quality genomic data and other health data across borders

- Sequenze genomiche;
- (meta)Dati clinici da associare ai genomi;
- Modi efficienti ed efficaci per analizzare condividere i suddetti dati
- Iniziative organiche nazionali e sovra-nazionali, per non disperdere gli sforzi

Cosa serve per la medicina di precisione?



- Infrastrutture dedicate
- Interoperabili a livello sovra-nazionale
- Il progetto GDI - coordinato dal ELIXIR: l'infrastruttura europea per le scienze della vita- ambisce a costruire questo tipo di infrastruttura

L'iniziativa europea "1+Million Genomes"

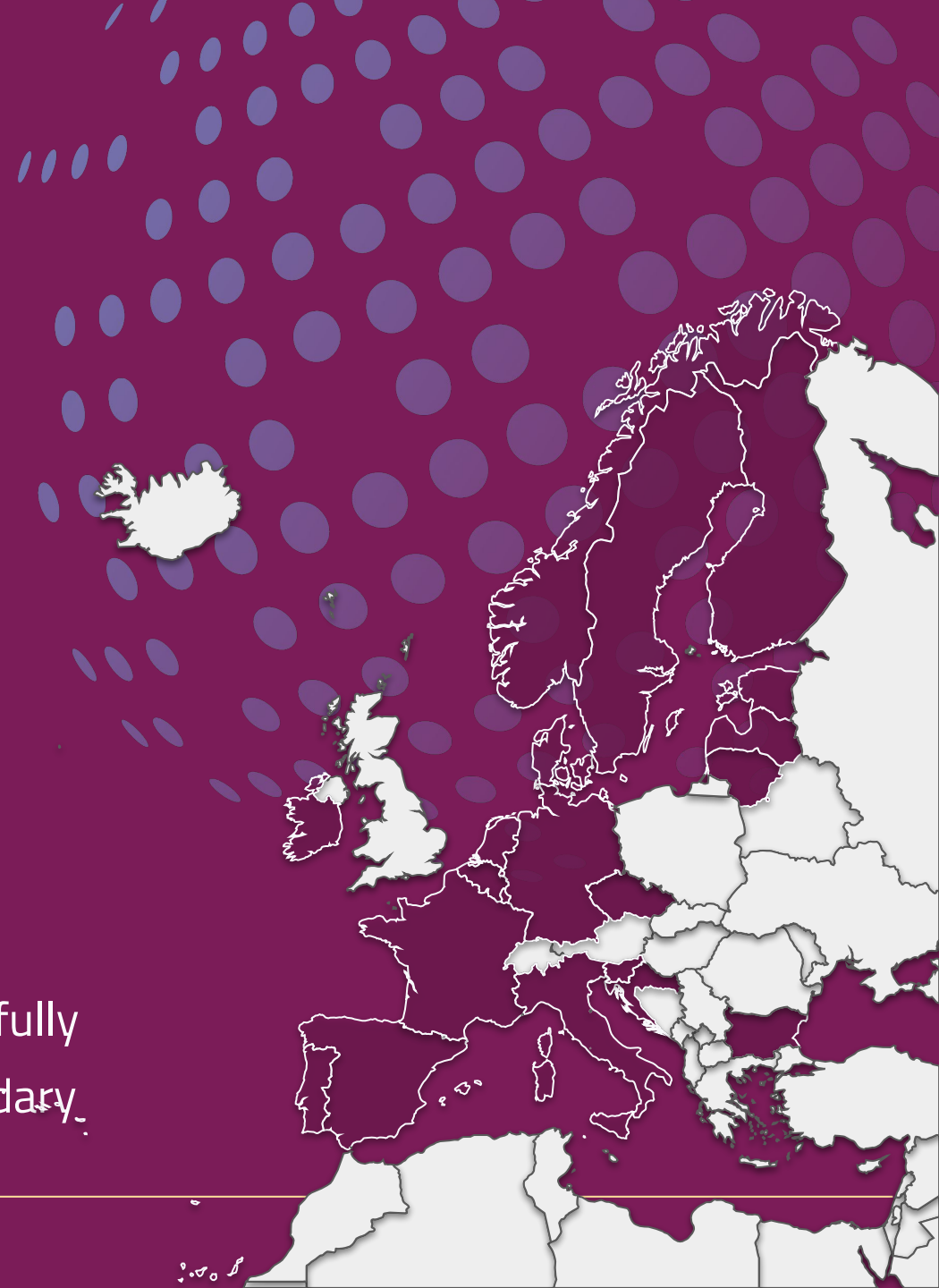
Nel 2018 è stata sottoscritta da 26 paesi europei la dichiarazione 1+M Genomes per sviluppare il contributo dei dati omici nella pratica clinica, sia attraverso il sequenziamento di almeno 1M di genomi di cittadini europei (inclusi 500K di soggetti sani) che con la realizzazione di una infrastruttura condivisa su scala europea per la gestione e condivisione sicura dei dati.





