Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

Data standards and global variant databases

Thomas Keane (@drtkeane)
Head of European Genome-phenome Archive
EMBL-EBI

Genomic standards

Global Alliance for Genomics & Health

Fast Healthcare Interoperability Resource

FHIR®

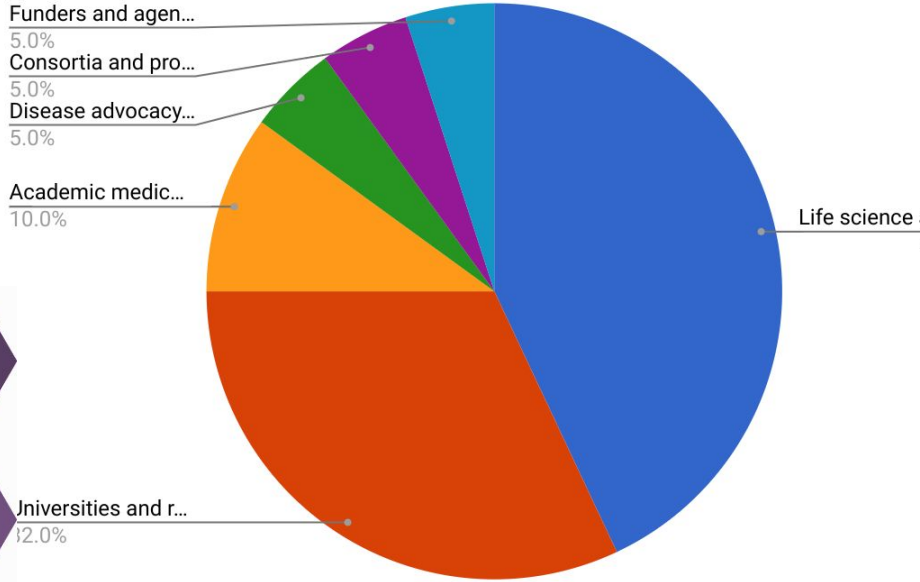
- Interoperability standard for exchange of healthcare information

MPEG

MPEG

- MPEG-G: standard for compression of genomic information

Global Alliance for Genomics and Health (GA4GH)



GA4GH mission

Global Alliance for Genomics & Health

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.

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

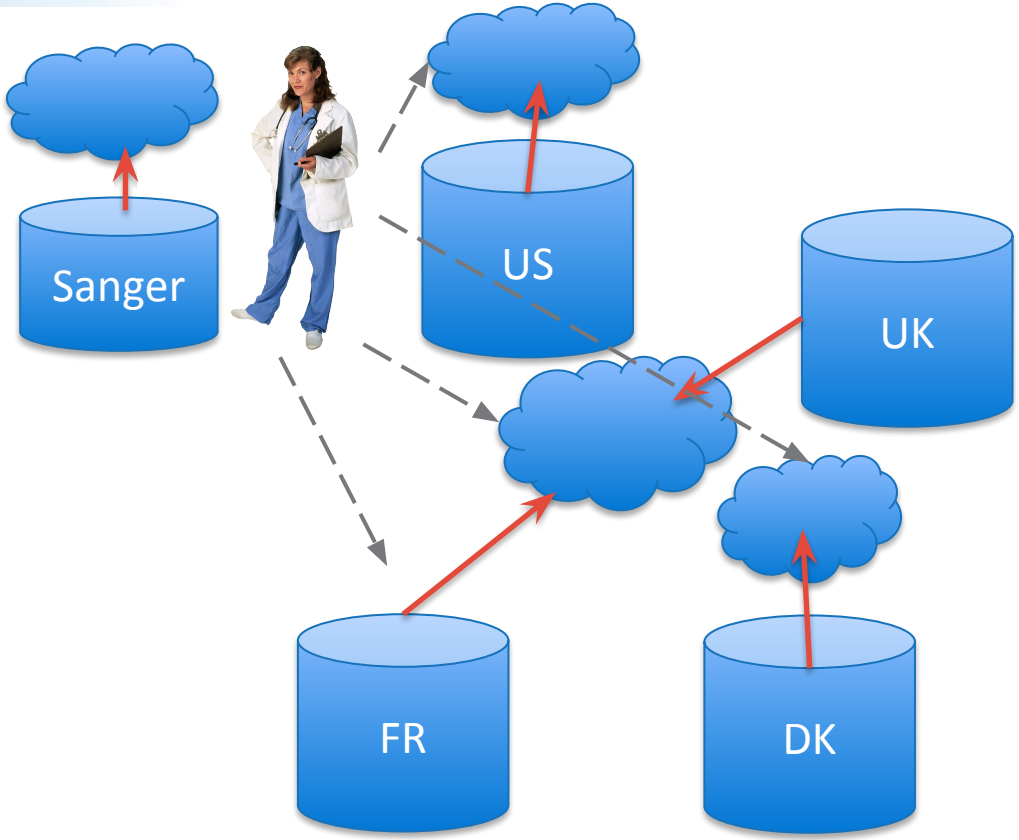
Peter Goodhand
Ontario Institute for Cancer Research
Toronto, Canada
Chief Executive Officer, GA4GH
Member, Steering Committee

