Genomics Medicine at EMBL-EBI and the Global Alliance for Genomics and Health

Maria J. Martin, PhD (martin@ebi.ac.uk)

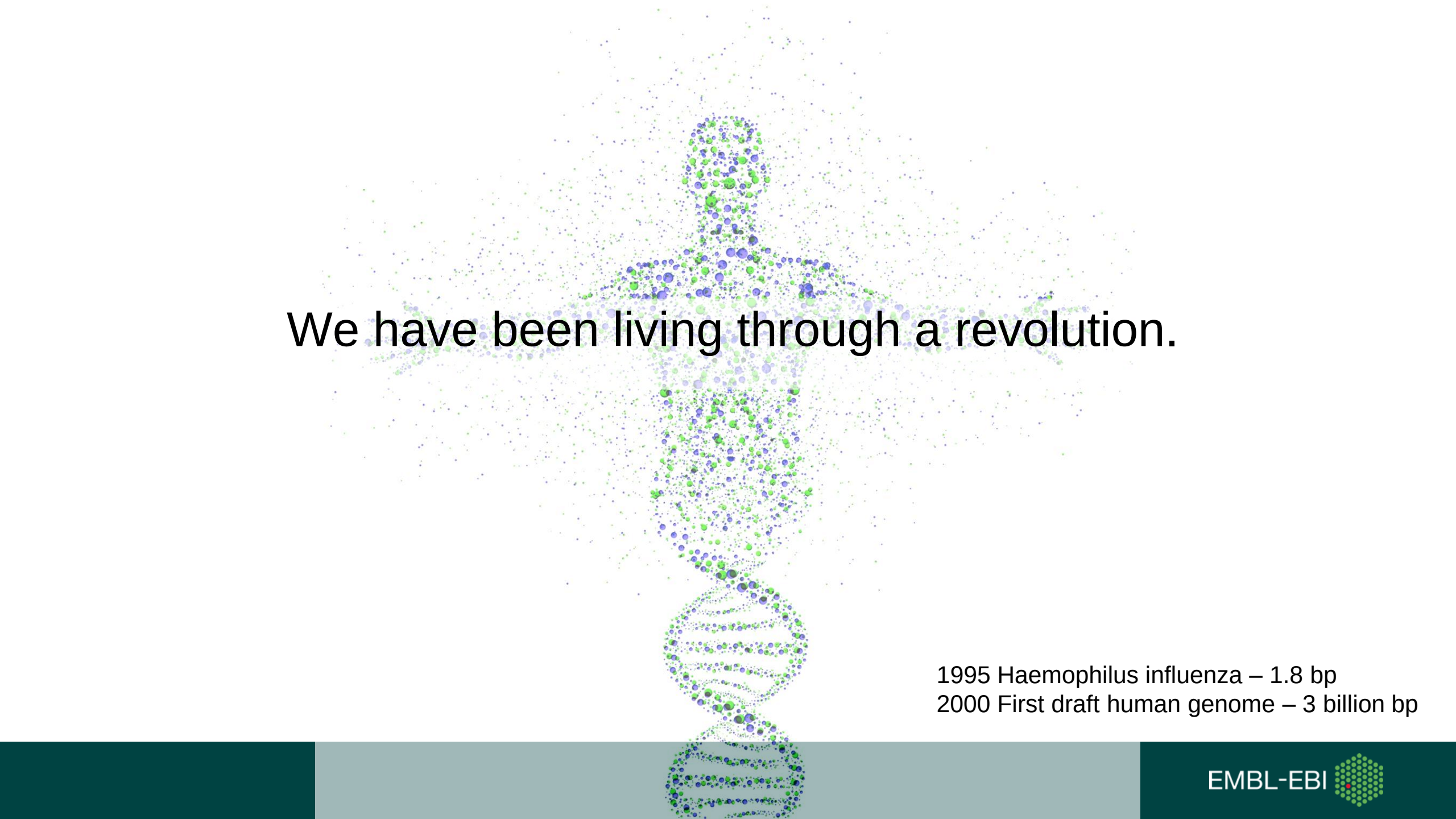
Team Leader

EMBL-European Bioinformatics Institute (EMBL-EBI)

Wellcome Genome Campus

Cambridge, UK

EMBL-EBI



We have been living through a revolution.

1995 Haemophilus influenza – 1.8 bp
2000 First draft human genome – 3 billion bp

Sequencing is now “cheap enough”

The cost of sequencing a genome in 2003

The cost of sequencing a genome in 2018

Between $200-300 /exome, and $800-$1000 for whole genome

$100 Genome within the next 5 years (likely 3 years)

More costs now in consent, DNA sample acquisition (storage and standard analysis low-ish, but not 0!)

Clinical Utility is present: Rare disease

- Rare diseases (i.e. those with a person frequency of 1 in 2,000 or lower (1)) often have a clear genetic component, often of high penetrance with only a few genes involved in each disease.
- Genomics provides a confident diagnosis for many rare diseases, which enables families and healthcare systems to manage the disease appropriately
- More diagnoses at lower costs

Diagnosing rare disease

Sequencing the genomes of children with suspected rare disease and their parents can yield a diagnosis for one in four patients..

20 - 30 %

Clinical Utility is present: Cancer

- When cancer patients have genomic sequence this informs clinical decisions ~10% of the time

- The clinical research community is confident this will increase steadily over the coming years

Genomics meets healthcare

Percentage of whole genomes and exomes that are funded by **healthcare** systems

Areas of clinical uptake: infectious disease, cancer, rare disease, common/chronic

Genomics in healthcare: GA4GH looks to 2022

Ewan Birney, Jessica Vamathevan, Peter Goodhand

doi: https://doi.org/10.1101/203554

The economics of healthcare is different from research

- Healthcare has far larger % GDP than research
- Healthcare often has strong cost control but…
- …when costs are justified and approved for a medical procedure, it is very likely to be applied to all patients where it is useful

Genomics enters healthcare

In **2017** active genomic medicine programmes are already underway in many countries. Finland, the UK, the US, and Australia are a few examples.

1 MILLION PEOPLE

The National Institutes of Health plans to recruit over 1 million people into its All of Us research programme.

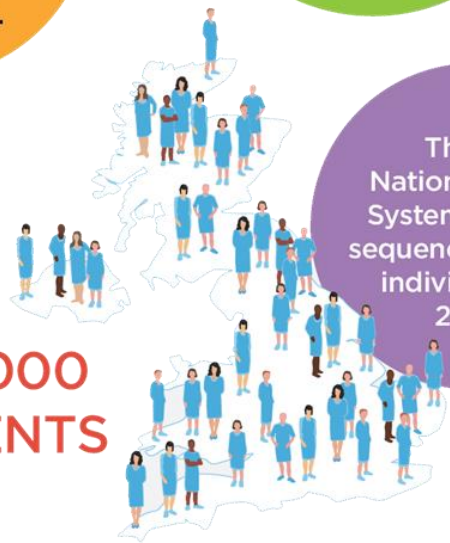
18,000 people by 2021

To improve patient outcomes and support research, the Australian healthcare system is building a Federation of clinical and genomic data.

10% of Finland's population expected to have some genomic data in healthcare by 2020.

10%

100,000 PATIENTS



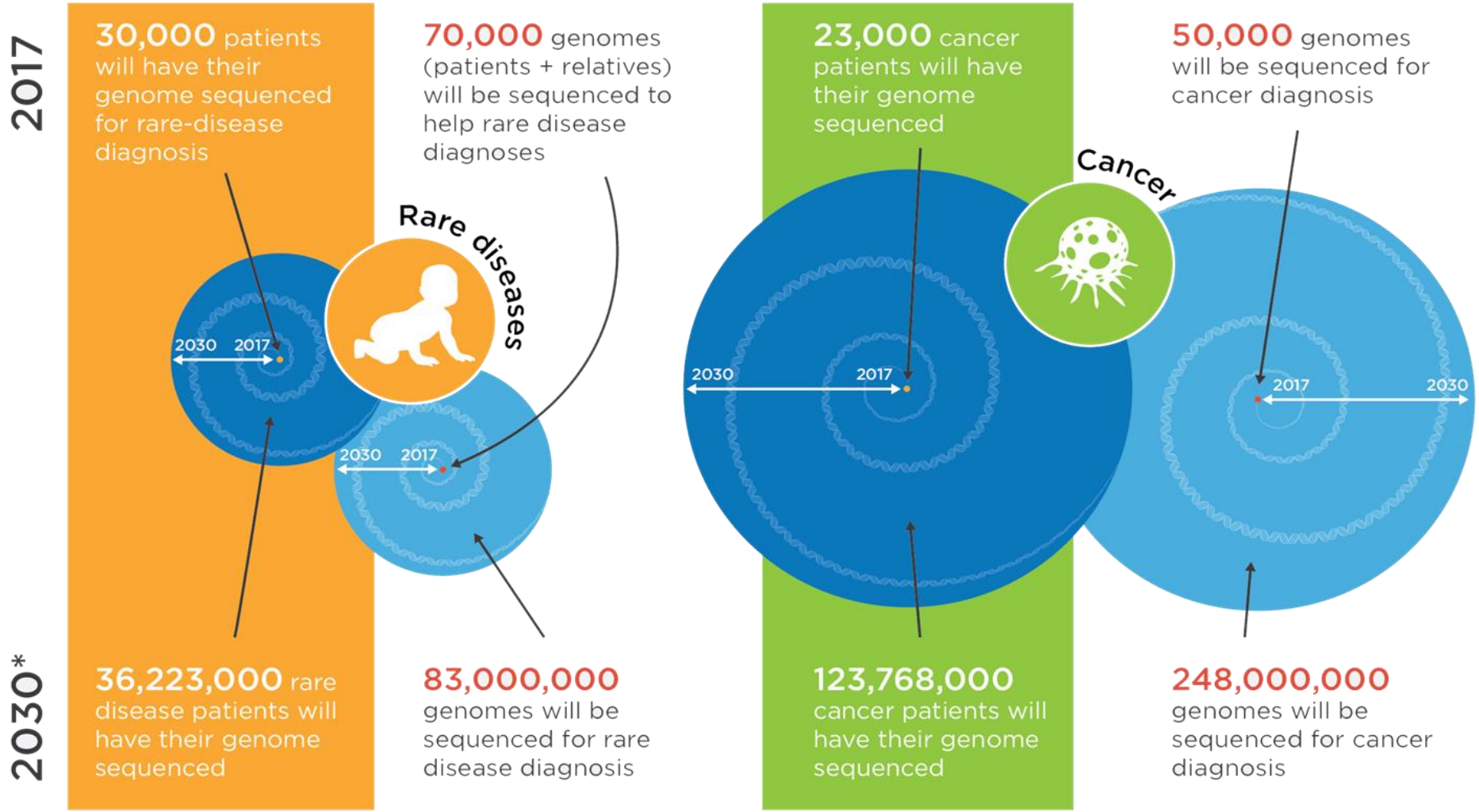
The UK National Health System plans to sequence 100,000 individuals by 2020.