GDI – Genomic Data Infrastructure

- **40 MEUR** over 4 years (2022–2027) to implement the infrastructure for the European 1+ Million Genome Project
50% co-funded by the member states
- 24 European countries with totally 60+ partners
- 502 persons (and increasing)
- Upcoming European Health Data Space (EHDS) which hopefully will provide a more clear regulation on access to and secondary use of human data in healthcare, research and innovation.





Countries' commitment to GDI by 2026

Fully operational and integrated into 1+MG infrastructure:

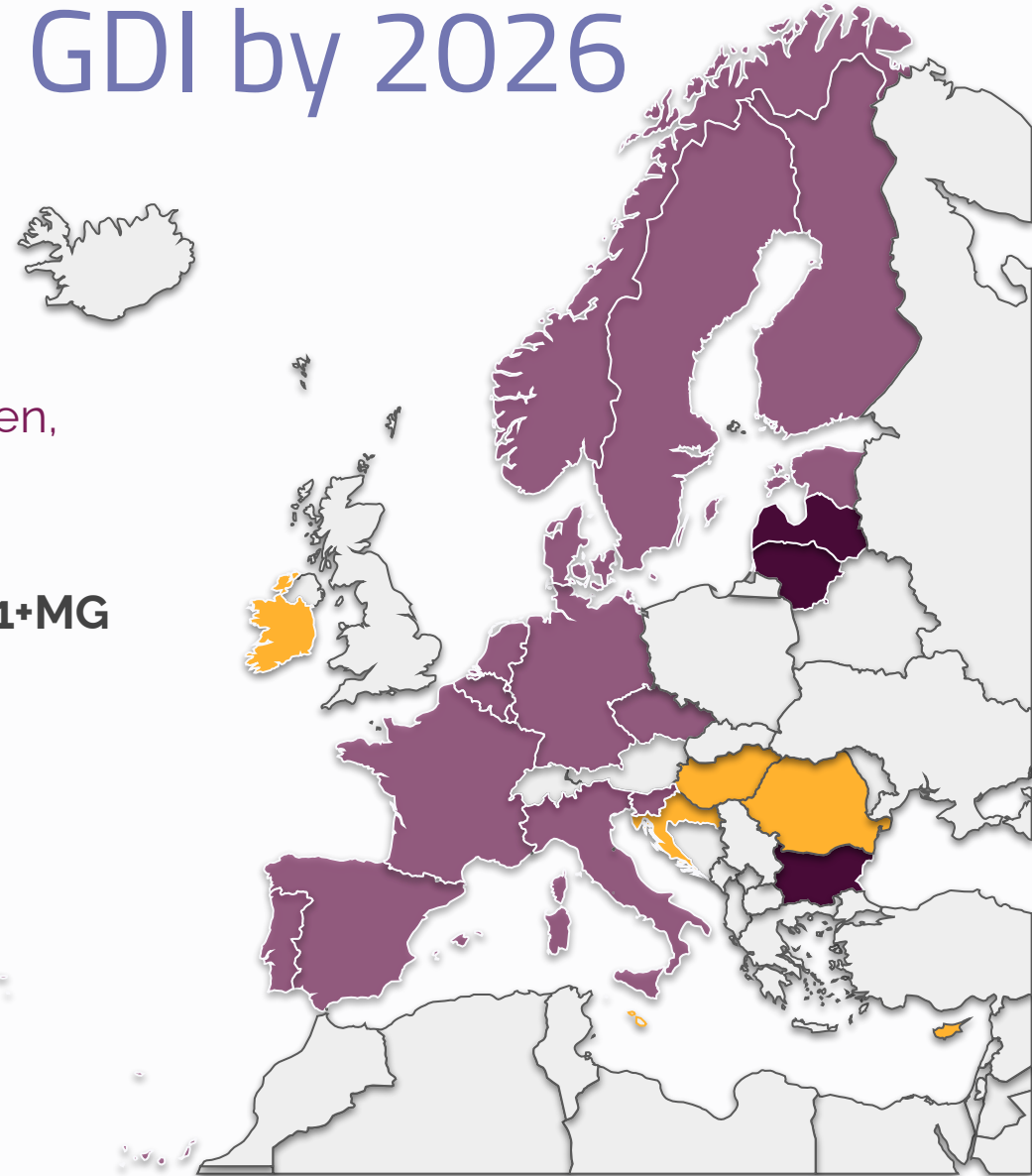
Belgium, Czech Republic, Denmark, Estonia, Finland, France, Germany, Italy, Luxembourg, Portugal, Slovenia, Spain, Sweden, The Netherlands, Norway

