INSTITUTE OF MOLECULAR BIOLOGY

ИНСТИТУТ БИОМЕДИЦИНЫ И ФАРМАЦИИ

ABI

BIG
BIOINFORMATICS GROUP

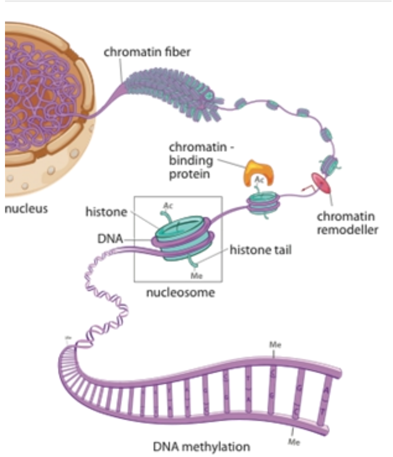
Human
Genomics
Laboratory

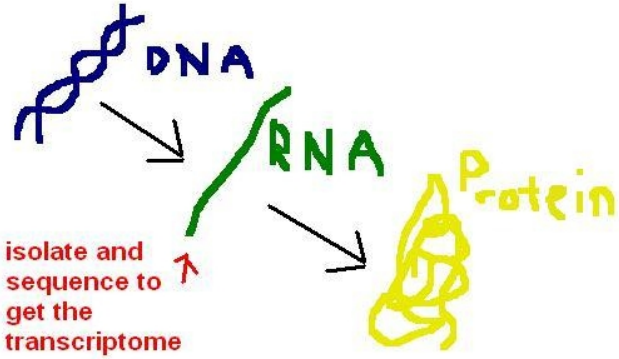
Intro to Bioinformatics

Arsen Arakelyan

11.06.2022, Avetis Informatics Fellowship Program

Information content in biological systems





Genome→ mutations and chromosomal aberrations

→ $3x10^{9}$ base pairs

→ $4x10^{6-7}$ SNPs

Epigenome→ chromatin remodelling

→ $28x10^{6}$ CpG sites

→ ~ 10^{7} nucleosomes

Transcriptome

→ > 10^{3} transcription factors

→ ~ $2x10^{4}$ coding genes

→ >10^{5} proteins

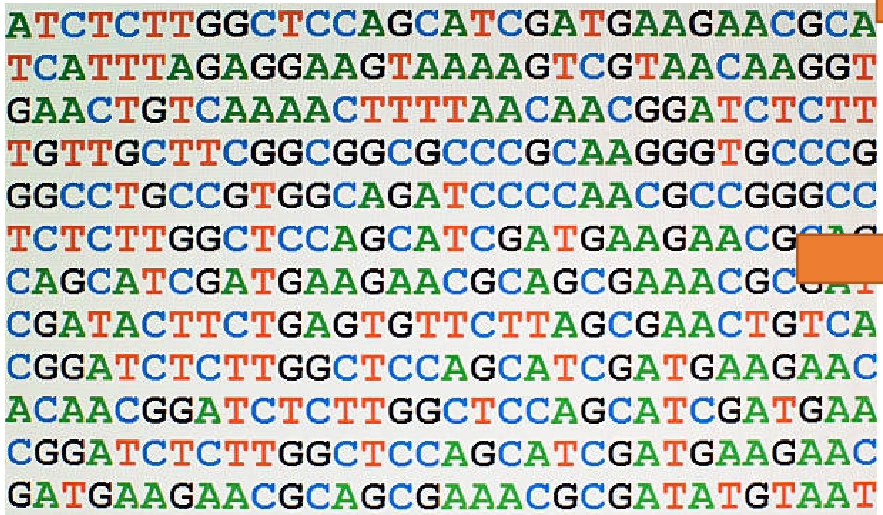
→ >10^{5} non-coding genes

Proteome

Metabolome

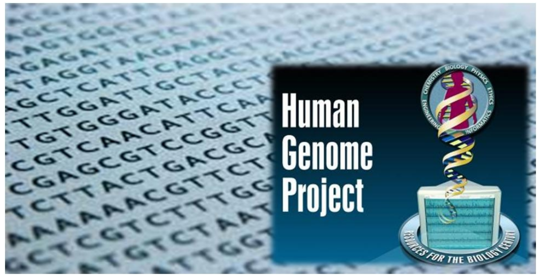
Credit: Hans Binder

Sequencing – uncovering information in DNA



```
ATCTCTTGGCTCCAGCATCGATGAAGAACGCA
TCATTTAGAGGAAGTAAAAGTCGTAACAAGGT
GAACTGTCAAAACTTTTAACAACGGATCTCTT
TGTTGCTTCGGCGGCGCCCGCAAGGGTGCCCG
GGCCTGCCGTGGCAGATCCCCAACGCCGGGCC
TCTCTTGGCTCCAGCATCGATGAAGAACG
CAGCATCGATGAAGAACGCAGCGAAACGC
CGATACTTCTGAGTGTTCTTAGCGAACTGTCA
CGGATCTCTTGGCTCCAGCATCGATGAAGAAC
ACAACGGATCTCTTGGCTCCAGCATCGATGAA
CGGATCTCTTGGCTCCAGCATCGATGAAGAAC
GATGAAGAACGCAGCGAAACGCGATATGTAAT
```

Human genome project and massively parallel data generation technologies



- Human genome project (draft completed in 2003):
  - determining the sequence of human DNA, identifying and mapping all of the genes of the human genome both a physically and functionally

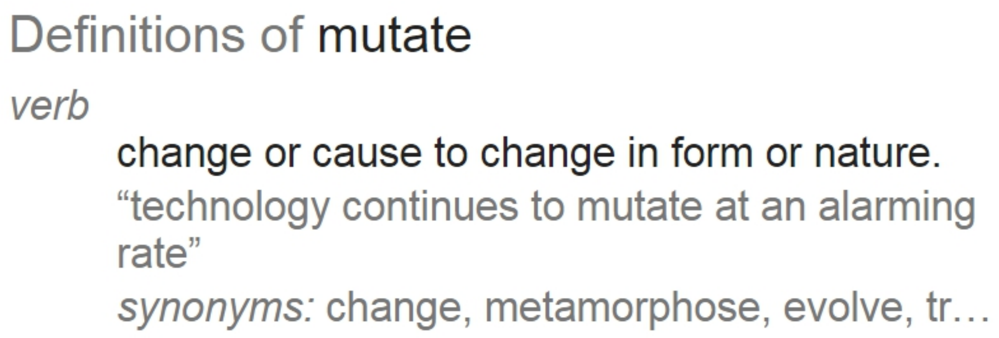
- Human genome project has led to an important inference:
  - Global outlook on complex biological processes occurring at cellular, tissue, or organism level is possible only with parallel assessment of a complete set of spatially and temporarily related molecular data

- Paradigm shift

  From: Few carefully selected parameters in a large sample (low-throughput)

  To: Dozens to hundreds of thousands of parameters from a single sample (high-throughput)

Mutations



Definitions of mutate

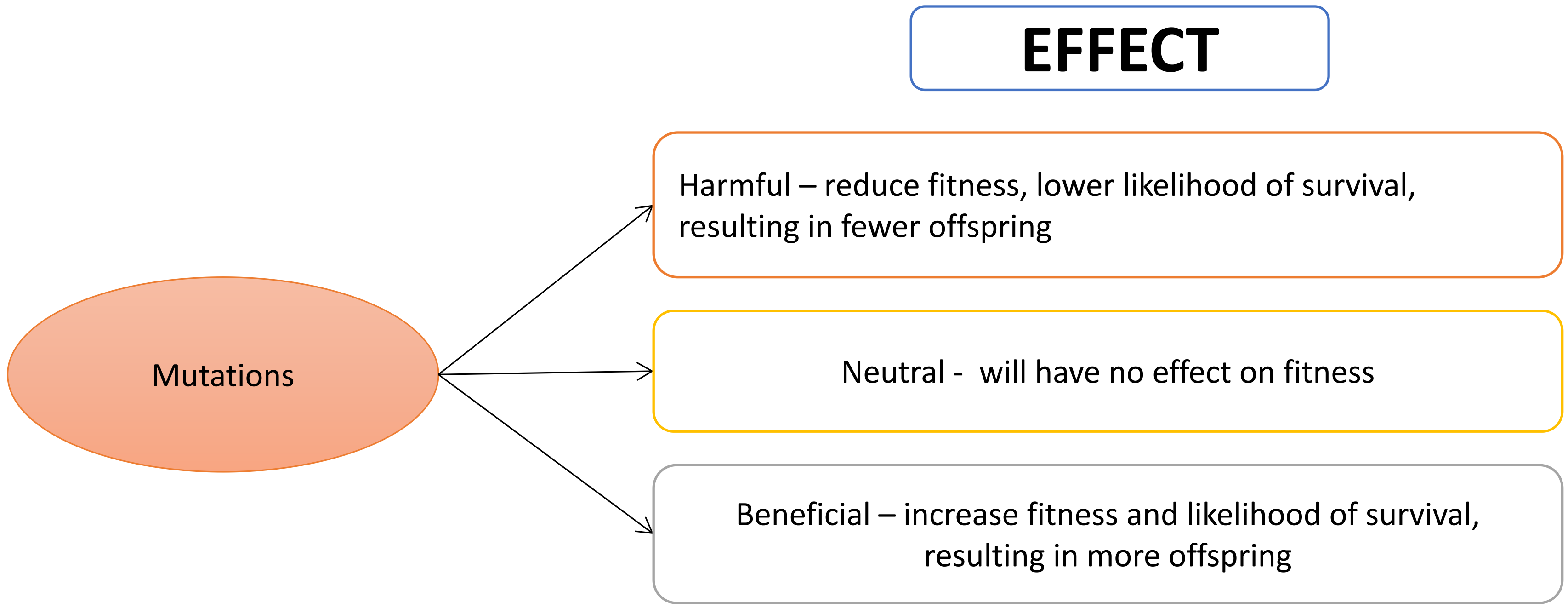
verb

change or cause to change in form or nature.