Genomics in healthcare: GA4GH looks to 2022

Ewan Birney, Jessica Vamathevan, Peter Goodhand

doi: https://doi.org/10.1101/203554

National Genomics Initiatives



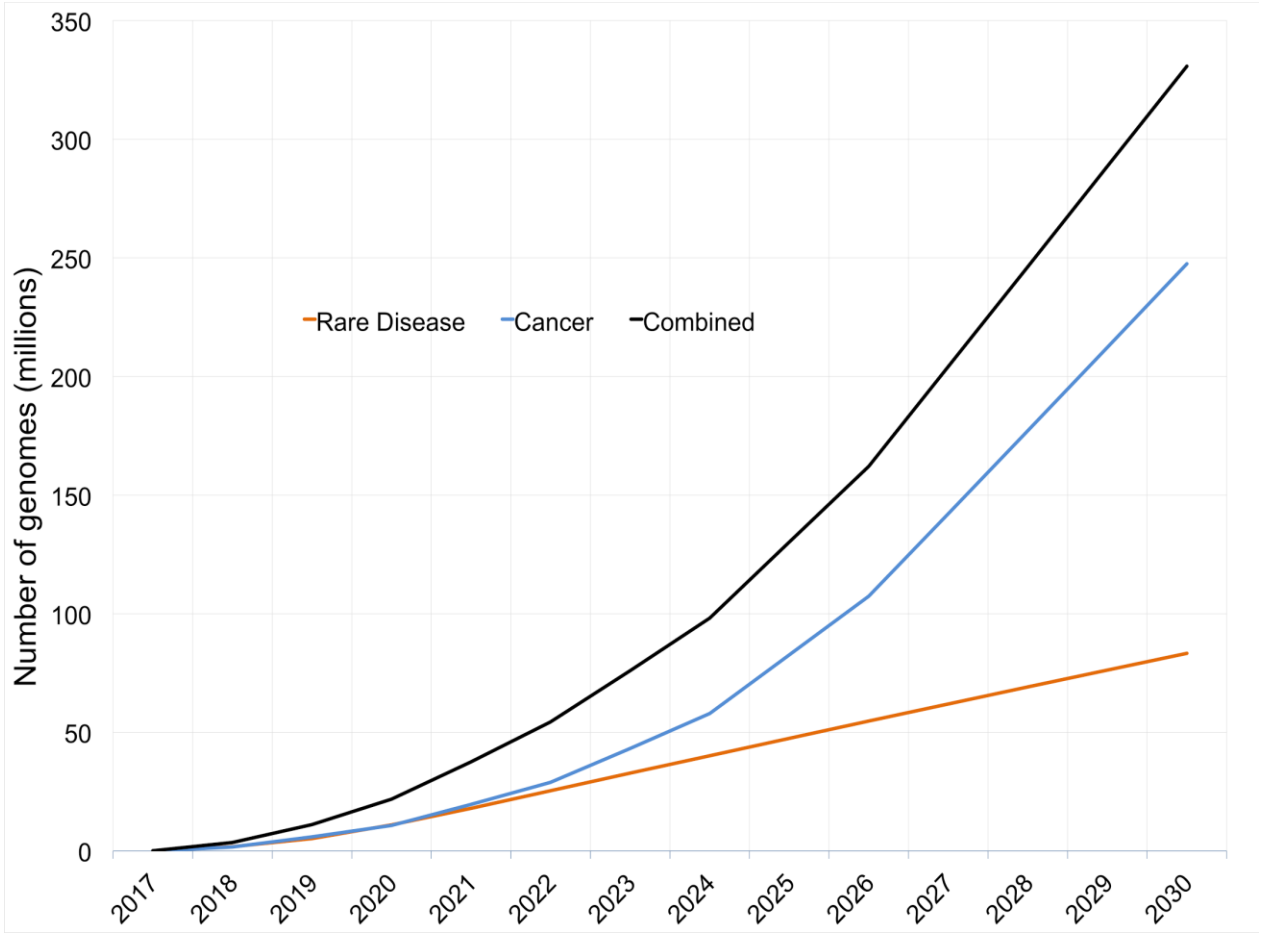
* Projected figures, based on current data and known status of genomics initiatives worldwide.

Genomics in healthcare: GA4GH looks to 2022

Ewan Birney, Jessica Vamathevan, Peter Goodhand

doi: https://doi.org/10.1101/203554

Big numbers!


350
300
250
200
150
100
50
0
Number of genomes (millions)
Rare Disease
Cancer
Combined
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030

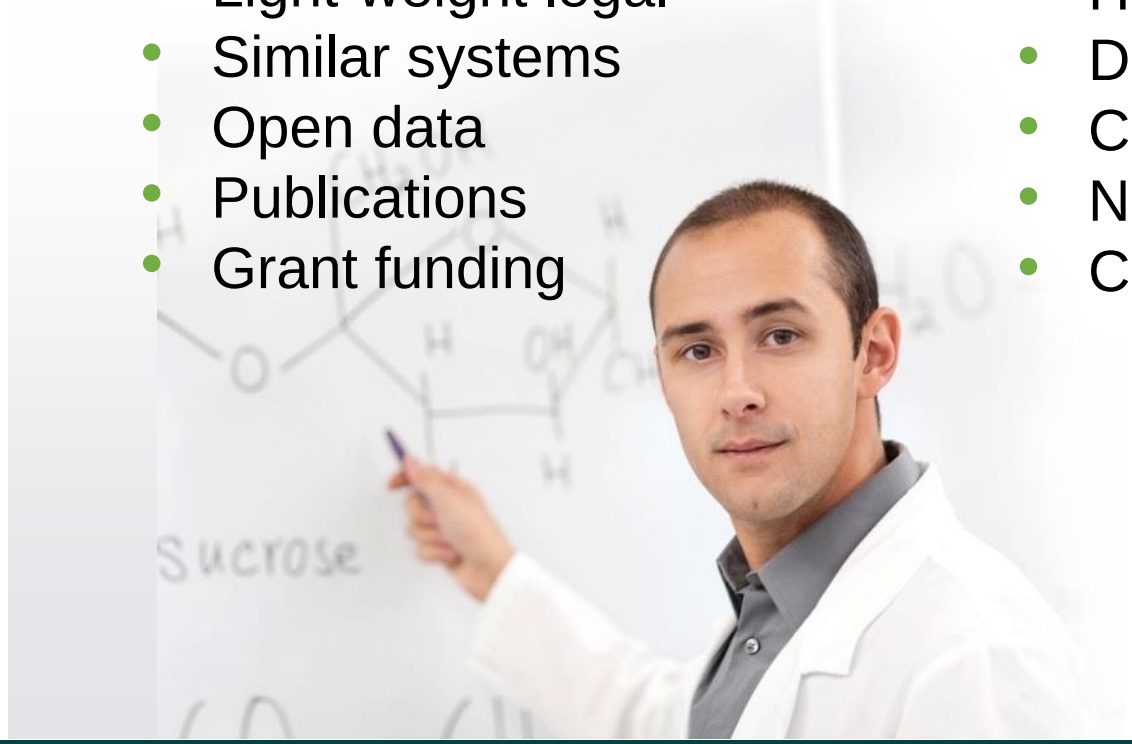
Opportunity

- If we can enable secondary use of clinical genomic data for research we will have a >60 million virtual cohort by 2025
- Data from millions of samples may be needed to show patterns in health/disease
- Humans will be the best studied organisms on the planet due to healthcare

Genomics: from research to healthcare

Research

- English language
- Light-weight legal
- Similar systems
- Open data
- Publications
- Grant funding



Practicing Medicine

- National language
- Heavy legal framework
- Different systems
- Closed data
- Not published
- Contract funding

The challenges are surmountable

- Healthcare not used to this type, amount of data; we must draw on skills, learnings of research
- Clinical data are not interoperable: cultural, regulatory, technical differences between international healthcare systems. Portable analysis routines must be developed.
- As genomic datasets grow from terabyte to petabyte to exabyte scale, the community must re-tool
- Different nations have unique regulatory environments, requiring broad, reciprocal data access methods that are as open as possible while respecting national processes and patient consent.