Fully operational national node but not yet integrated in the 1+MG infrastructure:

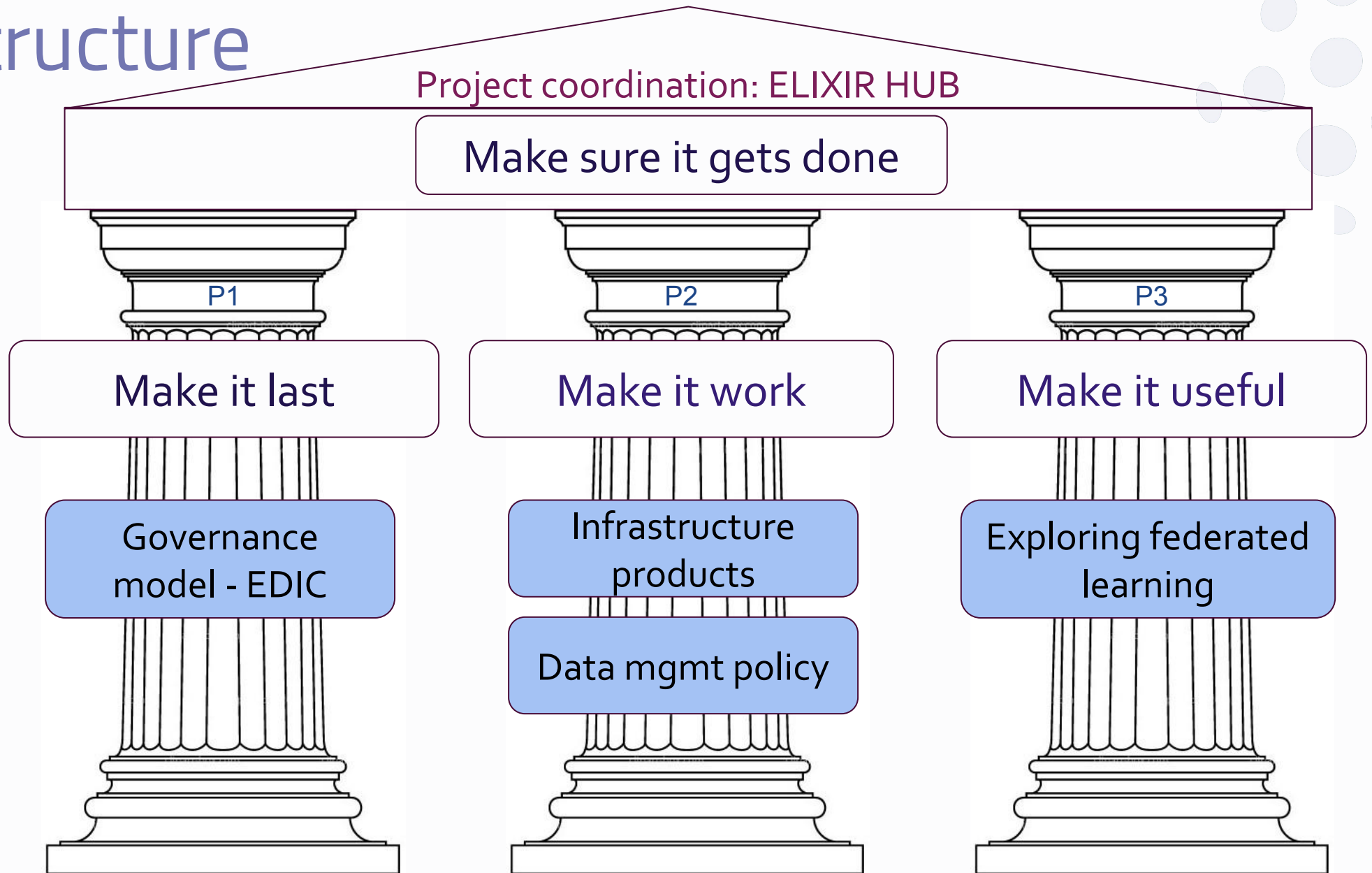
Bulgaria, Latvia, Lithuania

Onboarding:

Croatia, Cyprus, Hungary, Ireland, Malta, Romania



GDI structure





GDI Pillar II



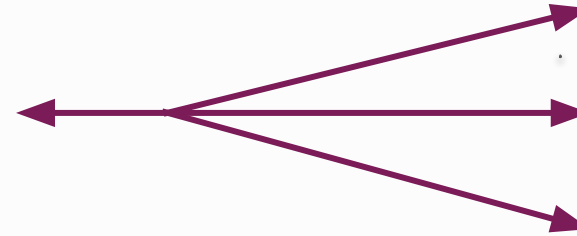
Design & Testing
2020 - 2023



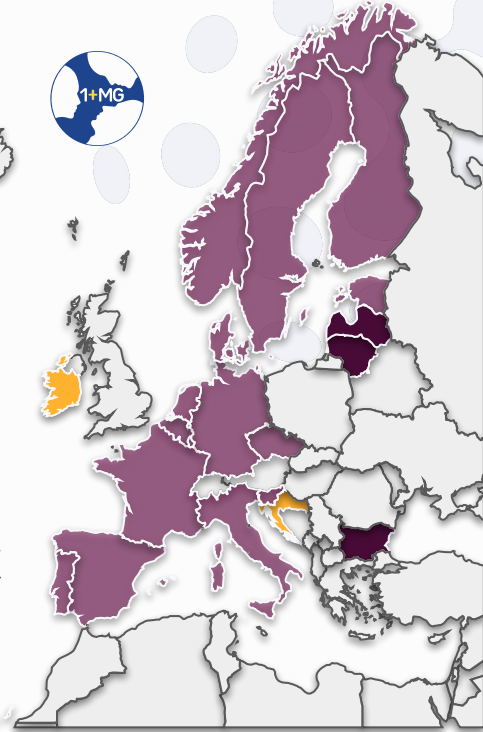
European
Genomic Data
Infrastructure

Scale up & Sustainability

2022 - 2026



Deployment & Feedback



Deployment of 1+MG national nodes

Operational

National nodes interconnected



Deployment

National node fully operational



Onboarding

Design and testing phase

Considerations & Feedback

- National Strategy
- Emerging technologies
- Pillar III Use Cases



GDI: Pillar II

L'obiettivo del GDI è quello di stabilire un'infrastruttura federata, sostenibile e sicura, basata su standard aperti della comunità, per accedere ai dati genomici e ai dati fenotipici e sanitari correlati in tutta Europa