David Haussler
University of California Santa Cruz
Santa Cruz, 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

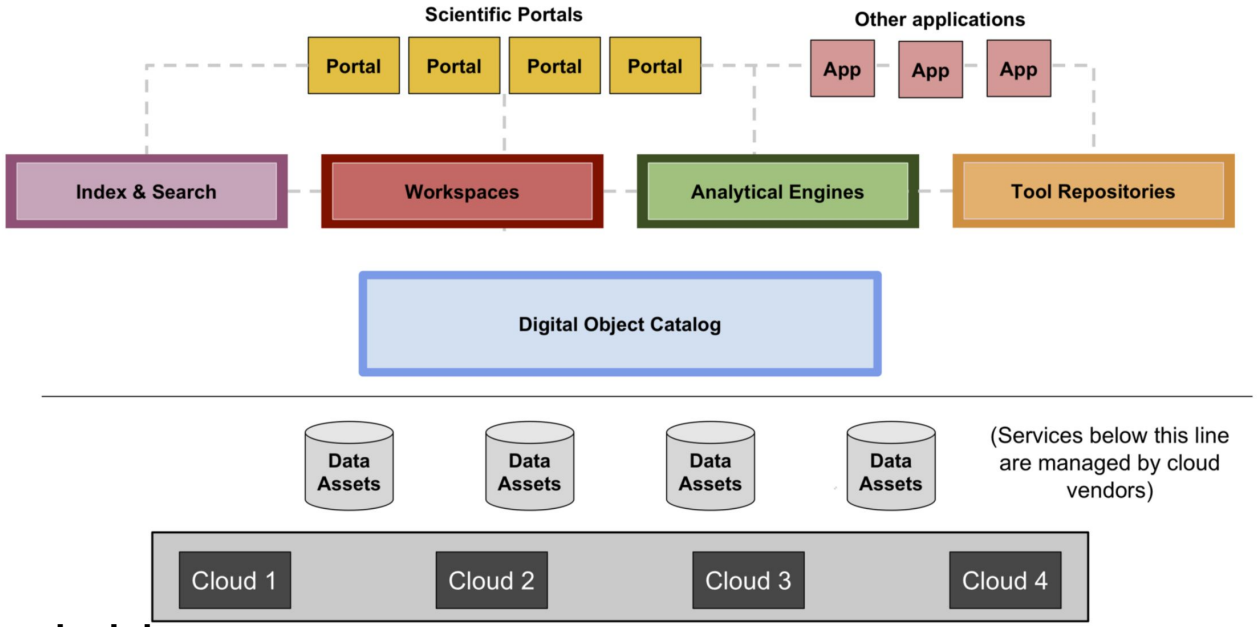
Where do we want to get to?