“technology continues to mutate at an alarming rate”

synonyms: change, metamorphose, evolve, tr…

- A mutation is a change in a DNA sequence.
- A mutation is a source of new alleles.
- A mutation may produce an allele that is selected against, selected for, or selectively neutral.



EFFECT

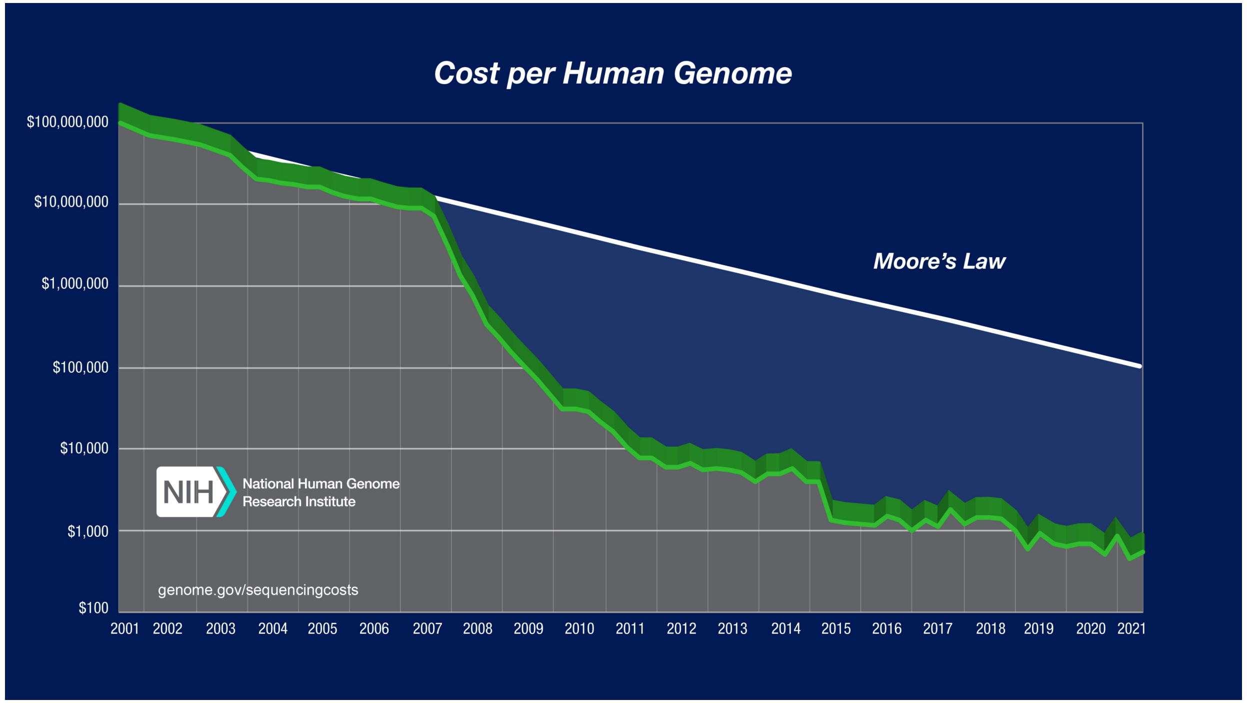
Mutations

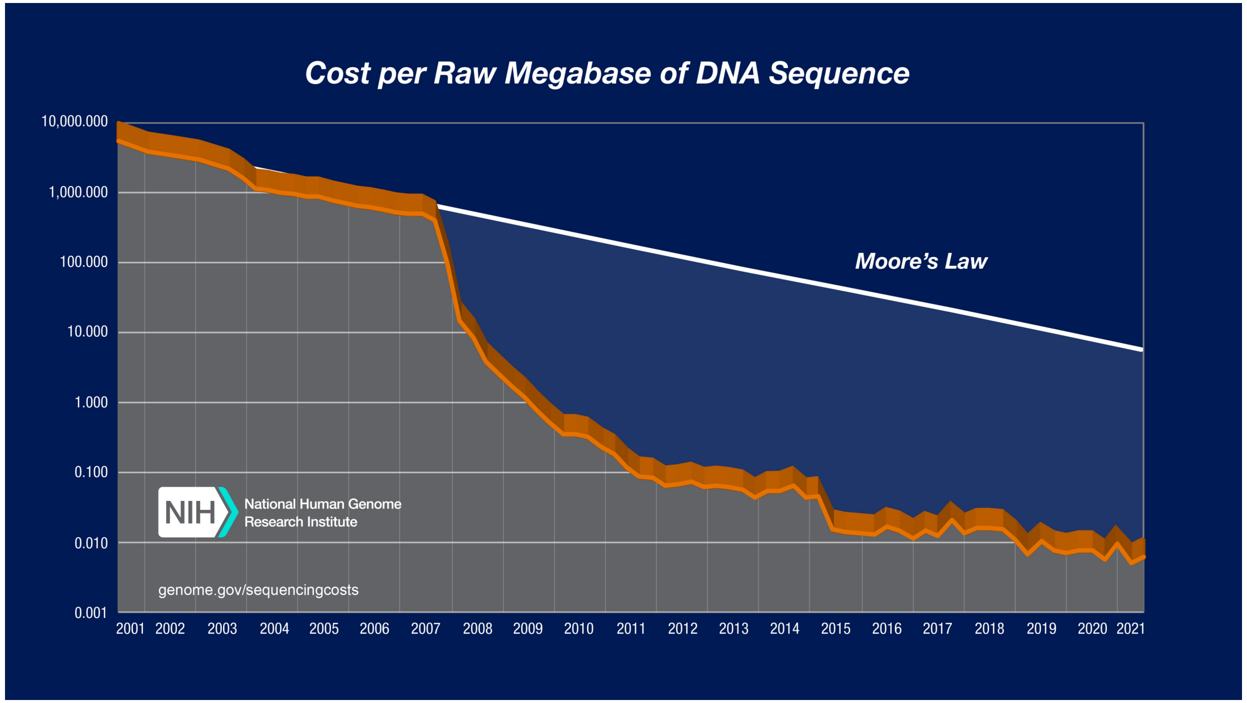
Harmful – reduce fitness, lower likelihood of survival, resulting in fewer offspring

Neutral - will have no effect on fitness

Beneficial – increase fitness and likelihood of survival, resulting in more offspring

Sequencing and microarrays revolutionized biotech





Sequencing technologies and applications in biomedicine

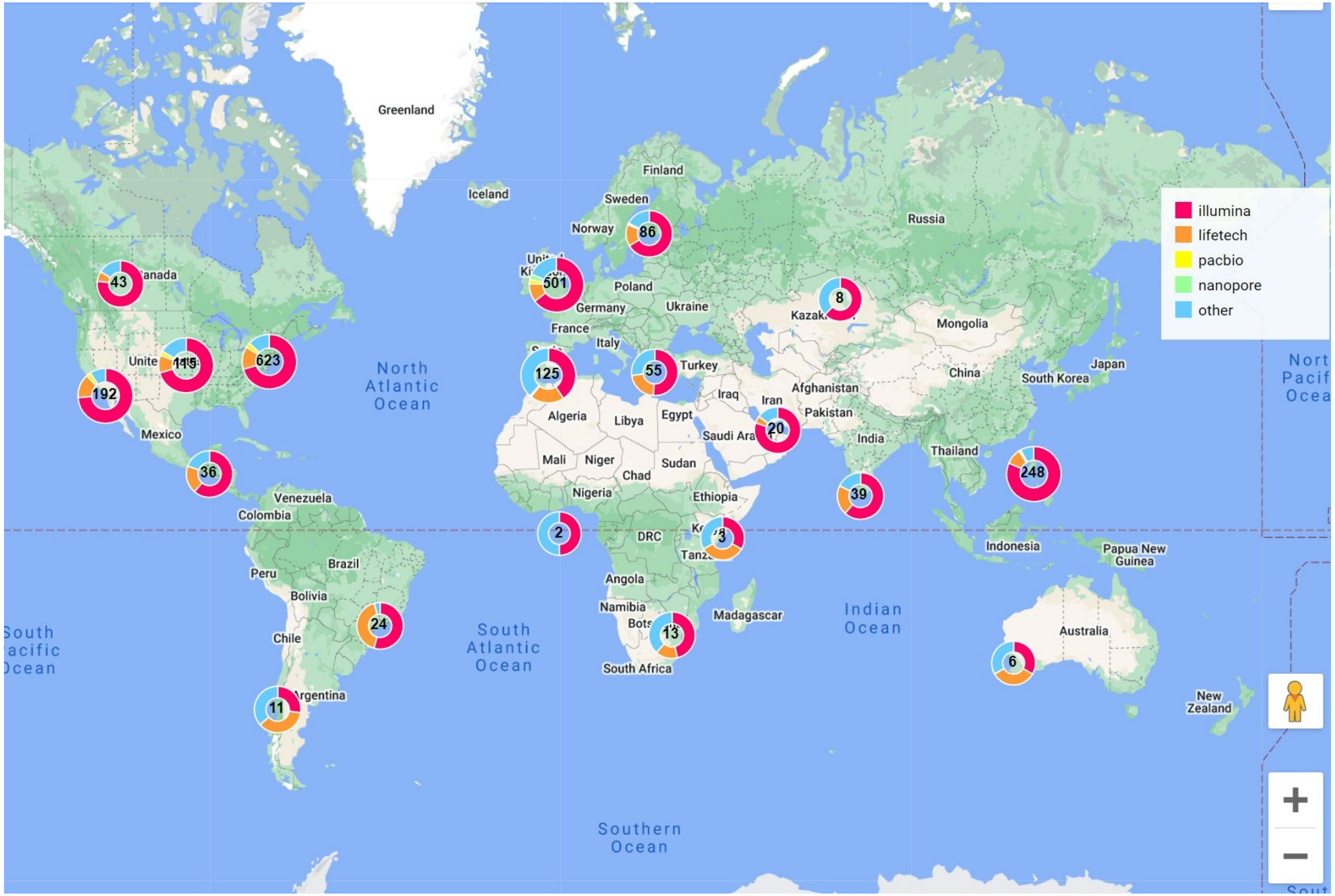
DNA sequencing

Cancer germline mutations
Cancer somatic mutations
Population Risk Score
Pharmacogenomics
Prenatal diagnostics
...