The challenges are surmountable

Impact of data sharing:

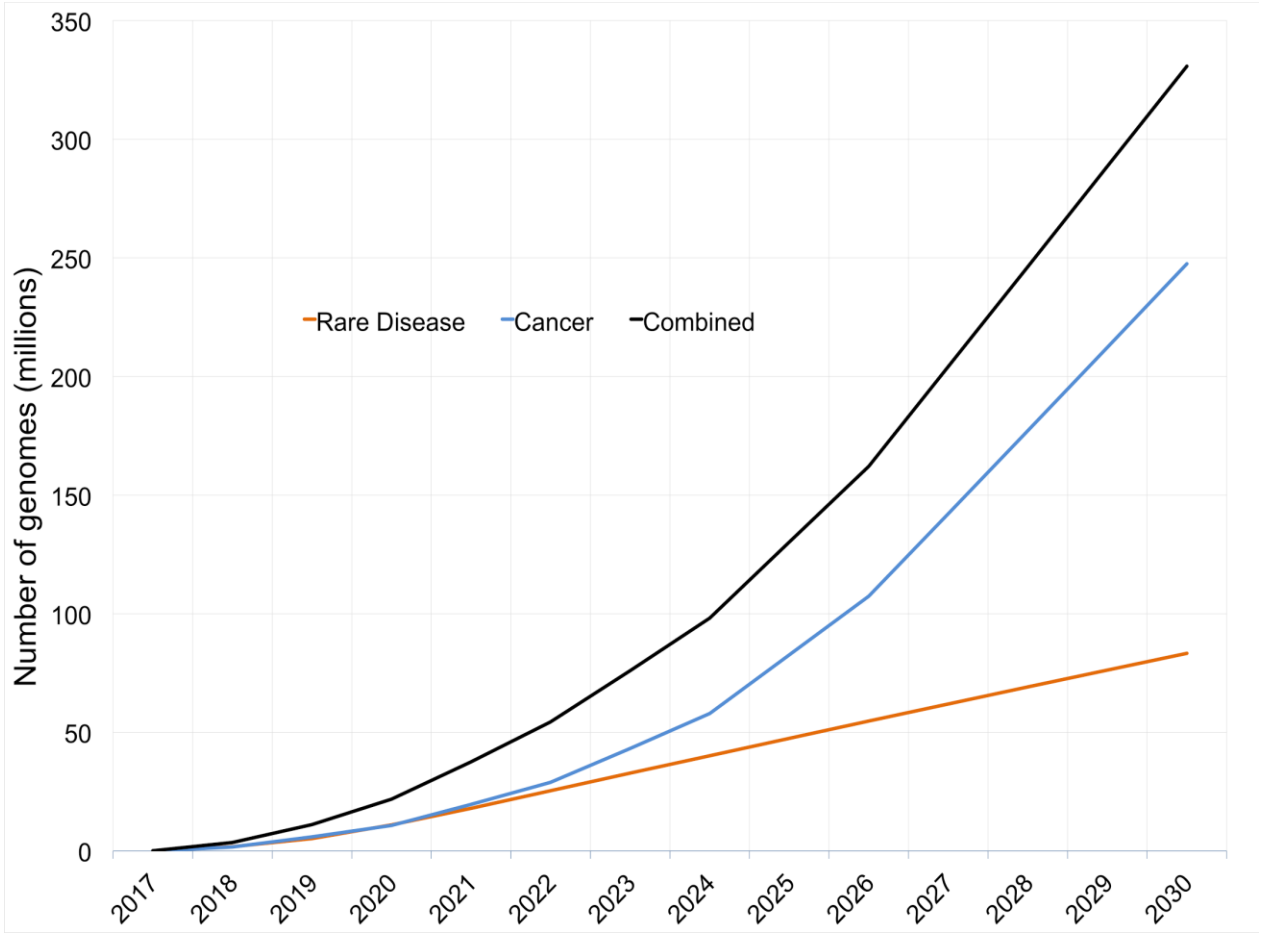
- Increases statistical significance of analyses
- Helps find that other, similar rare disease patient
- Leads to ‘stronger’ variant interpretations backed by group consensus
- More informed clinical decisions; peace of mind

Don’t act: an overwhelming mass of fragmented data, such as electronic medical records in many countries
Collective Action: achieve the interoperability of the www or global telecommunications / smartphones

Responsibility

- Technical knowhow around genomics is in the research community
- Technical knowhow around clinical features and diagnosis is in the clinical community
- We have a joint responsibility to make this work for patients

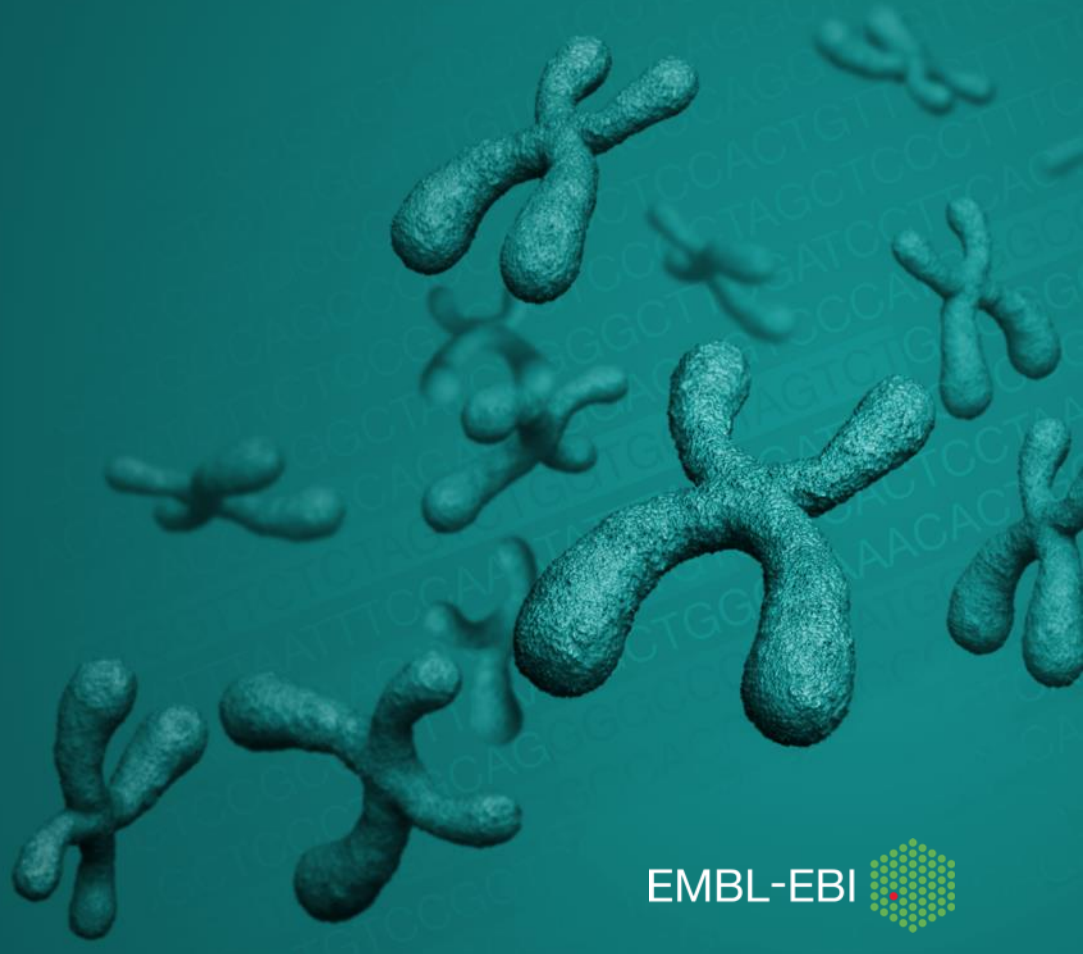
Big numbers!


350
300
250
200
150
100
50
0
Number of genomes (millions)
Rare Disease
Cancer
Combined
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030