Global Alliance
for Genomics & Health



A Data Biosphere for Biomedical Research

Global Alliance for Genomics & Health



Data Biosphere principles

- (1) modular, functional components + interfaces
- (2) community-driven
- (3) open, open-source licenses
- (4) standards-based, to ensure interoperability

https://medium.com/@benedictpaten

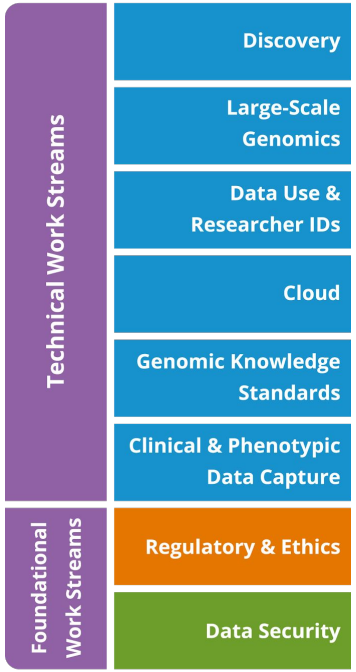
Core Mantra for GA4GH

Global Alliance for Genomics & Health

Standardise on interfaces, compete on implementations

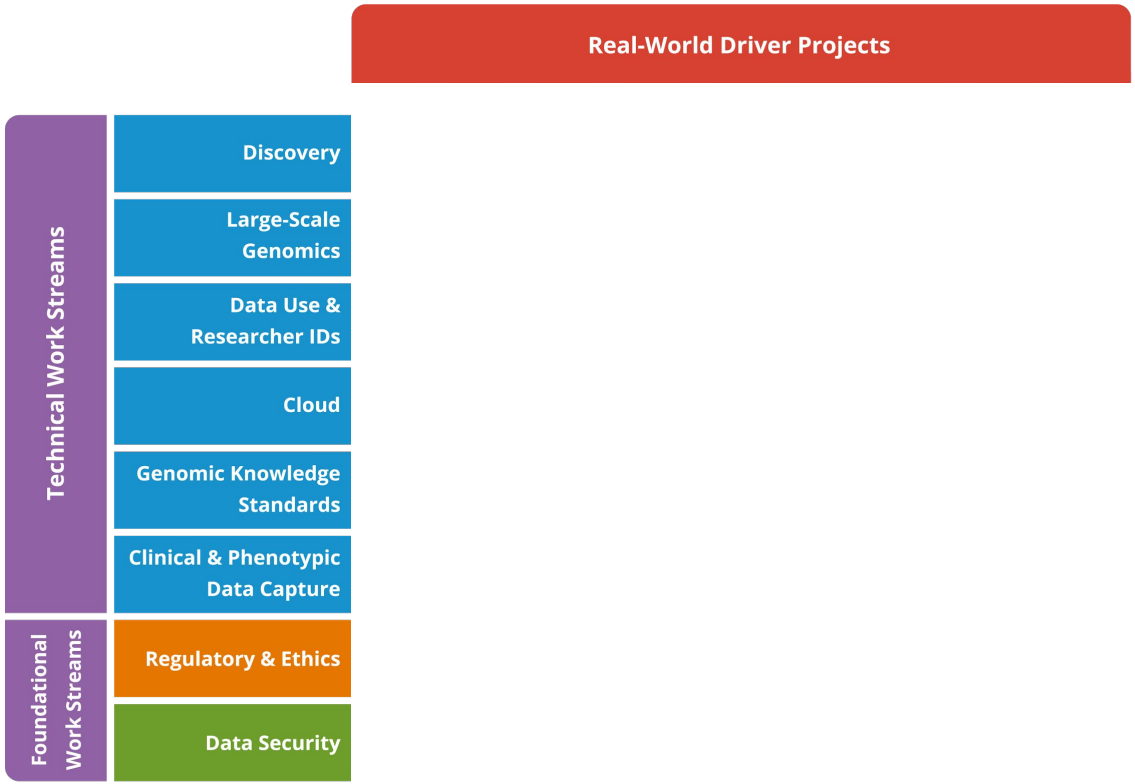
GA4GH Connect

Global Alliance for Genomics & Health



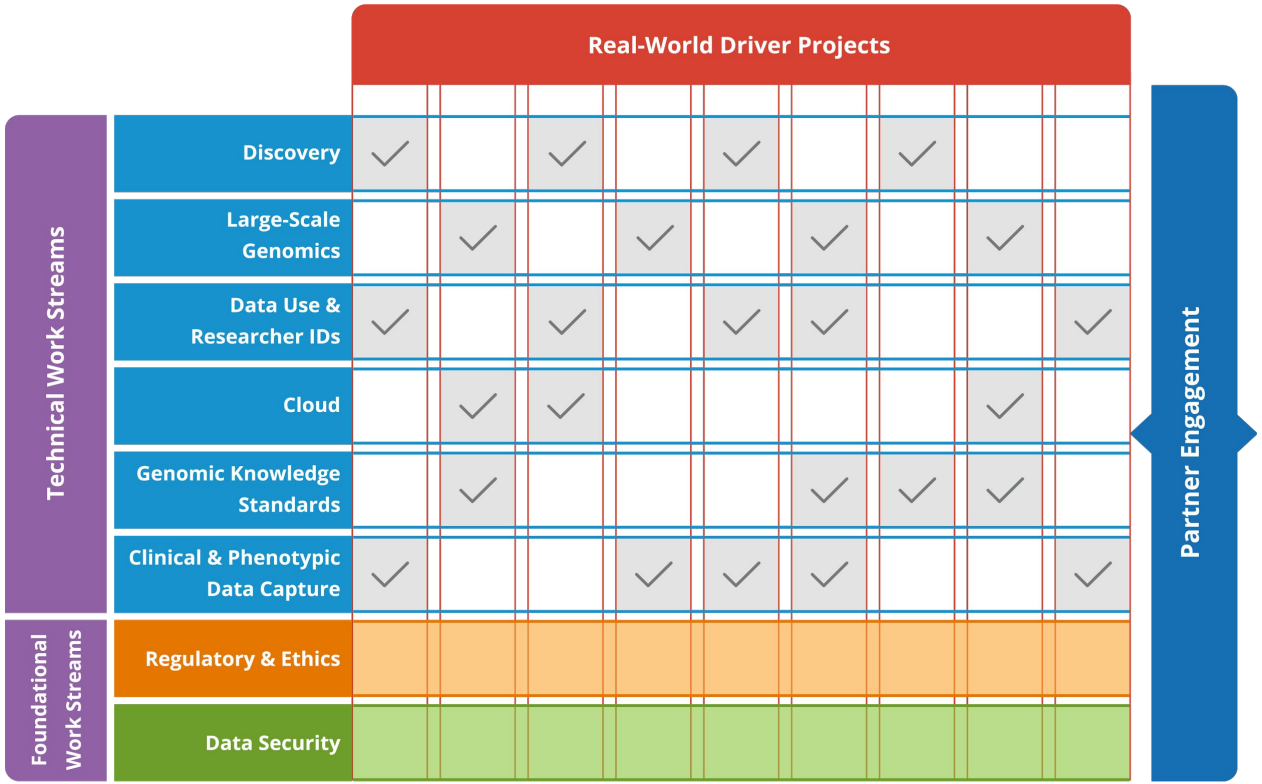
GA4GH Connect

Global Alliance for Genomics & Health



GA4GH Connect

Global Alliance for Genomics & Health



2017 Driver Projects

Global Alliance for Genomics & Health

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 Genomic Data Commons
United States