Data discovery



Access management tools



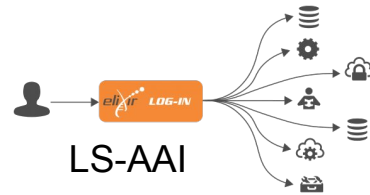
Data processing



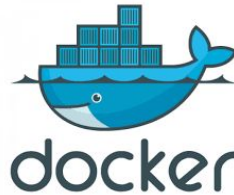
Data reception



Storage and interfaces



rems



Containerized and federated compute



Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.



Federated
European
Genome-phenome
Archive

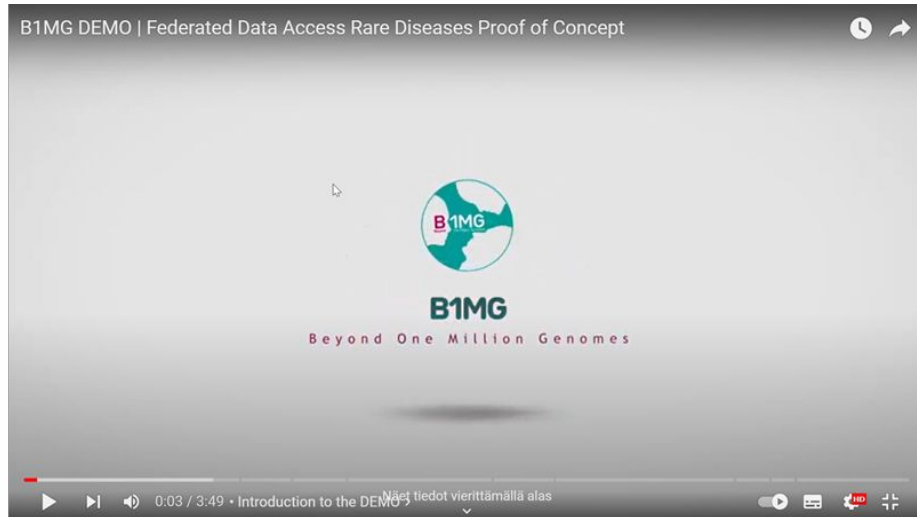
EGA-archive/
crypt4gh



GA4GH cryptographic tools

4 Contributors 37 Used by 1 Discussion 18 Stars 12 Forks

Applicazioni pratiche



<https://www.youtube.com/watch?v=6MtIJA4xXdU>

GDI Italia

Total budget: GDI 40 M€; GDI Italy :1.4 M€ (cofunding 50%)

WP1	Coordination group and communication	
Pillar I		
WP2	Long term sustainability	UCSC, CNR
Pillar II - 1+MG infrastructure deployment		
WP3	Deployment of 1+MG national nodes	UCSC, CNR, HSR, IIT
WP4	European level operations	UCSC, CNR, HSR, IIT
WP5	Technical coordination, training, and outreach	UCSC, CNR, HSR, IIT
WP6	Data management and technical implementation of governance	UCSC, CNR, HSR, IIT
Pillar III - Application and innovation solution		
WP7	GDI use cases	HSR, OPBG
WP8	Transversal and innovative technologies for biomedical data	HSR, OPBG



Genome of Europe



Genome of Europe (GoE) è il più grande progetto genomico finanziato dall'Unione Europea, che ha come obiettivo rendere possibile prevenzione e cure personalizzate per tutti i cittadini europei.

GoE, che svolgerà un ruolo cruciale nelle future scoperte genetiche per la cura della salute e la prevenzione, sosterrà i programmi genomici nazionali e faciliterà l'integrazione della genomica nello Spazio Europeo dei Dati Sanitari (un ecosistema per la salute composto da regole, standard e pratiche comuni, infrastrutture e un quadro di governance per l'accesso ai dati sanitari elettronici e alla loro condivisione/controllo).