RNA sequencing

Cancer monitoring
Cancer fusion genes
Epidemiological surveillance
...

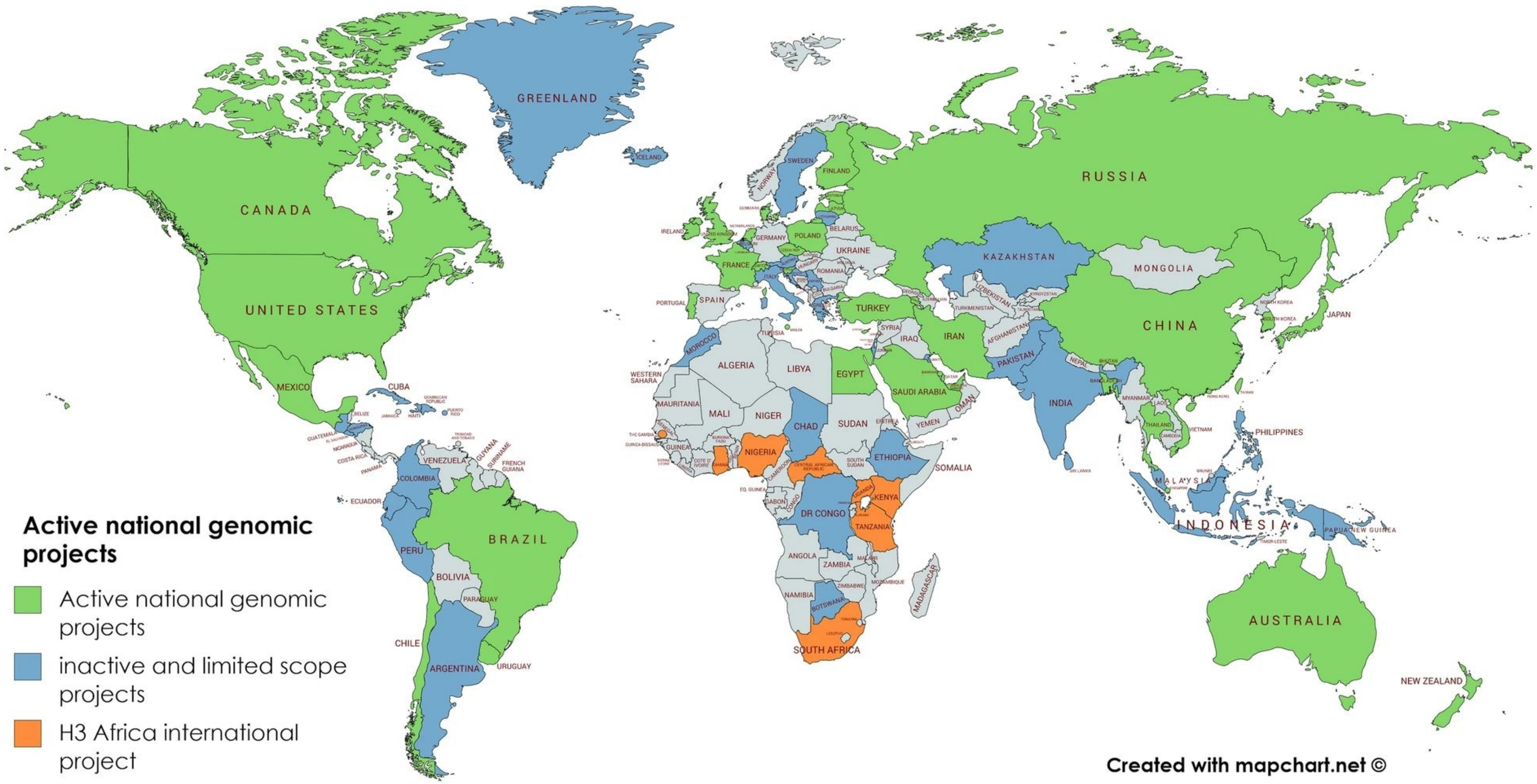
Big data producers



http://enseqlopedia.com/ngs-mapped/

7389 sequencing machines in **1027** centers

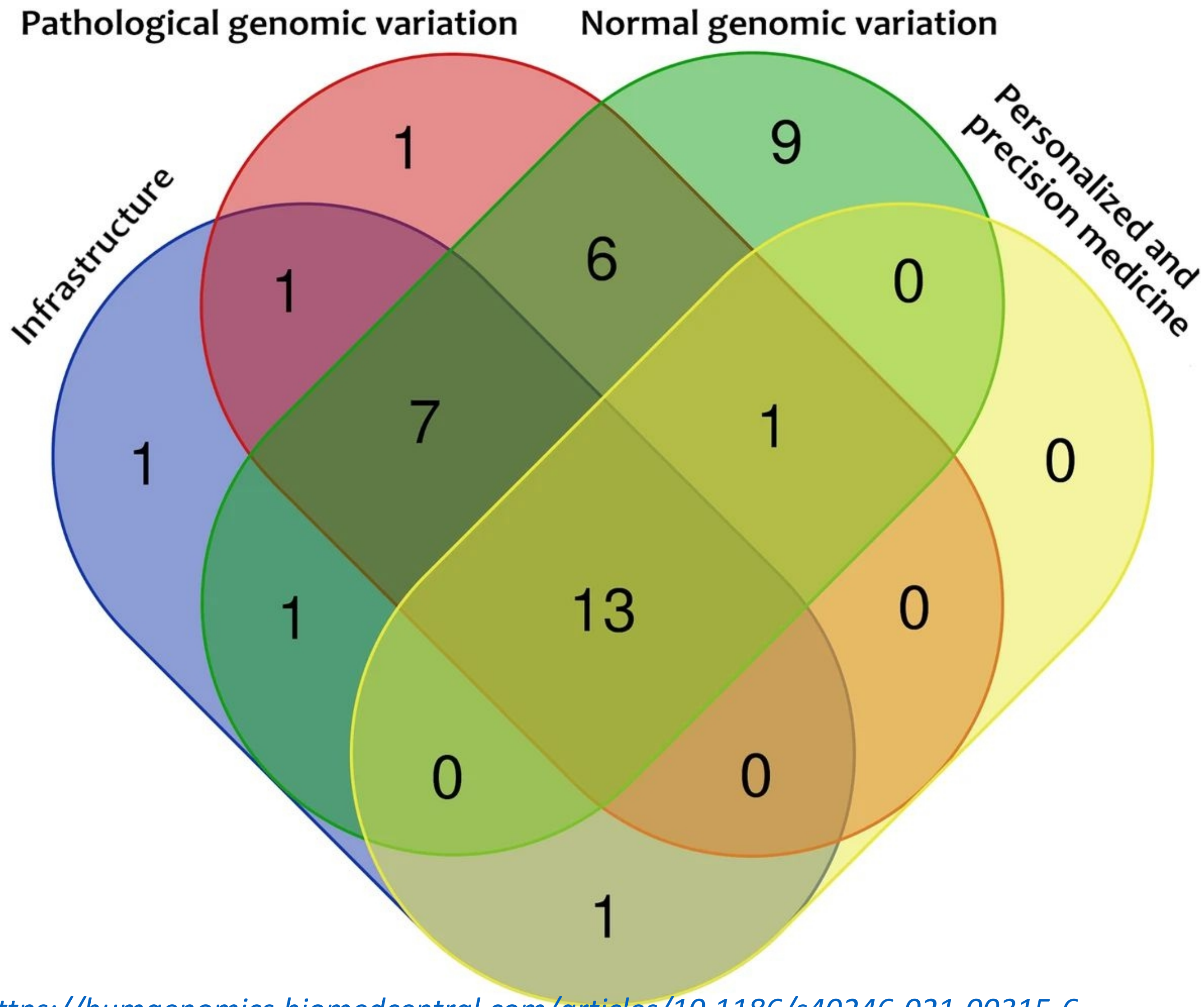
National genomic projects across the world



“Additionally, many country-specific aims were also identified, such as history/ethnic studies (**Armenia**, Brazil, Chile, Hong Kong, Iran, Malta, Mexico, New Zealand, Russia, Singapore, Vietnam) [20,21,22, 25, 31, 34, 41, 42, 45, 56, 61]”.