TOPMed
United States

Variant Interpretation for Cancer Consortium
International

Data Use and Researcer ID Workstream

Global Alliance for Genomics & Health

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

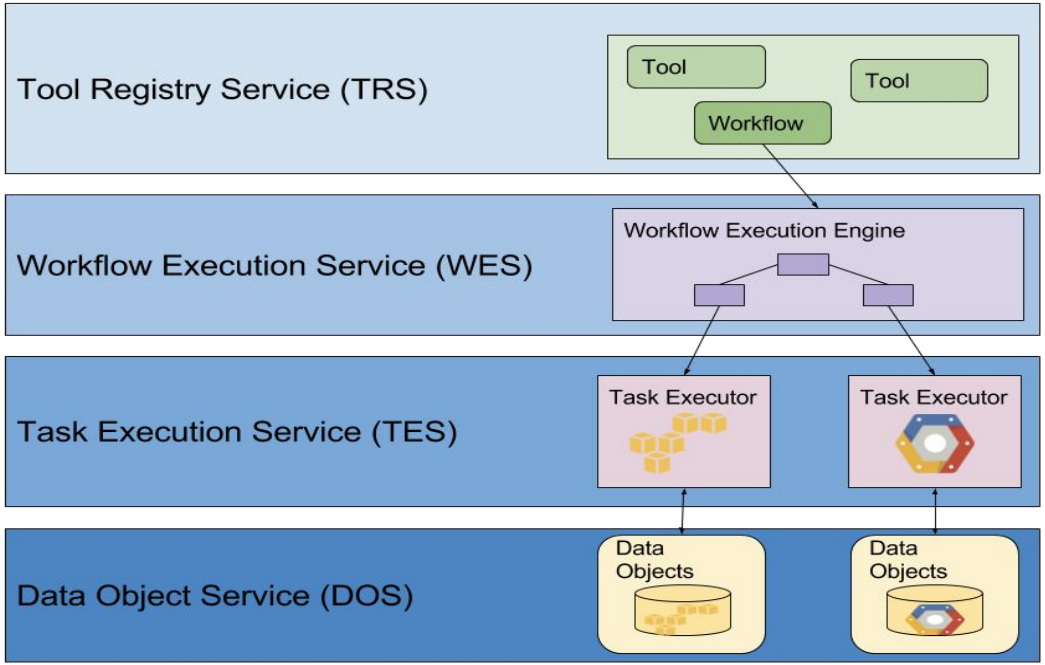
Authentication and Authorization Infrastucture (AAI): Who are you?

"Only consortium members can access this data."

Ravi Pandya (Microsoft)
Anthony Philippakis (Broad)

Cloud Workstream

Global Alliance for Genomics & Health



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)

Discovery Workstream

Global Alliance for Genomics & Health

Beacon Network

Matchmaker Exchange
Genomic discovery through the exchange of phenotypic & genotypic profiles

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)

Genome Knowledge Standards Workstream

Global Alliance
for Genomics & Health

- Focus Areas and Driver Project feedback
  - Variant Representation: create a standard model for computer readable variant representation
  - Variant Annotation: represent and link annotations, including their evidence and provenance, to variants

- **June 2018**: Continue current VMC spec development and start to extend into structural variants, engage DPs as pilot projects.
- **Dec 2018**: Refine due to pilot feedback, expand structural variant capability, expand approach to allele equivalence
- **June 2019**: Review and expand VMC based on feedback, publish manuscripts with both structural & allele equivalence foci.