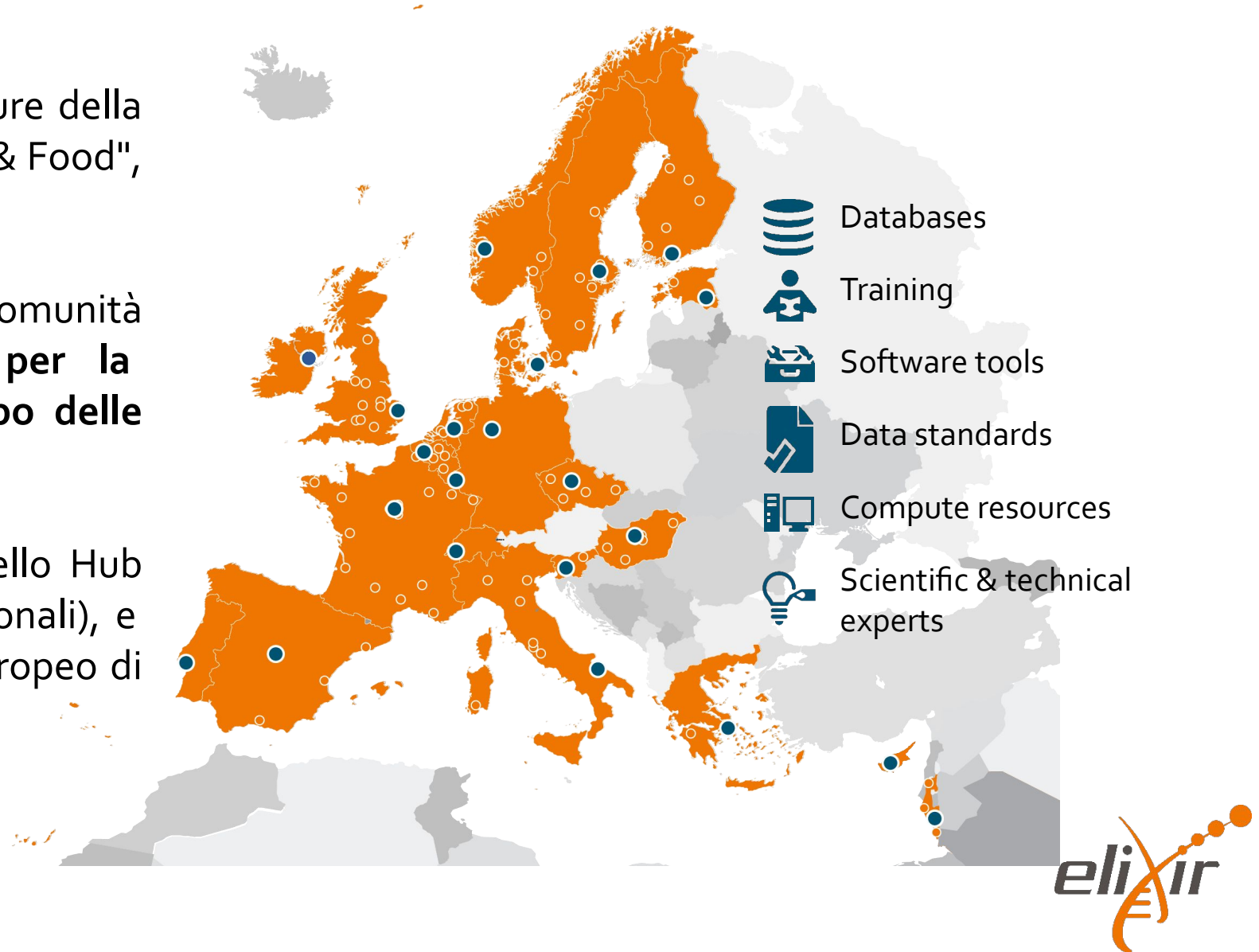
GoE genererà una banca dati, cui parteciperanno 27 Paesi UE, che conterrà inizialmente un minimo di 100.000 genomi interamente sequenziati, rappresentativi dei cittadini europei provenienti dalle principali 40 sottopopolazioni (etnie). In una seconda fase la banca dati crescerà fino a contenere 500.000 genomi. GoE genererà un genoma di riferimento europeo e una serie di genomi "nazionali", su cui basare la ricerca clinica e di laboratorio per nuovi farmaci, nuove terapie mirate, nuove strategie di prevenzione e diagnosi più accurate.