The European Bioinformatics Institute

The home for big data in biology

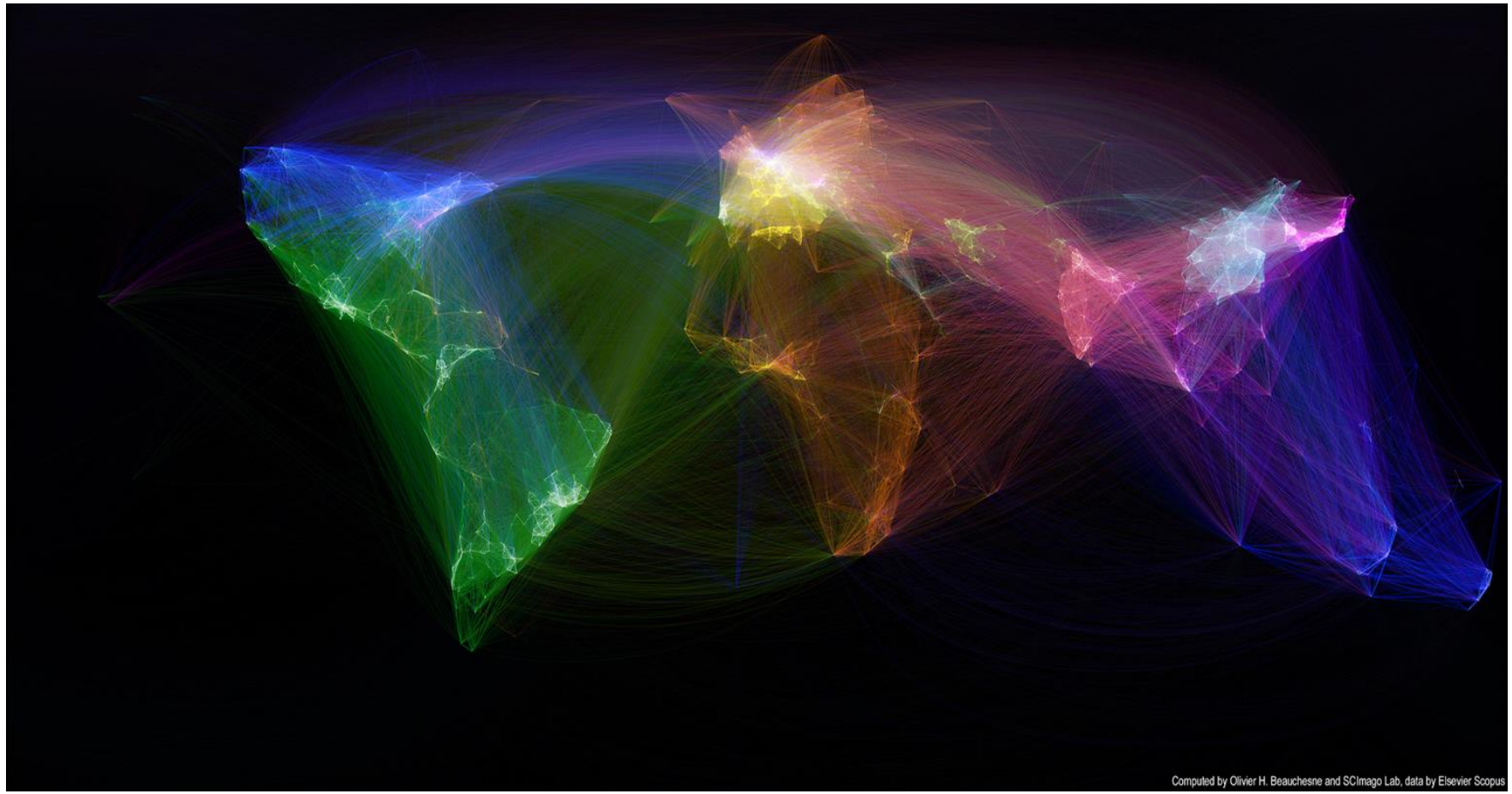
www.ebi.ac.uk



EMBL-EBI

Research / EMBL- EBI ideal world

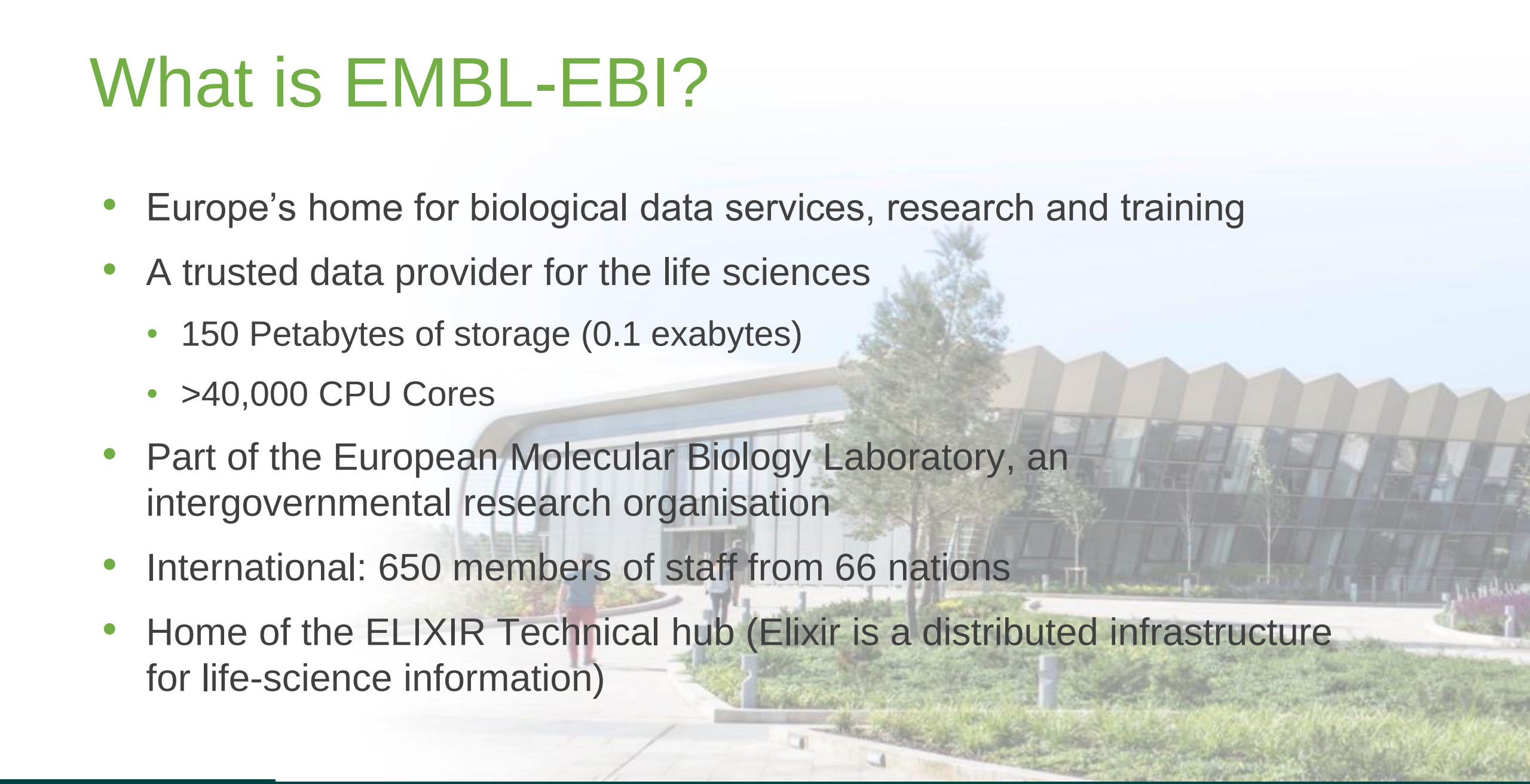
- Open
- collaboration



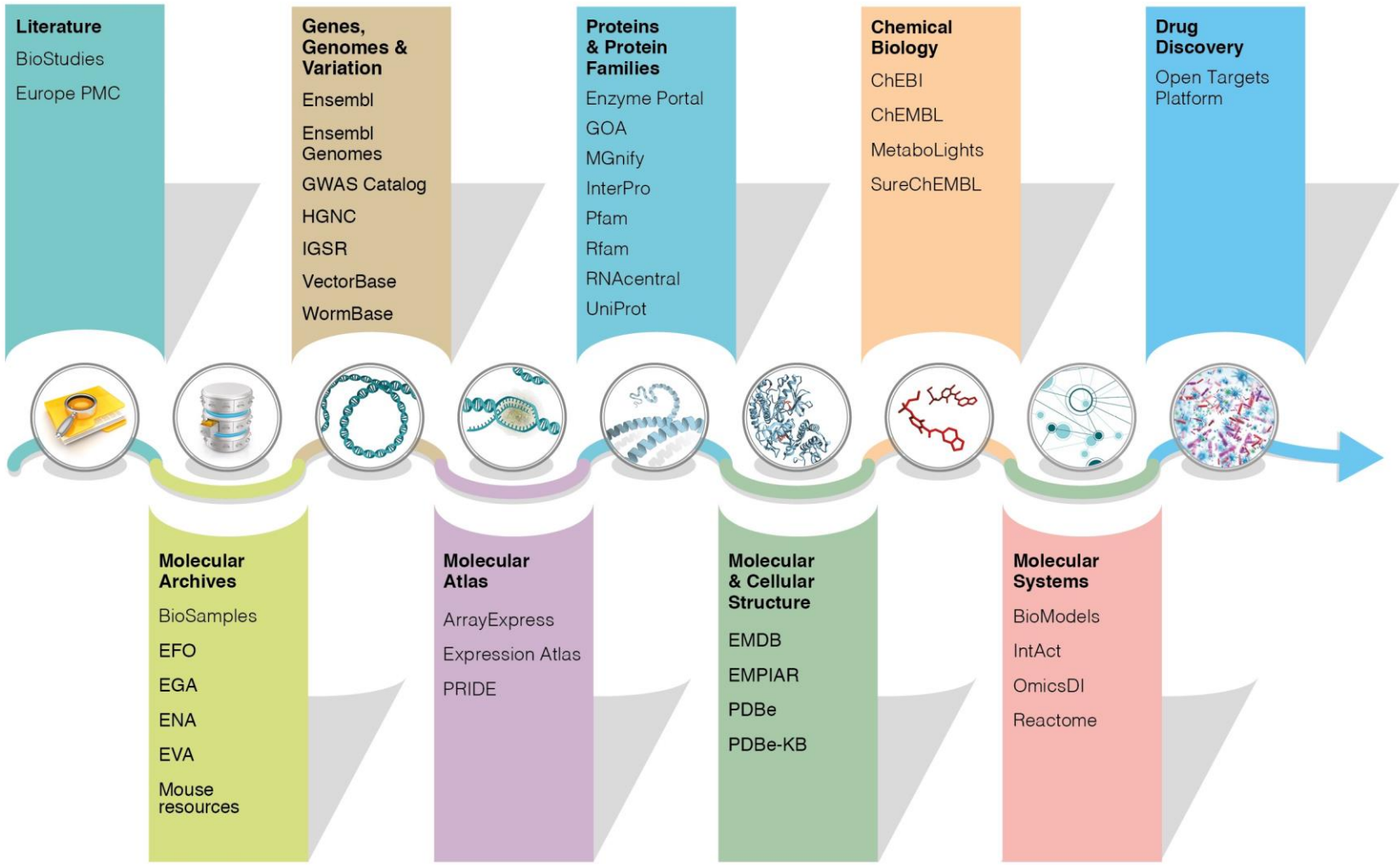
Map of Scientific Collaborations from 2008-2014