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

Clinical and Phenotypic Data Capture Workstream

Global Alliance for Genomics & Health

- Focus Areas: 3 sub-groups
  - Data representation & interoperability (Melanie Courtot, Tudor Groza, Allison Heath)
  - Data exchange & interoperability (Grant Wood, Michael Baudis, Gil Alterovitz)
  - Implementation/Education/Engagement (Andrea Ramirez, John Mattison)

- How were the following deliverables prioritized based on engagement with Driver Projects and other Work Streams
  - Leads and sub-group leads directly linked to Driver Projects
    - Melissa Haendel & Tudor Groza → *Monarch Initiative*
    - David Hansen → *Australian Genomics*
    - Michael Baudis → *ELIXIR Beacon*
    - Andrea Ramirez → *All of Us*

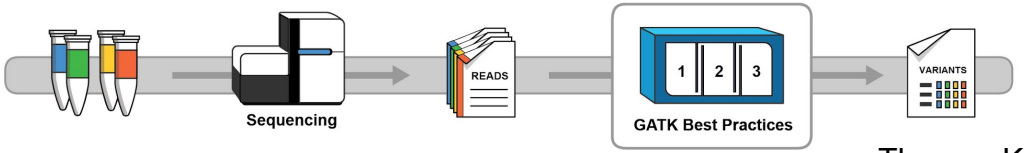
David Hansen (Australian Genomics)
Melissa Haendel (Monarch Initiative)

Large Scale Genomics Workstream

Global Alliance for Genomics & Health

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

- Key stakeholders
  - Platform vendors, clinical genomics platforms, genome sequencing providers, bioinformatics developers, academic sequencing centers



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

NGS data formats (2009-date)

Global Alliance for Genomics & Health

Legacy of the 1000 Genomes Project

- Standardisation at the file level
- Few population scale projects

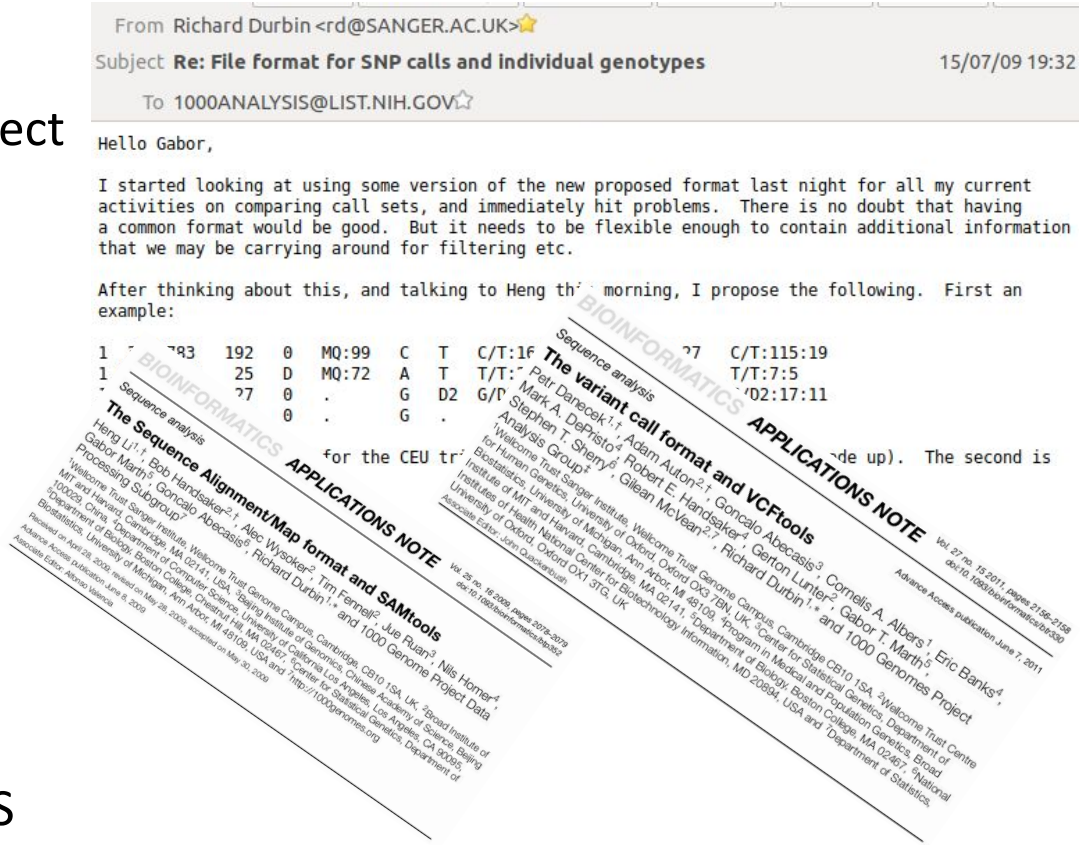
Sequencing reads

- SAM/BAM (2009-)
- CRAM (2016-)