ELIXIR Europe

ELIXIR, è una delle principali infrastrutture della roadmap ESFRI afferent all'area "Health & Food", avviata formalmente nel 2016.

La mission di ELIXIR è offrire alla comunità scientifica **risorse "bioinformatiche" per la ricerca e le sue applicazioni nel campo delle Scienze della Vita e della Salute.**

ELIXIR è organizzata secondo un modello Hub (Cambridge, UK) & Spoke (23 nodi nazionali), e formalmente afferisce al Laboratorio Europeo di Biologia Molecolare (EMBL).





European Genomic Data Infrastructure




ELIXIR-IT is the Italian Node of **ELIXIR**, the European life science infrastructure, supporting basic and translational research by providing bioinformatics services and facilitating access to data in the field of biological sciences.



ELIXIR-IT includes several thematic communities and **six operational Platforms: Compute, Data, Interoperability, Omics, Tools, and Training.**

COMPUTE

The Compute platform offers specialized and customizable bioinformatics analysis services such as workflow managers and advanced tools, partnering with technological partners.

- Collaboration with ReCas-Bari and CINECA for computation and storage services
- In-cloud Laniakea platform for on-demand workflow management systems
- IaaS and PaaS platform at ReCas-Bari to host and deploy bioinformatics tools and services

INTEROPERABILITY

The Interoperability platform focuses on services related to standardization and transparent access to extensive research data, enacting internationally recognized FAIR principles, and promoting life science communities to adopt standardized formats and identifiers.

- Implementation of FAIR principles for data management
- Best practices and guidelines for data interoperability
- Development of protocols and ontologies for data integration

DATA

The Data platform facilitates access to a collection of databases by promoting the principles of open science, implementing the FAIR principles and therefore promoting access to large quantities of data through which companies can develop applications, products, and services for their customers.

- Databases for proteomics, genomics, transcriptomics, molecular interactions and metabolomics curated by experts
- Extensive expertise in managing omic data
- Expertise in data curation and FAIRification

OMICS

The Omics platform provides services for omic data generation, focusing on massive nucleic acids sequencing and high-throughput characterization of metabolomes and proteomes.

- Whole genome, exome and transcriptome sequencing
- Single-cell genome and transcriptome sequencing also at spatial resolution
- Microbiome analysis through metagenomics and DNA metabarcoding sequencing
- Optical mapping of chromosomes
- Epigenetic analysis
- Metabolomics and proteomics analysis

TOOLS

The Tools platform provides software, workflows and libraries for omic and molecular data analysis. It provides reliable, expert-developed, optimized, reproducible bioinformatics software for biological and biomedical data analysis.

- Characterization and integration of omic data
- Automation of omic data analysis workflows
- Standardisation and optimisation of analysis tools

TRAINING

The Training Platform is specialized in delivering tailored, practical training designed to meet the diverse needs of business and industry. By integrating multi-platform expertise, it offers courses delivered by international experts, ensuring a high-quality learning experience.

- Customized Training
- Theoretical and Practical Courses
- Highly Qualified International Instructors

Website

<https://elixir-italy.org>

ELIXIR IT Newsletter
Subscription



1. GDI ambisce a facilitare la condivisione dei dati genomici;
2. (non ne ho parlato) ma gli aspetti ELSI e la tutela del dato rappresentano il punto più critico;
3. La genomica e la medicina di precisione necessitano di HPC, e non necessariamente/non solo per sviluppare nuovi sofisticati modelli di AI;
4. A livello nazionale (ma anche sovranazionale) ELIXIR rappresenta un punto di riferimento per la gestione dei dati e la messa a sistema degli sforzi già prodotti