https://humgenomics.biomedcentral.com/articles/10.1186/s40246-021-00315-6

Types of Genome Projects


Pathological genomic variation
Normal genomic variation
Infrastructure
Personalized and precision medicine
1
9
1
6
0
1
7
1
0
1
13
0
0
0
1

https://humgenomics.biomedcentral.com/articles/10.1186/s40246-021-00315-6

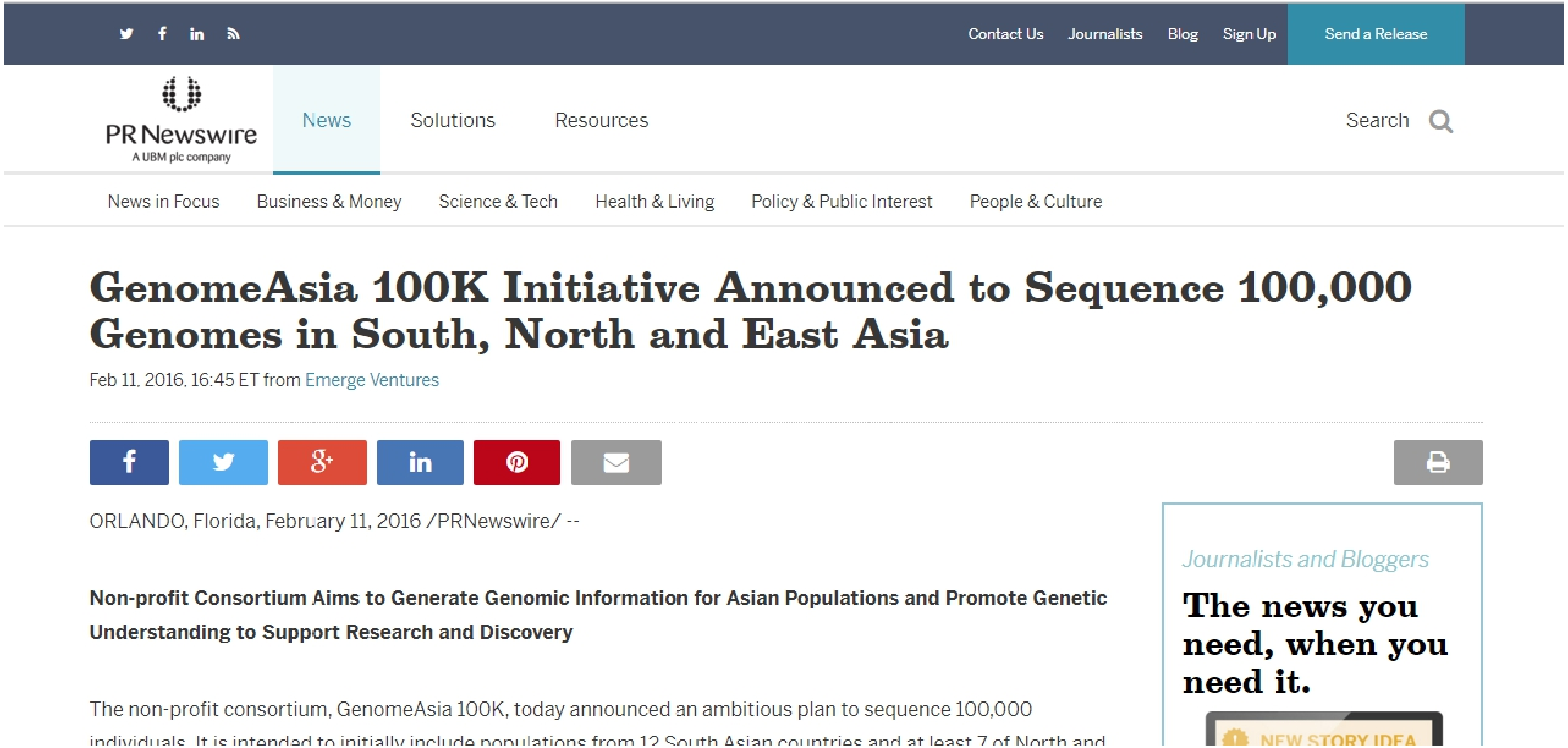
Determining normal genomic variation
cohorts based on demographic data and criteria for identifying healthy individuals

Determining pathological genomic variation
determine pathological genomic variation through the sequencing of clinical cohorts (rare diseases, cancers)

Infrastructure
data generation, data management, establishing standards of analyses, and education

Personalized and precision medicine
tailored diagnosis and treatment according to the information from the patient’s own genome and specific environmental factors

Genome Asia: 100 000!, 2016



GenomeAsia 100K Initiative Announced to Sequence 100,000 Genomes in South, North and East Asia

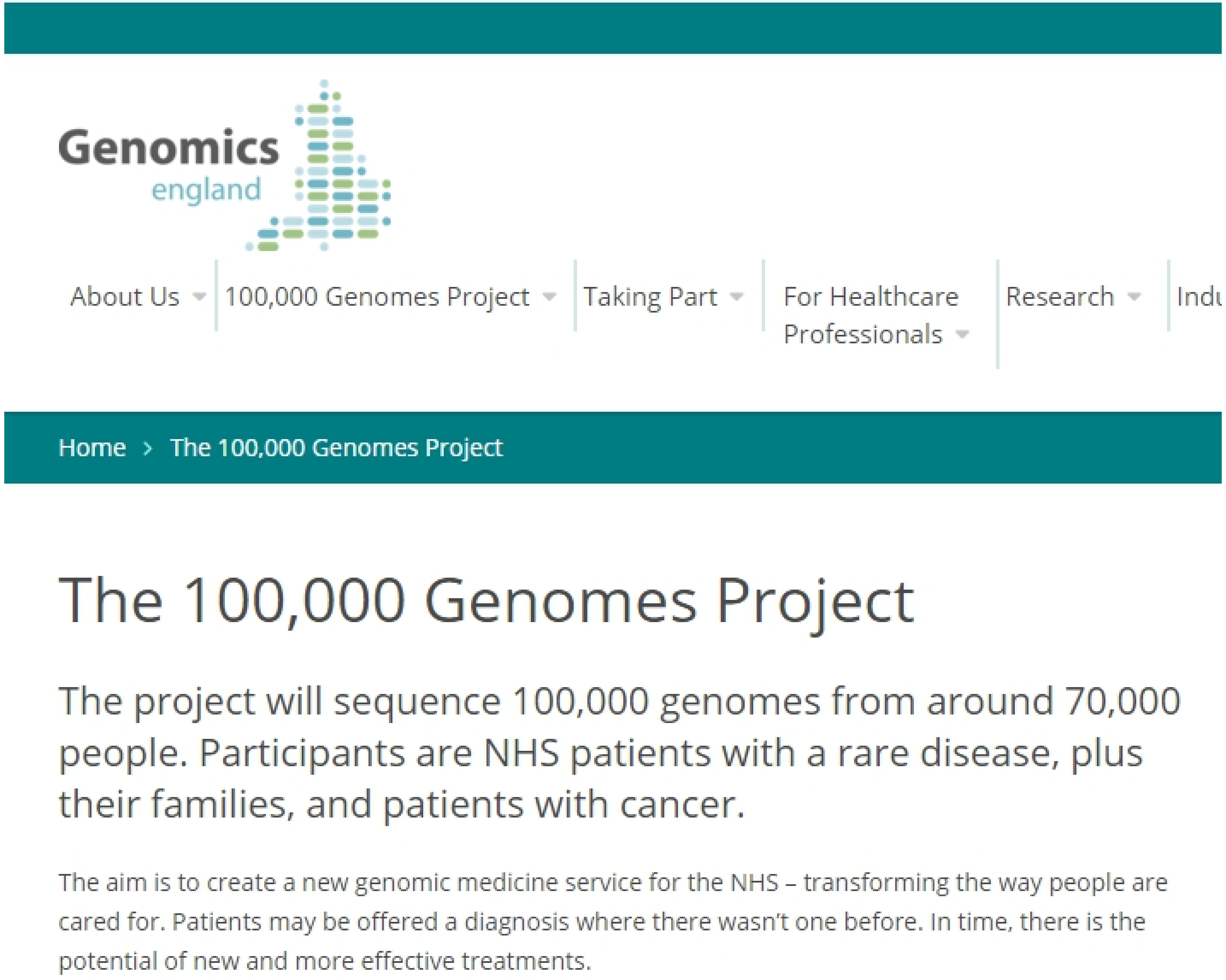
Feb 11, 2016, 16:45 ET from Emerge Ventures

ORLANDO, Florida, February 11, 2016 /PRNewswire/ --

Non-profit Consortium Aims to Generate Genomic Information for Asian Populations and Promote Genetic Understanding to Support Research and Discovery

The non-profit consortium, GenomeAsia 100K, today announced an ambitious plan to sequence 100,000 individuals. It is intended to initially include populations from 12 South Asian countries and at least 7 of North and