Genetic variation

- VCF/BCF (2010-)

Enabled many applications of NGS


From Richard Durbin <rd@SANGER.AC.UK>

Subject **Re: File format for SNP calls and individual genotypes** 15/07/09 19:32

To 1000ANALYSIS@LIST.NIH.GOV

Hello Gabor,

I started looking at using some version of the new proposed format last night for all my current activities on comparing call sets, and immediately hit problems. There is no doubt that having a common format would be good. But it needs to be flexible enough to contain additional information that we may be carrying around for filtering etc.

After thinking about this, and talking to Heng th morning, I propose the following. First an example:

1	783	192	0	MQ:99	C	T	C/T:16	27 C/T:115:19
1		25	D	MQ:72	A	T	T/T:	T/T:7:5
	27	0	.		G	D2	G/D	/02:17:11
		0	.		G	.		

for the CEU tr de up). The second is

BIOINFORMATICS APPLICATIONS NOTE

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,1,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing, 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

Received on April 28, 2009; revised on May 28, 2009; accepted on May 30, 2009

Advance Access publication June 8, 2009

Associate Editor: Ronald Vaienda

BIOINFORMATICS APPLICATIONS NOTE

Sequence analysis

The variant call format and VCFtools

Petr Danecek[1,†], Adam Auton[2,1,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4,5], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, [3]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141, [5]Department of Biology, Boston College, MA 02467, [6]National Center for Biotechnology Information, MD 20894, USA and [7]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK