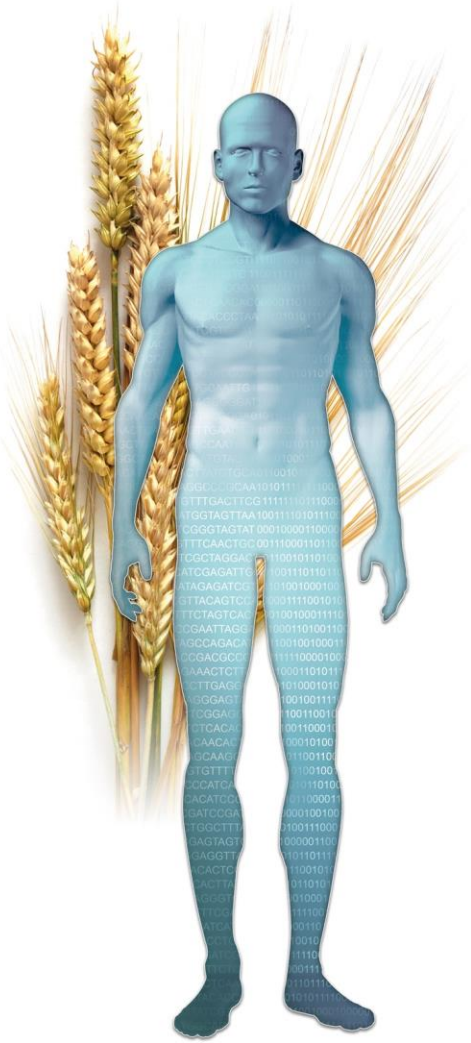
What is EMBL-EBI?

- Europe’s home for biological data services, research and training
- A trusted data provider for the life sciences
  - 150 Petabytes of storage (0.1 exabytes)
  - >40,000 CPU Cores
- Part of the European Molecular Biology Laboratory, an intergovernmental research organisation
- International: 650 members of staff from 66 nations
- Home of the ELIXIR Technical hub (Elixir is a distributed infrastructure for life-science information)



Data resources at EMBL-EBI





EMBL-EBI role in Genomics medicine

- Providing reference datasets for clinical research
- Working collaboratively with partners in technology transfer
  - Building interactions with relevant players both institutes and companies
  - International coordination for open human disease data initiatives worldwide
  - Developing software for managing clinical data and depositing research results
- Communication and encouragement in Europe's nation states
  - Ideal: "Institute/network for Biomedical informatics" in each country
  - Build a global international integrated Medical Bioinformatics Research Infrastructure with strong pan-European participation

EUROPEAN GENOME-PHENOME ARCHIVE
10TH ANNIVERSARY

EGA Original model

The EGA provides a service for the permanent **archiving and distribution** of **personally identifiable genetic and phenotypic data** resulting from biomedical research projects

IHEC International Human Epigenome Consortium

UK 10K

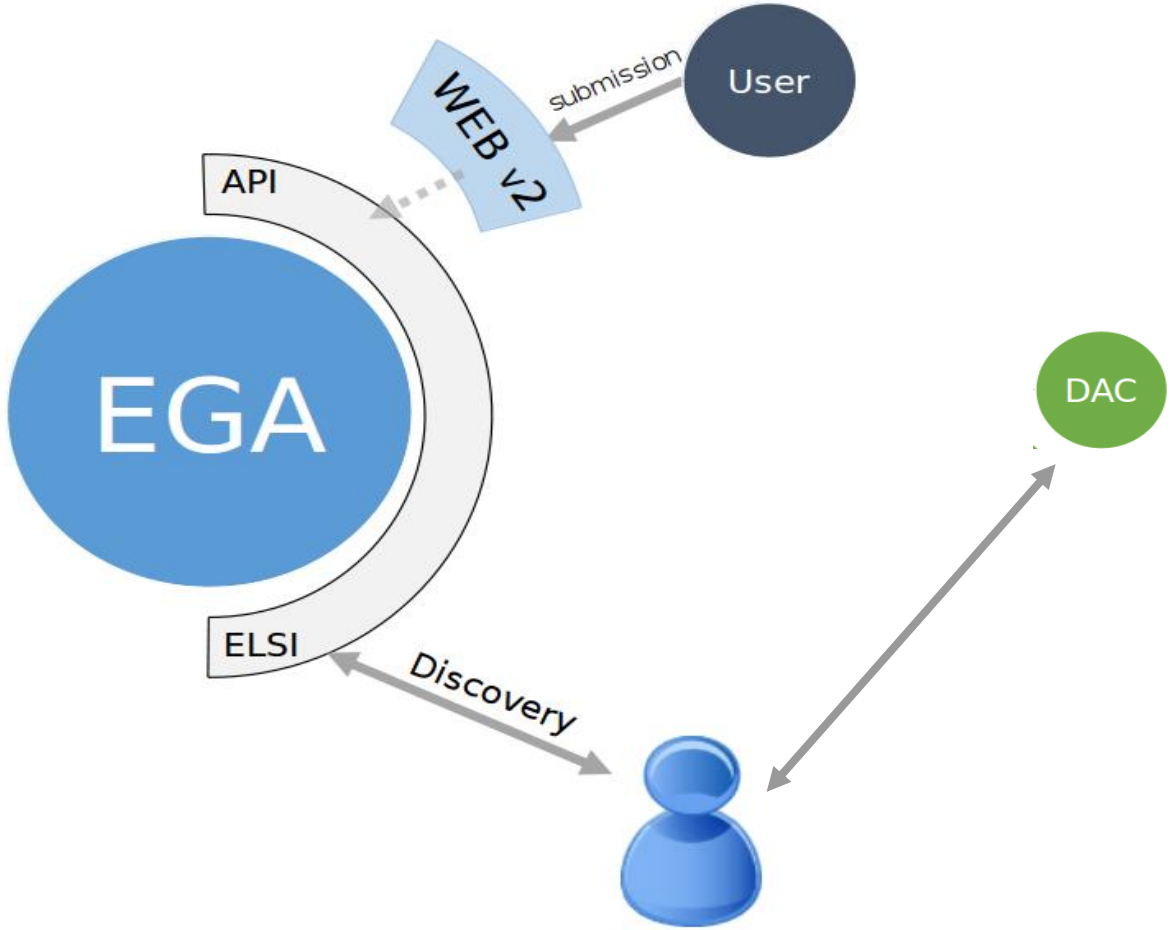
biobank uk
Improving the health of future generations