The UK 100,0000 Genomes Project, 2015

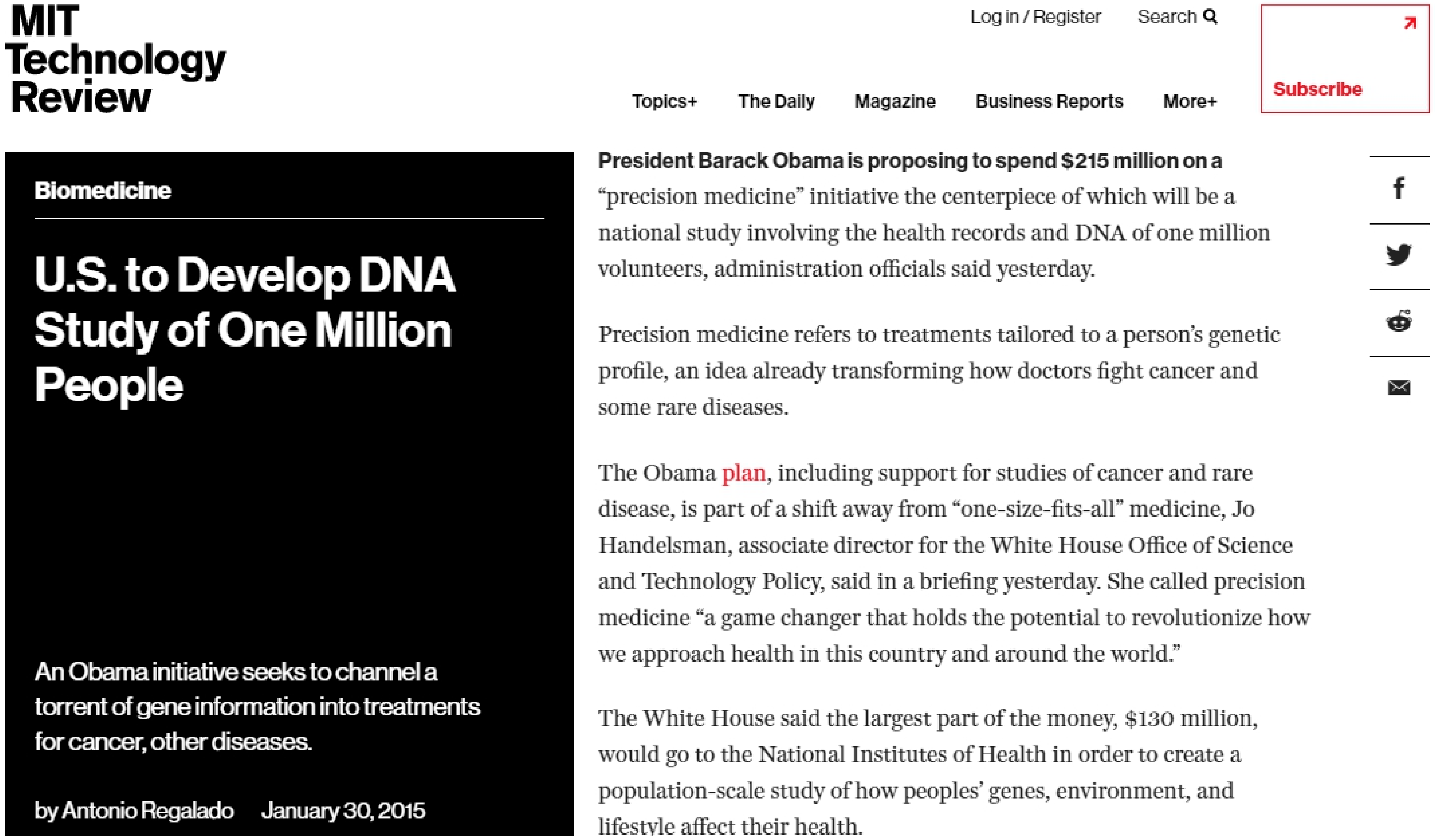


The 100,000 Genomes Project

The project will sequence 100,000 genomes from around 70,000 people. Participants are NHS patients with a rare disease, plus their families, and patients with cancer.

The aim is to create a new genomic medicine service for the NHS – transforming the way people are cared for. Patients may be offered a diagnosis where there wasn't one before. In time, there is the potential of new and more effective treatments.

All of US genome project: 1 mln! start 2015



Biomedicine

U.S. to Develop DNA Study of One Million People

An Obama initiative seeks to channel a torrent of gene information into treatments for cancer, other diseases.

by Antonio Regalado January 30, 2015

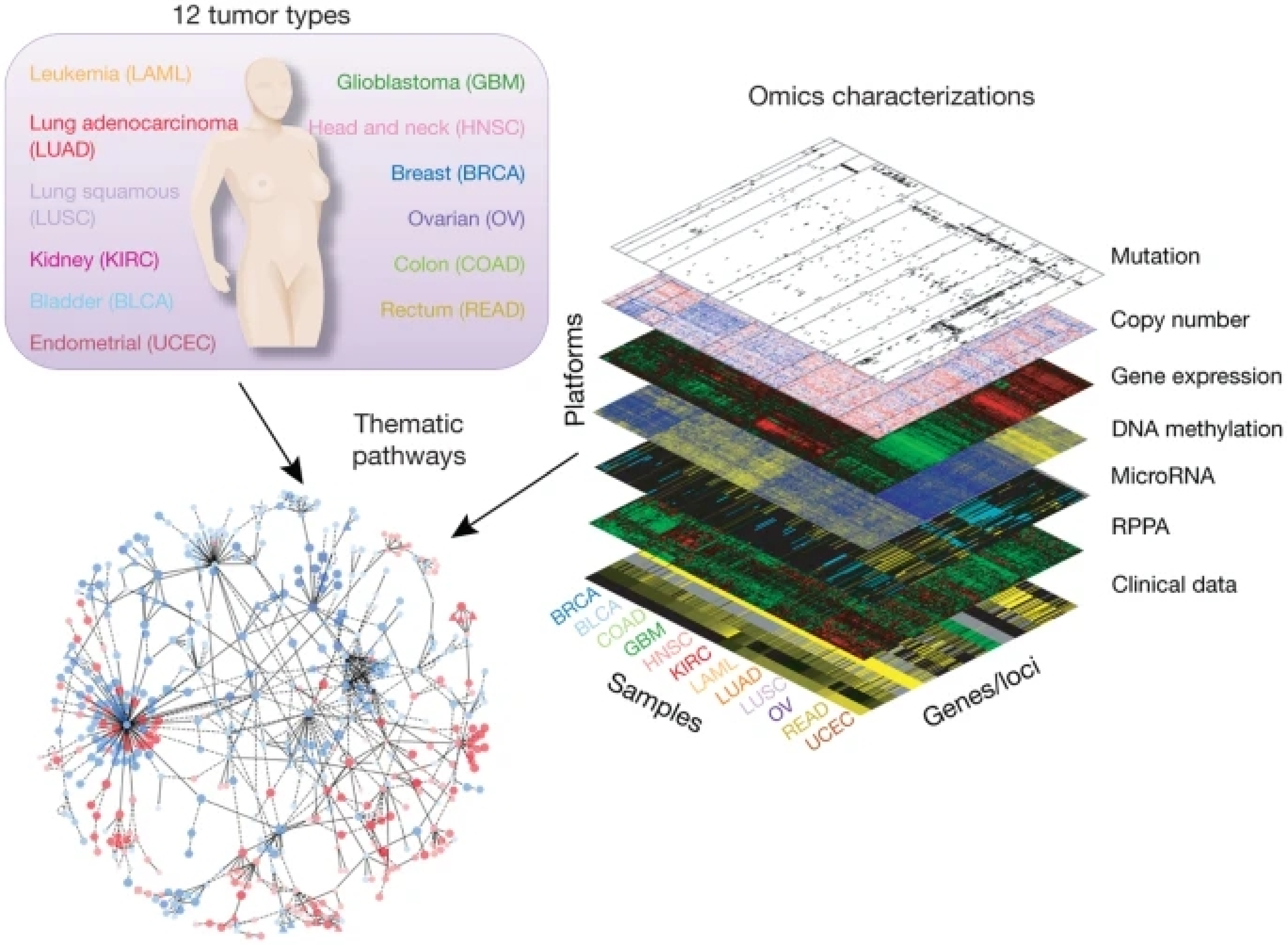
President Barack Obama is proposing to spend $215 million on a “precision medicine” initiative the centerpiece of which will be a national study involving the health records and DNA of one million volunteers, administration officials said yesterday.

Precision medicine refers to treatments tailored to a person’s genetic profile, an idea already transforming how doctors fight cancer and some rare diseases.

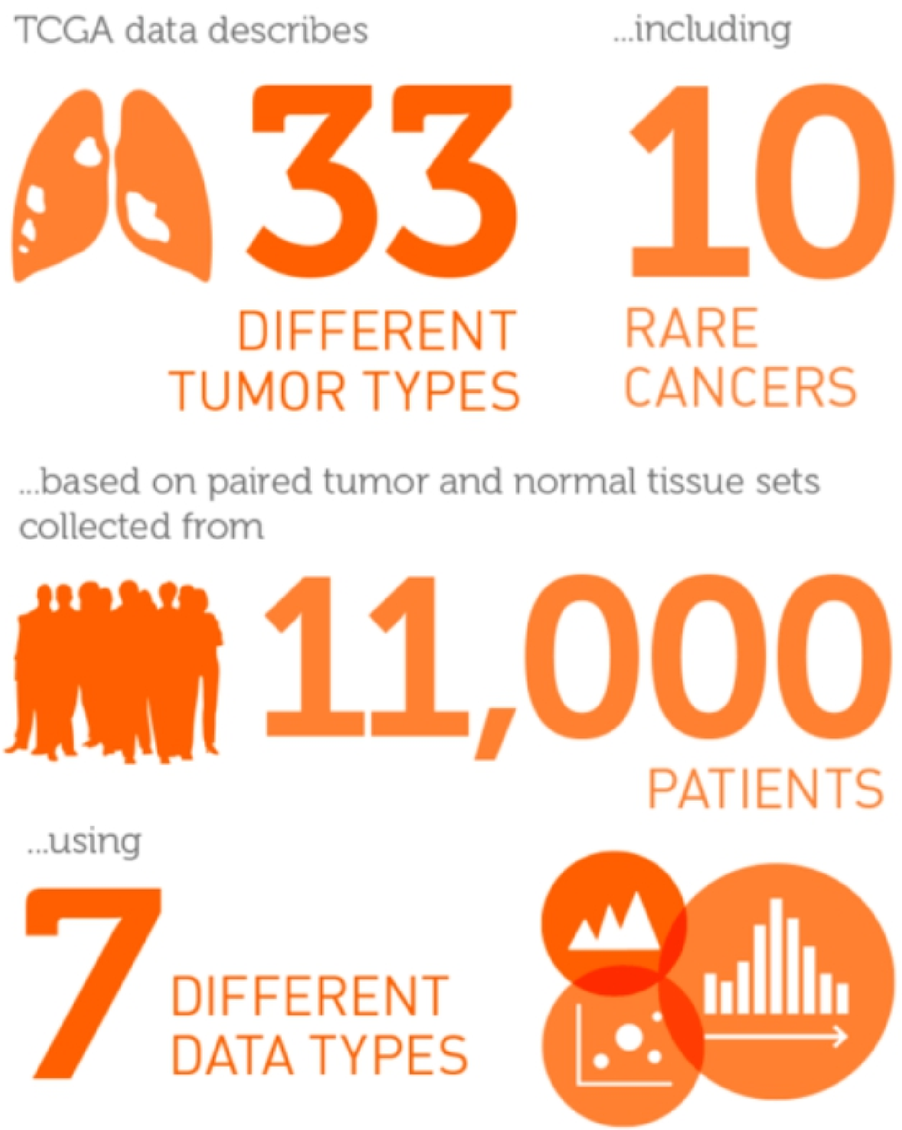
The Obama plan, including support for studies of cancer and rare disease, is part of a shift away from “one-size-fits-all” medicine, Jo Handelsman, associate director for the White House Office of Science and Technology Policy, said in a briefing yesterday. She called precision medicine “a game changer that holds the potential to revolutionize how we approach health in this country and around the world.”