Vol. 27 no. 15 2011, pages 2156–2158

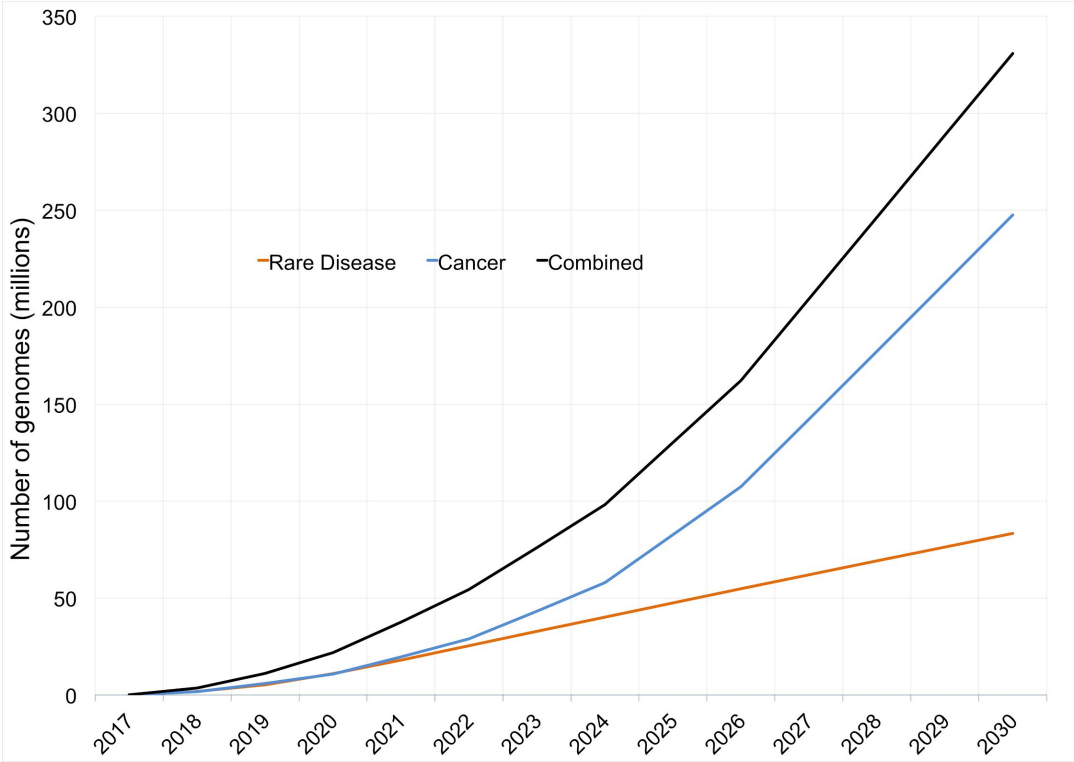
doi:10.1093/bioinformatics/btr330

Advance Access publication June 7, 2011

Associate Editor: Jonathan Wren

How Many Genomes?

Global Alliance for Genomics & Health



Genomics in healthcare: GA4GH looks to 2022

Ewan Birney, Jessica Vamathevan, Peter Goodhand

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

Htsget Streaming API

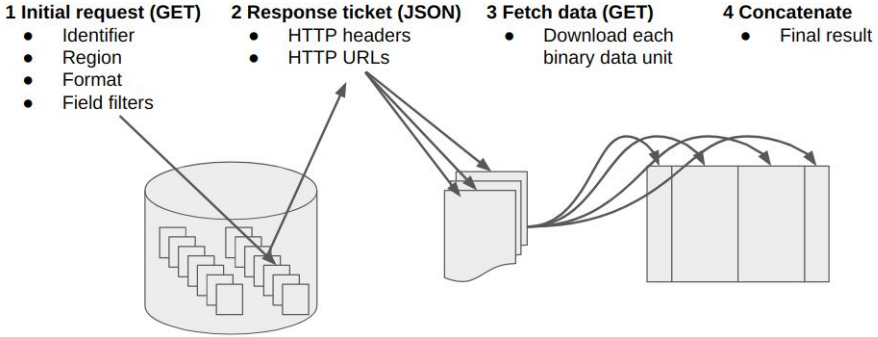
Global Alliance for Genomics & Health

Securely accessing or streaming genomic data

- htsget: A standardised non-file based API for securely streaming genomic data
- Spec maintainer: Mike Lin (DNAnexus)

Milestones

- Launched 1.0 (Oct 17)
- VCF support (Feb 18)
- POST support (Mar 18)
- htsget paper (Jun 18)



	Servers									
	WSI		DAWS		DAZ		GCP		EGA	
Clients	BAM	CRAM	BAM	CRAM	BAM	CRAM	BAM	CRAM	BAM	CRAM
htsget								n/a		n/a
WSI								n/a		n/a
EGA								n/a		n/a
Samtools								n/a		n/a

Table 1: Interoperability status for clients (side) and servers (top). Green boxes indicate fully passing interoperability (see 2 above for details), yellow indicates combination not supported by the server. WSI=Wellcome Sanger Institute, DAWS=DNAnexus on Amazon AWS, DAZ=DNAnexus on Azure, GCP=Google Cloud Platform, EGA=European Genome-phenome Archive. Server URLs are given in section 1 above.

Htsget Streaming API

Global Alliance for Genomics & Health

Who has implemented it so far?

DNAnexus

wellcome trust sanger institute

Google Cloud Platform

Samtools
BCFtools
HTSlib

verily

EUROPEAN GENOME-PHENOME ARCHIVE

igv Integrative Genomics Viewer *

European Variation Archive

ENA European Nucleotide Archive

UNIVERSITY OF OXFORD

Genome Maps

*coming soon

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.

Thank you

https://www.ga4gh.org/

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.