International Cancer Genome Consortium

Wellcome Trust Case Control Consortium

HipSci

BLUEPRINT epigenome

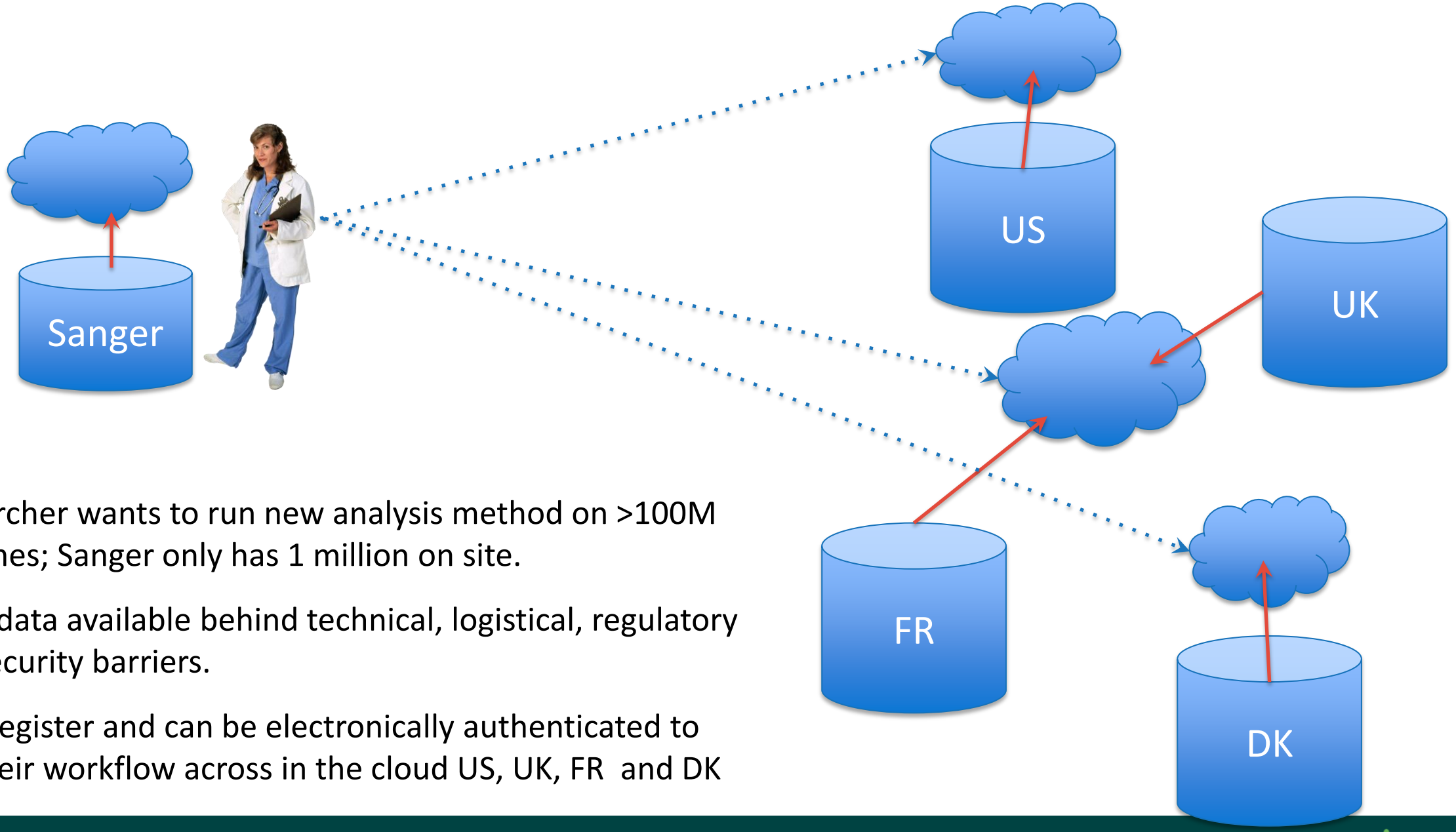


The solution is federation

“a grouping of autonomous organisations and datasets with a centralized control”

It allows us to….

1. move analysis to data, not aggregate data close to each researcher
2. have broad, reciprocal data access methods which respect national processes and patient consent
3. transfer methods and skills into the healthcare sector
4. leverage healthcare data to make more discoveries on humans



- Researcher wants to run new analysis method on >100M genomes; Sanger only has 1 million on site.
- More data available behind technical, logistical, regulatory and security barriers.
- They register and can be electronically authenticated to run their workflow across in the cloud US, UK, FR and DK

The GA4GH Mission

Global Alliance for Genomics & Health

LAUNCH of GA4GH in 2013

The Global Alliance for Genomics and Health aims to accelerate progress in genomic science and human health by developing standards and framing policy for responsible genomic and health-related data sharing.

The GA4GH Ecosystem

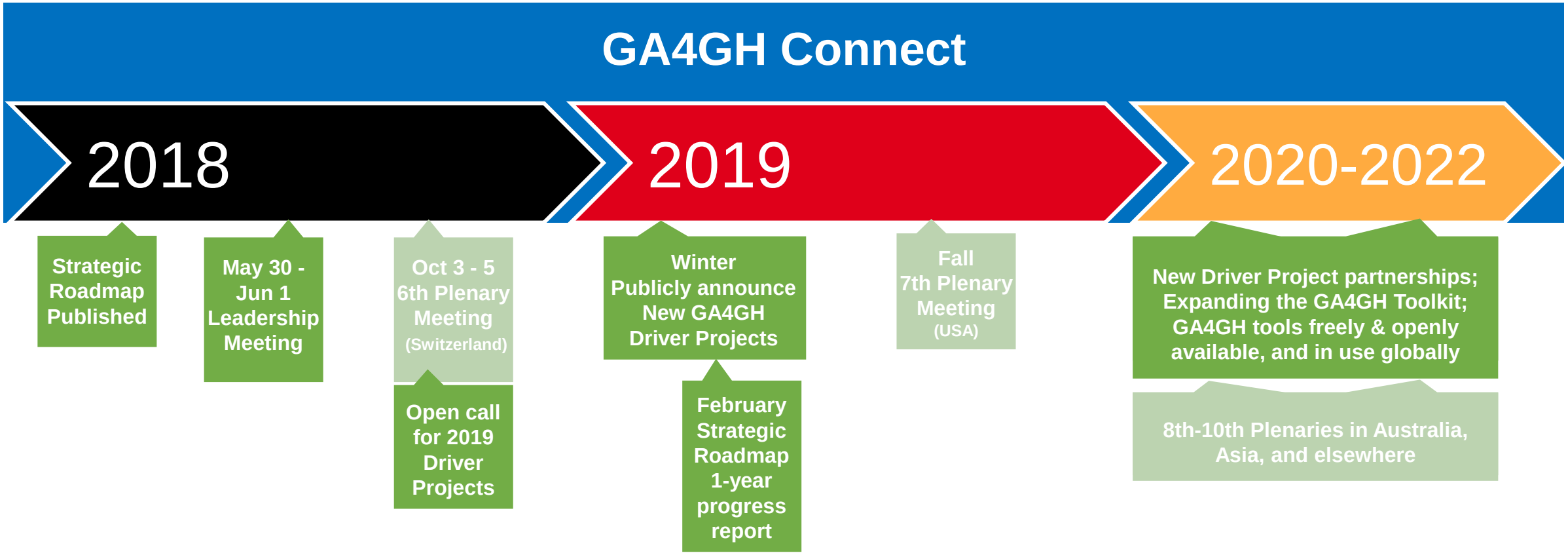
Global Alliance for Genomics & Health

Global Alliance members include:

1. Universities and research institutes (22%)
1. Academic medical centers and health systems (10%)
1. Disease advocacy organizations and patient groups (4%)
1. Consortia and professional societies (13%)
1. Funders and agencies (5%)
1. Life science and information technology companies (46%)