The White House said the largest part of the money, $130 million, would go to the National Institutes of Health in order to create a population-scale study of how peoples’ genes, environment, and lifestyle affect their health.

The Cancer Genome Atlas



Nature Genetics volume 45, pages1113–1120 (2013)



TCGA data describes 33 DIFFERENT TUMOR TYPES ...including 10 RARE CANCERS

...based on paired tumor and normal tissue sets collected from 11,000 PATIENTS

...using 7 DIFFERENT DATA TYPES

https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/tcga-by-the-numbers-infographic

BrainSeq: Neurogenomics to Drive Novel Target Discovery for Neuropsychiatric Disorders

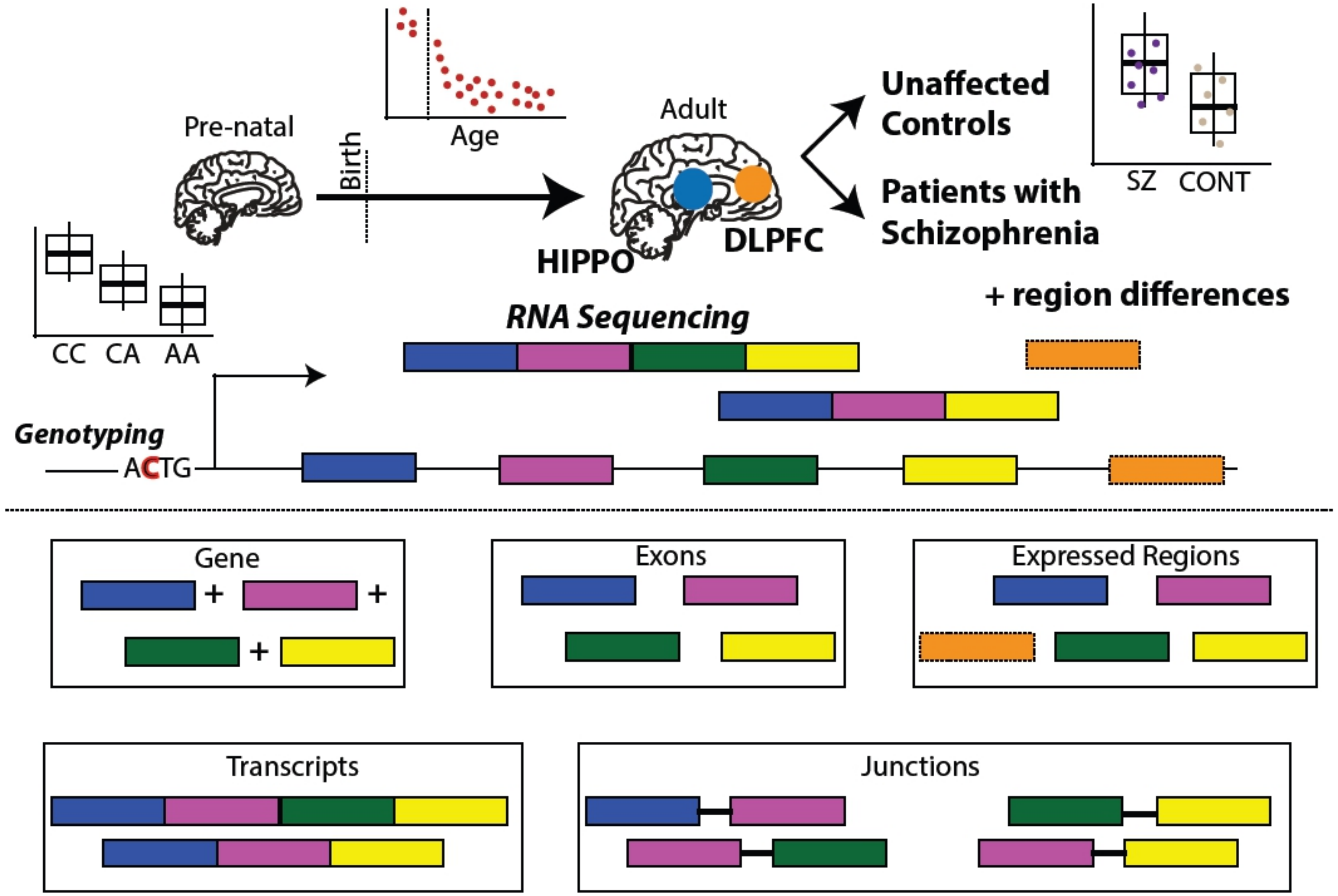
AstraZeneca
Johnson & Johnson
Roche
LIBD LIEBER INSTITUTE for BRAIN DEVELOPMENT MALTZ RESEARCH LABORATORIES
Lilly
Lundbeck
Pfizer
astellas

Schizophrenia cases

	DLPFC	HIPPO	total
adult	152	132	284
prenatal	0	0	0
0 <= age < 18	1	1	2
total	153	133	286

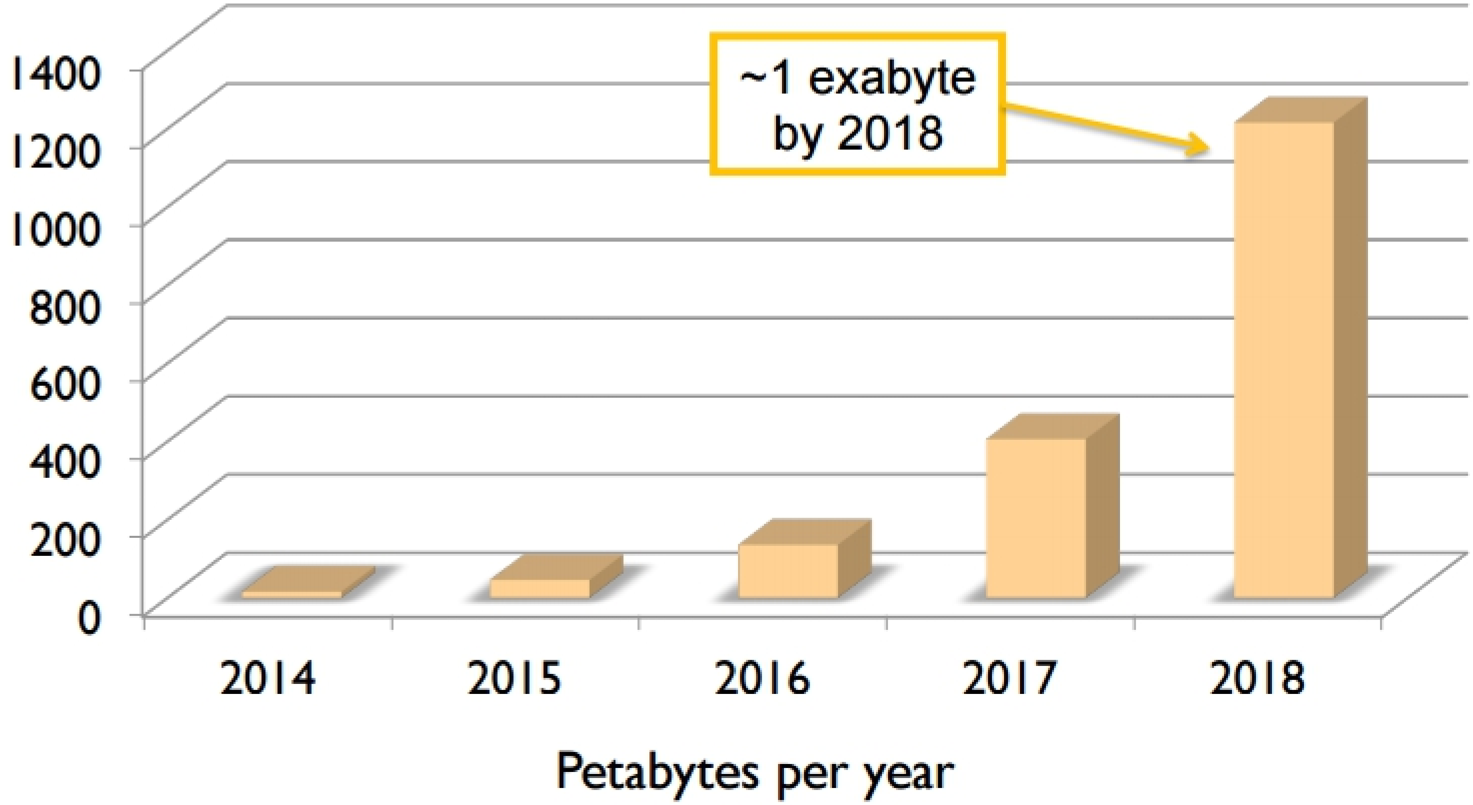
Non-psychiatric controls

	DLPFC	HIPPO	total
adult	222	238	460
prenatal	29	28	57
0 <= age < 18	49	48	97
total	300	314	614