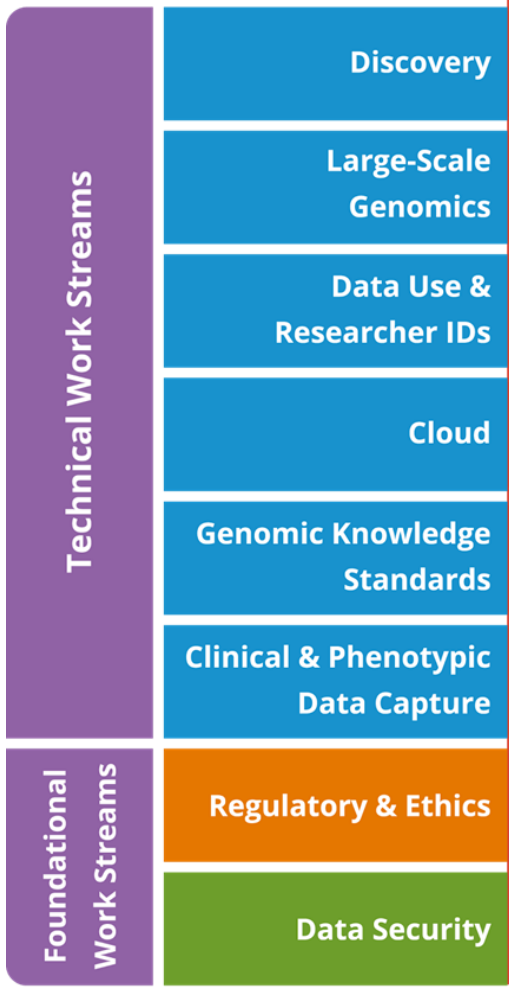
GA4GH: 2018 Onwards

Global Alliance for Genomics & Health



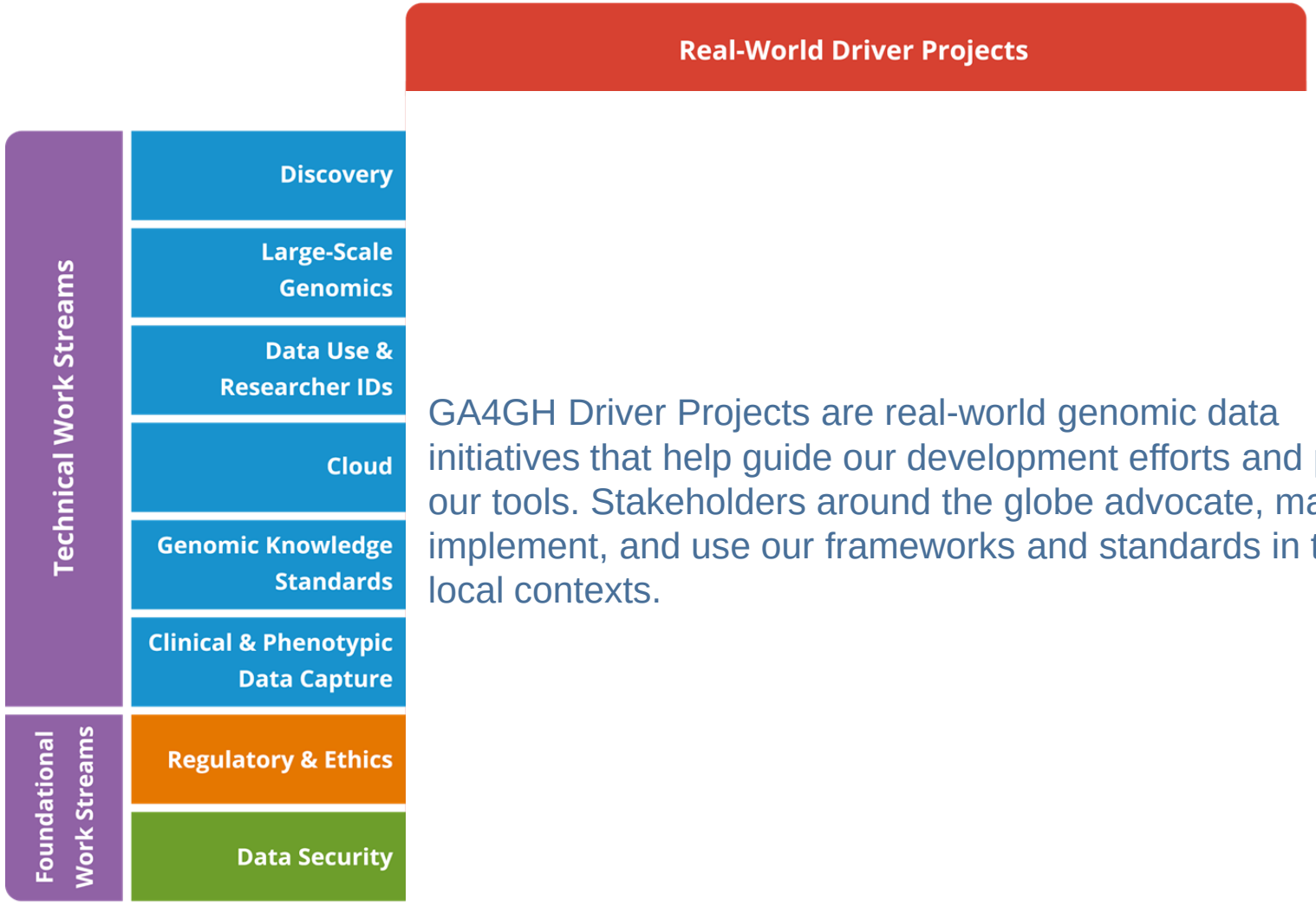
EMBL-EBI

GA4GH Connect



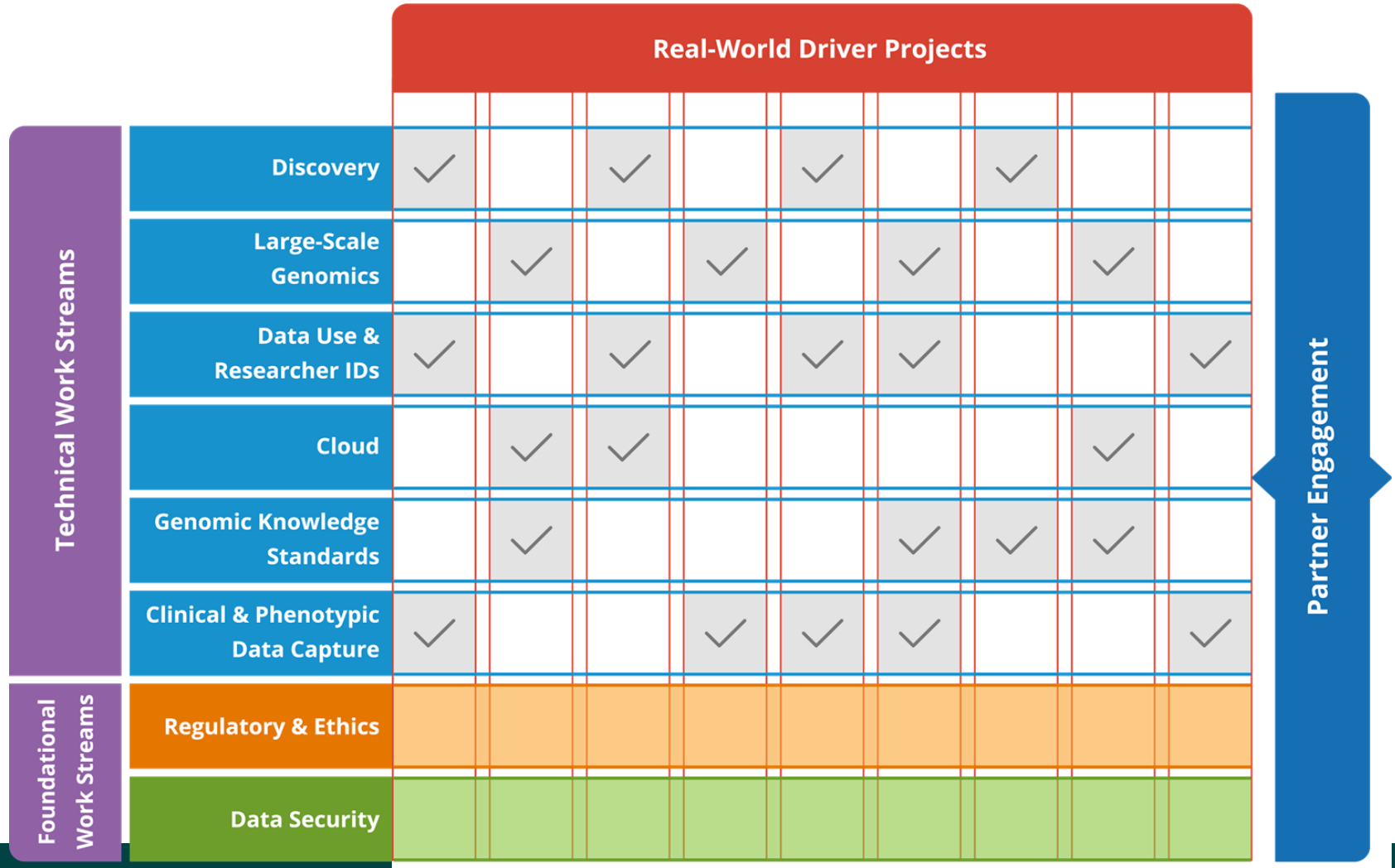
- Develop standards, tools, and frameworks that are designed to overcome technical and regulatory hurdles to international genomic data-sharing
- "Community-minded" leaders with bandwidth to ensure delivery of tools at the expected rate
- Foundational Work Streams provide guidance to both Technical Work Streams and Driver Projects in the areas of regulatory, ethics, and data security in genomics.

GA4GH Connect



GA4GH Driver Projects are real-world genomic data initiatives that help guide our development efforts and pilot our tools. Stakeholders around the globe advocate, mandate, implement, and use our frameworks and standards in their local contexts.

GA4GH Connect



The GA4GH Partner Engagement initiative facilitates two-way dialogue with the international community, including national initiatives, major health care centres, and patient advocacy groups.

GA4GH 2019 Driver Projects

All of Us Research Program
United States

Australian Genomics
Australia

BRCA Challenge
International

CanDIG
Canada

ClinGen
United States

ELIXIR Beacon
Europe

ENA/EVA/EGA
Europe

Genomics England
United Kingdom

Human Cell Atlas
International

ICGC-ARGO
International

Matchmaker Exchange
International

Monarch Initiative
International

National Cancer Institute (NCI)
United States

TOPMed
United States

VICC
International

EUCAN Cancer
International

Autism Sharing Initiative
International

EpiShare
International

GEM Japan
Japan

Swiss Personalized Health Network
Switzerland

European Joint Program For Rare Diseases
Europe

H3Africa
Pan-Africa

The Current GA4GH Toolkit

Genomic Data

- Read File Formats – **BAM/SAM/CRAM**
- Variant File Formats - **VCF/BCF**
- GA4GH Streaming API – **htsget**
- Reference Sequence Retrieval API – **refget API**
- **Beacon API** – querying for variants
- **Data Use Ontology (DUO)** – tag datasets with usage restrictions
- **Workflow Execution Service (WES)** – running portable workflows

The Current GA4GH Toolkit

Regulatory & Ethics

- Framework for Responsible Sharing of Genomic and Health Related Data
- Accountability Policy
- Data Sharing Lexicon
- Consent Policy

Data Security

- Security Infrastructure
- Privacy and Security Policy

Discovery Workstream

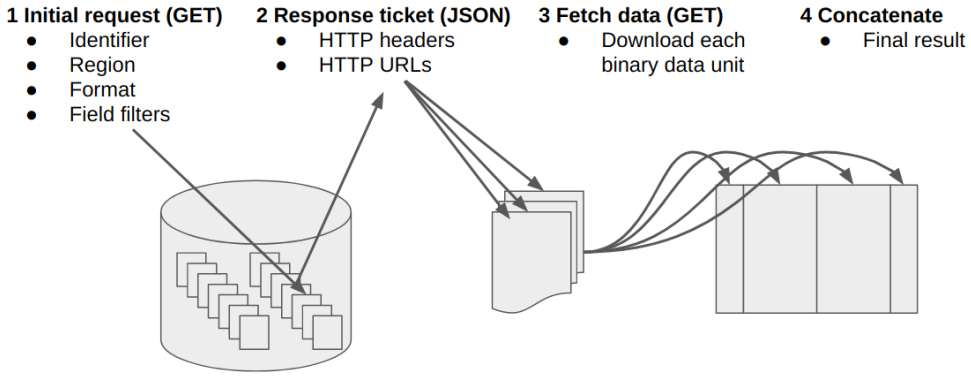
Our focus is on building standards for **federated, secured networks** of data and services, forming an "Internet of Genomics" (IoG), and **asking meaningful questions** across it.