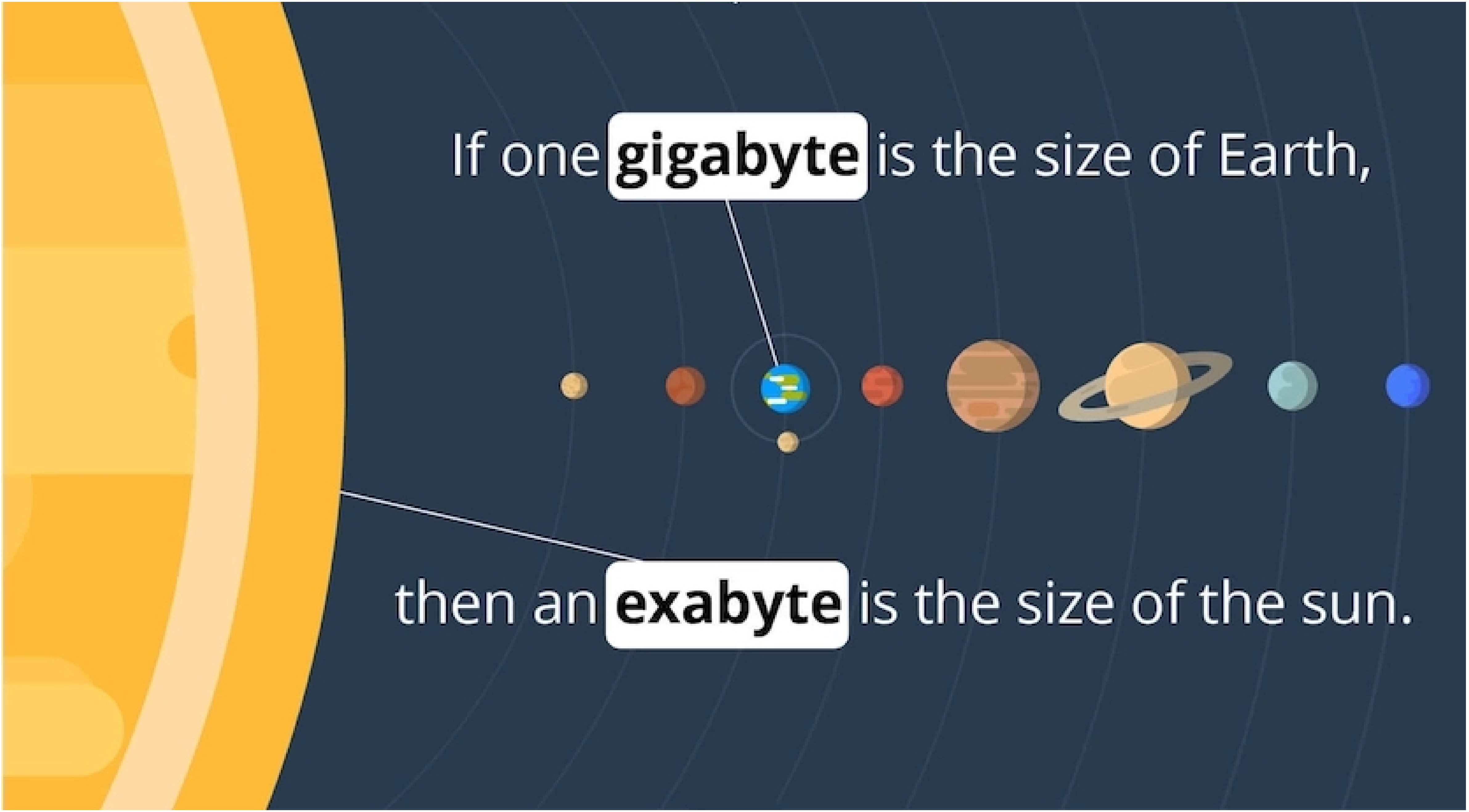
Explosion of biological data

Current world-wide sequencing capacity is growing at ~3x per year!



~1 exabyte by 2018

http://schatzlab.cshl.edu/presentations/2014.03.24.Keystone%20BigData.pdf

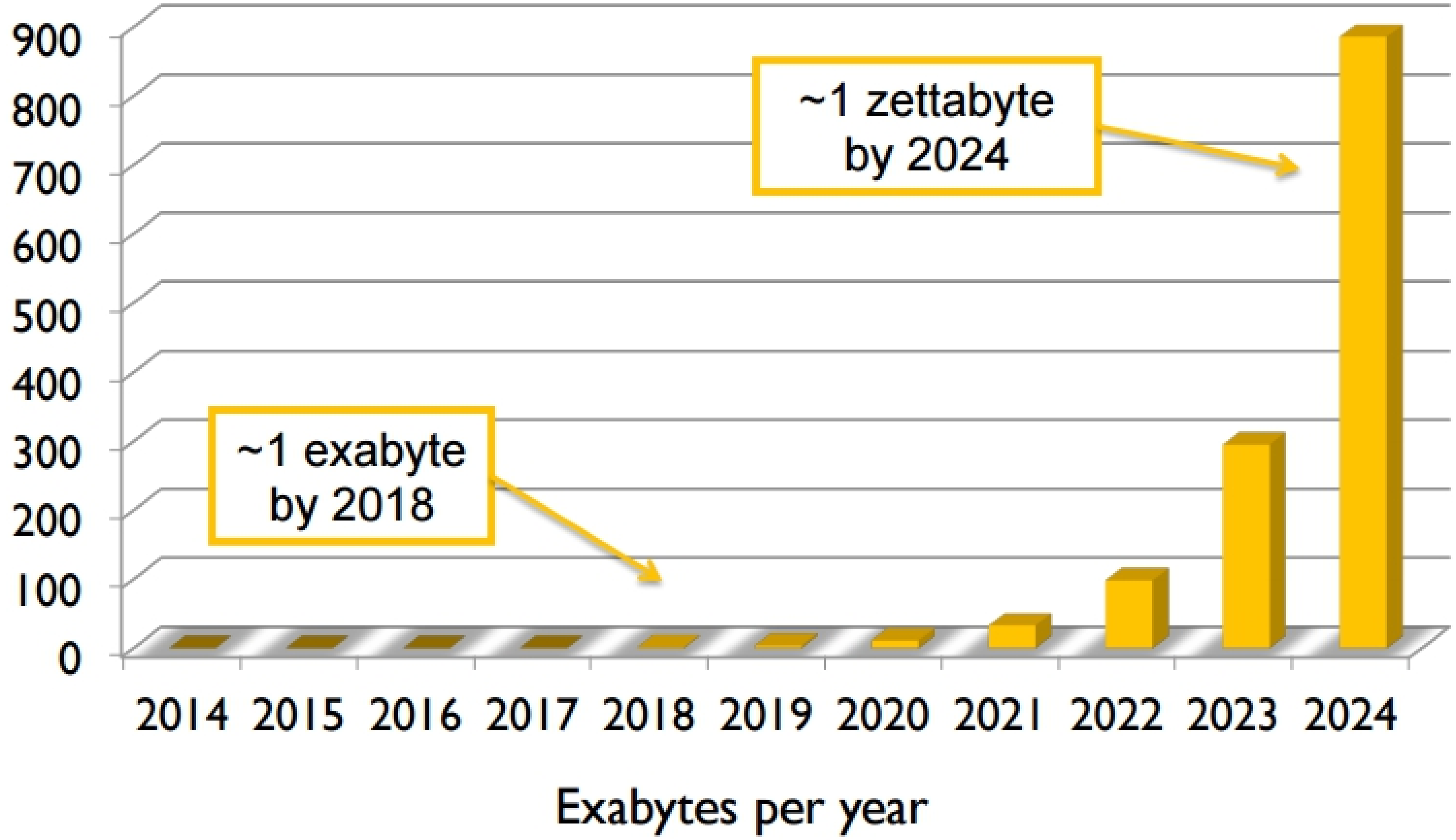


If one **gigabyte** is the size of Earth,

then an **exabyte** is the size of the sun.

Explosion of biological data

Current world-wide sequencing capacity is growing at ~3x per year!

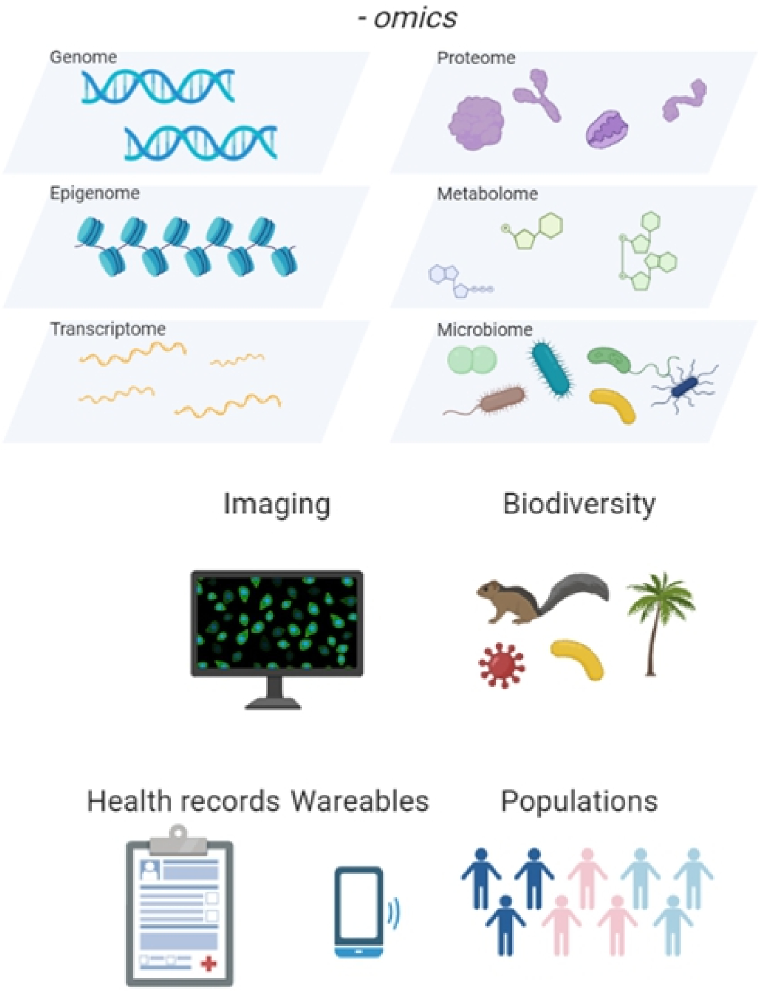


http://schatzlab.cshl.edu/presentations/2014.03.24.Keystone%20BigData.pdf

Bioinformatics as a genomics driver

***Bioinformatics** serves to organize, annotate and analyze the data in the most informative and creative manner to study biology, and synthesize or modify living matter/organisms for a better world.*

THE MULTIVERSE OF DATA TYPES



L Nersisyan, ABI, 2021

Biological questions to answer

- What is the genome sequence?
- How do different genomes vary?
- What variations are inked to diseases?
- How are genes activated and regulated?
- How genomes changed during evolution?
- What causes development of diseases?
- How does an organism respond to different drugs?