Mark Fiume (DNAstack)
Harindra Arachchi (Broad)

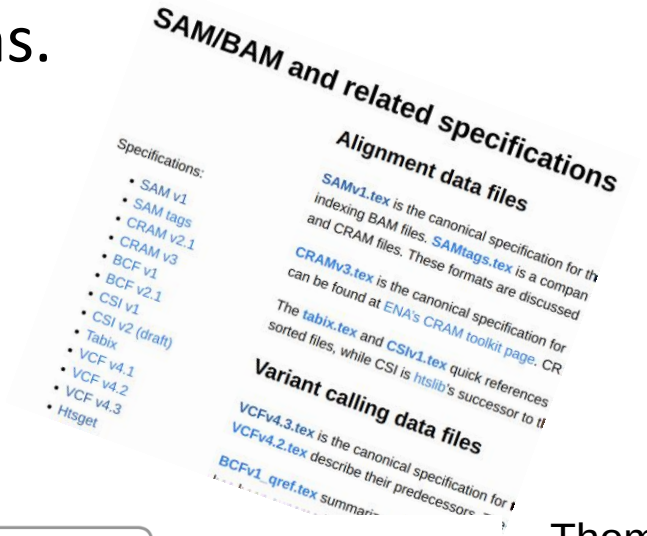
Large Scale Genomics Workstream

- *Create standardized methods for accessing large-scale genomic data by file-based, API-based, cloud-based, and distributed access.*

- Deliverables prioritized based on engagement with driver projects, key partners, and other work streams.



htsget v1.0



Thomas Keane, EMBL-EBI
Oliver Hofmann, University of Melbourne

EMBL-EBI

Data Use & Researcher Identity Workstream

Goal: Define the standards, both regulatory and technical, to facilitate both axes of access control (AAI & Data Use)

Data Use Restrictions: What are you doing with the data?

"The donor wants her data used only for non-commercial cancer research"

		Satisfies Data Use Restrictions	
		Yes	No
Appropriate Permissions	Yes	Access	No access
	No	No access	No access

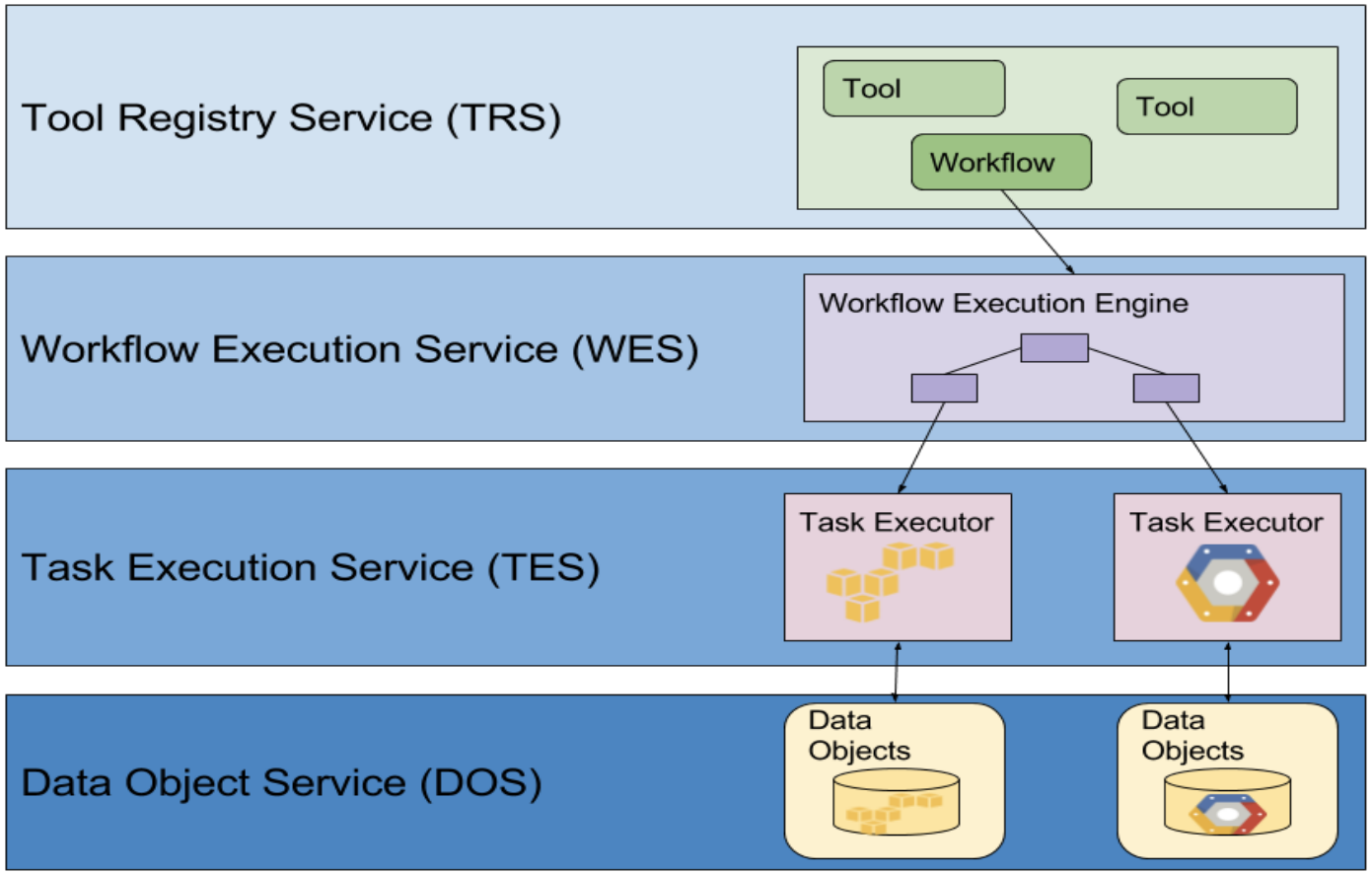
Authentication and Authorization Infrastucture (AAI): Who are you?

"Only consortium members can access this data."

Ravi Pandya (Microsoft)
Anthony Philippakis (Broad)

Cloud Workstream

Bringing algorithms to the data' by creating standards for portable workflows



v1.0.0 - Dockstore.org

v0.1 - CWL workflow-service

v0.2 - Funnel

Pre-release - inspired by HCA, ICGC, GDC, etc

David Glazer (Verily)
Brian O'Connor (UCSC)

Genome Knowledge Standards Workstream

Standards-based components for exchange of genomic information

- **Variant annotation: data model** to guide the linkage of annotations and structured clinical interpretations to variant data
- **Variant representation: data model/specification** extensible data model and message schema specification for the representation of variants

Andy Yates (EMBL-EBI)
Bob Freimuth (Mayo clinic)

GA4GH Leadership 2018

Exec

Ewan Birney
EMBL-European Bioinformatics Institute
Hinxton, United Kingdom
Chair, GA4GH
Member, Steering Committee

Peter Goodhand
Ontario Institute for Cancer Research
Toronto, Canada
Chief Executive Officer, GA4GH
Member, Steering Committee

Heidi Rehm
The Broad Institute
Cambridge, United States
Vice-Chair, GA4GH
Member, Steering Committee

Kathryn North
Murdoch Childrens Research Institute
Melbourne, Australia
Vice-Chair, GA4GH
Member, Steering Committee
Lead, Partner Engagement Initiative
Driver Project Champion, Australian Genomics

Work Stream Leads & Driver Project Champions

Brian O'Connor (Cloud | TopMed)
David Glazer (Cloud | All of Us)
Marc Fiume (Discovery)
Michael Baudis (Discovery | ELIXIR Beacon)
Thomas Keane (Large-Scale |EVA/ENA/EGA)
Oliver Hofmann (Large-Scale)
Tommy Nyrönen (DURI)
Moran Cabili (DURI)
Robert Freimuth (Genomic Knowledge)
Andy Yates (Genomic Knowledge)
David Hansen (Clinical & Phenotypic)
Melissa Haendel (Clinical & Phenotypic | Monarch Initiative)
Dixie Baker (Data Security)
Jean-Pierre Hubaux (Data Security)
Bartha Knoppers (Regulatory & Ethics)
Madeleine Murtagh (Regulatory & Ethics)

Amanda Spurdle (BRCA Challenge)
Gunnar Rätsch (BRCA Challenge)
Melissa Cline (BRCA Challenge)
Kym Boycott (Matchmaker Exchange)
Ada Hamosh (Matchmaker Exchange)
Heidi Rehm (ClinGen)
Serena Scollen (ELIXIR Beacon)
Ilkka Lappalainen (ELIXIR Beacon)
Obi Griffith (VICC)
Malachi Griffith (VICC)
David Tamborero (VICC)
Augusto Rendon (Genomics England)
Peter Counter (Genomics England)
Anthony Philippakis (All of Us)
Mike Brudno (CanDIG)
Steven Jones (CanDIG)
Guillaume Bourque (CanDIG)

Robert Grossman (NCI GDC)
Kathryn North (Australian Genomics)
Clara Gaff (Australian Genomics)
Christina Yung (ICGC-ARGO)
Lincoln Stein (ICGC-ARGO)
Goncalo Abecasis (TOPMed)
Tim Tickle (Human Cell Atlas)
Laura Clarke (Human Cell Atlas)
Jordi Rambla (ENA/EVA/EGA)
Dylan Spalding (ENA/EVA/EGA)
Cristina Gonzales (ENA/EVA/EGA)
Peter Robinson (Monarch Initiative)
Tudor Groza (Monarch Initiative)

EMBL-EBI


Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.
HOME
ACCOMMODATION
VENUE
GA4GH 7th Plenary Meeting
October 21-23, 2019
The Hynes Convention Center
Boston, USA
REGISTER

EMBL-EBI

Conclusions

- We are on the cusp of a massive adoption of genomics across health providers
- EMBL-EBI provides infrastructure to ensure that the data can be re-used to create knowledge, and worldwide data analysis network
- EMBL-EBI is involved in Driver Projects through its molecular archives (ENA, EGA and EVA), membership in ELIXIR and collaboration in the Human Cell Atlas data-coordination platform
- GA4GH is developing and testing the data standards needed to bring genomics into the clinic and enable precision medicine