And many more …

Why (where) bioinformatics matters

Precision medicine

- Precision drugs
- Diagnostics
- Microbiome

Precise, early, non-invasive diagnostics and personalized treatments

Epidemiology

- Early detection
- Diagnostics
- Vaccine development

Faster, accessible detection of infectious agents
Efficient screening for vaccine candidates

Bioengineering

- Agriculture, wine
- Biomaterials, biomimicry
- Genetic engineering

Engineer better soil, better crops, better biomaterials
Identify events of genetic engineering

Ecosystem management

- Biodiversity
- Gene drive

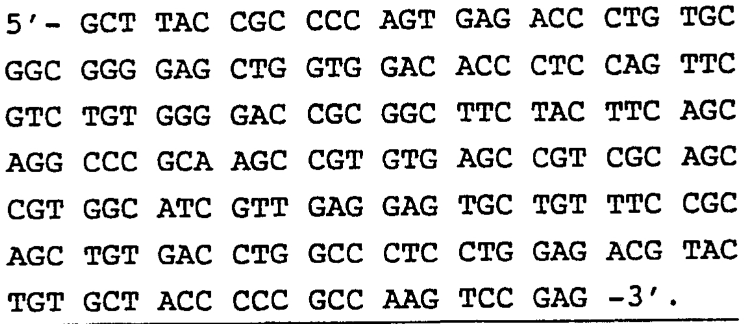
Engineer better soil, better crops, better biomaterials
Find traces of genetic modifications

Big issues of bioinformatics

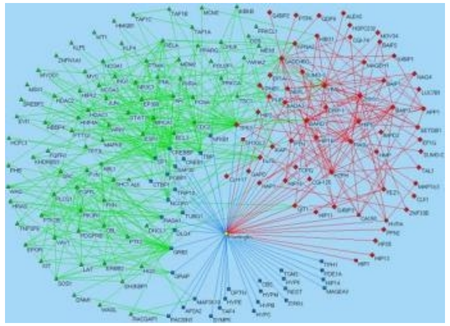
Storage	Exascale biology is certain, zettascale on the horizon More aggressive compression algorithms needed Streaming
Computing systems	Parallel computing GPU-computing FPGA-computing Cloud computing
Data analysis	Batch effects HDLSS algorithms Genomic privacy Visualization
Visualization	Complex data requires simple visualization VR technologies to facilitate visualization

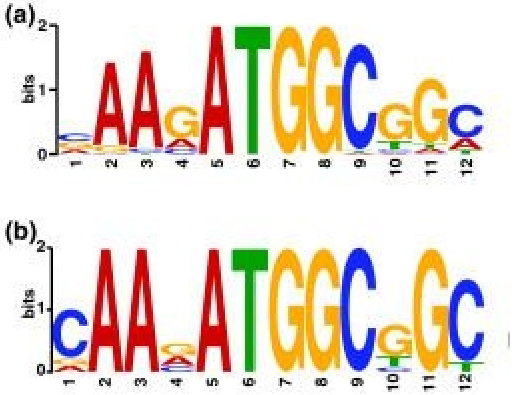
Biological data types

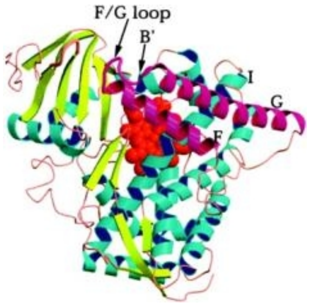
- *Sequences*
- *Graphs*
- *High-dimensional data*
- *Geometric information*
- *Patterns*
- *Constraints*
- *Images*
- *Spatial information*
- *Models*
- *Literature*

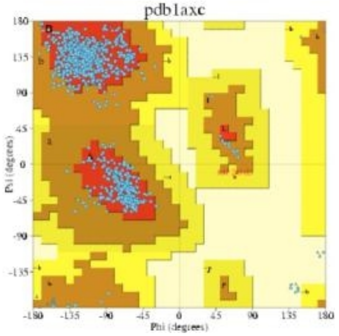


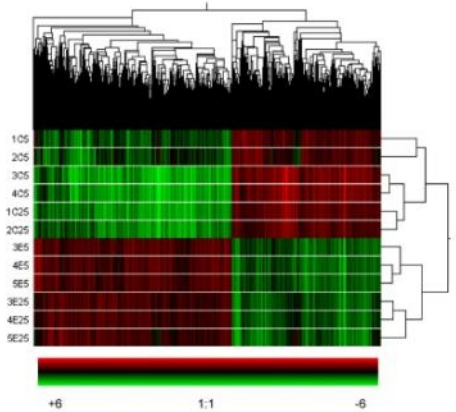
5′- GCT TAC CGC CCC AGT GAG ACC CTG TGC
GGC GGG GAG CTG GTG GAC ACC CTC CAG TTC
GTC TGT GGG GAC CGC GGC TTC TAC TTC AGC
AGG CCC GCA AGC CGT GTG AGC CGT CGC AGC
CGT GGC ATC GTT GAG GAG TGC TGT TTC CGC
AGC TGT GAC CTG GCC CTC CTG GAG ACG TAC
TGT GCT ACC CCC GCC AAG TCC GAG -3′.











Biological big data issues

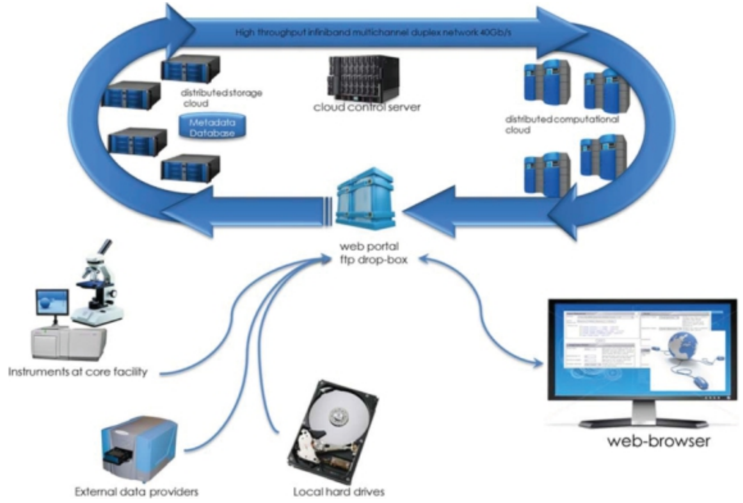
- Storage
- Computing systems
- Algorithms
- Data analysis
- Knowledge generation

UniProt

KEGG Kyoto Encyclopedia of Genes and Genomes

GEO Gene Expression Omnibus

ENA European Nucleotide Archive


High throughput infiniband multichannel duplex network 40Gb/s
distributed storage cloud
Metadata Database
cloud control server
distributed computational cloud
web portal ftp drop-box
Instruments at core facility
External data providers
Local hard drives
web-browser

Galaxy

Storage

- European Molecular Biology Laboratory – European Institute of Bioinformatics ~ 20 Pb

- National Centre for Biotechnology Information (US) ~ 25 Pb

- NCBI GEO – 2,081,388 samples

- NCBI SRA – 2,340,690 samples

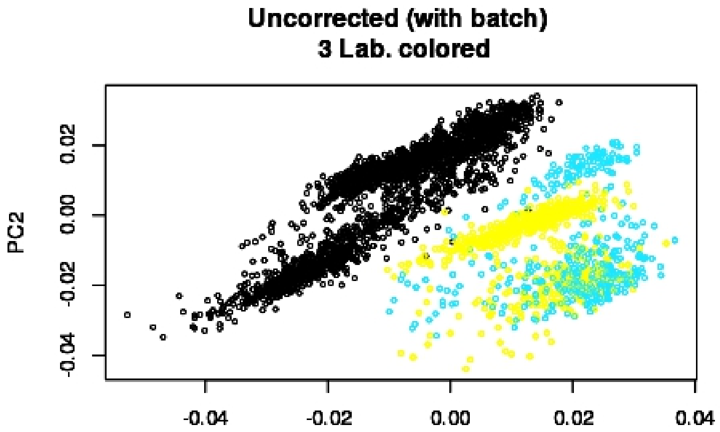
Data analysis issues

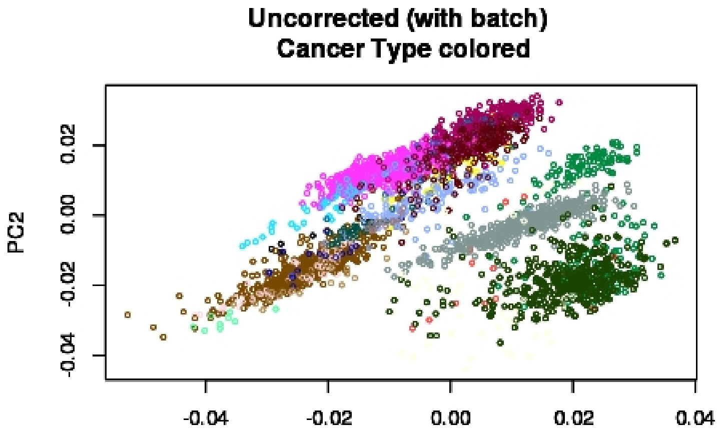
- Batch effects
- HDLSS algorithms
- Genomic privacy
- Visualization

Batch effects

- Data produced in different labs are different

- Standardization is practically impossible
  - ~2500 different microarray platforms
  - > 10 sequencing platforms
  - recommended RNA amounts 2.5-20 pg
  - in single cell seq every cell is a batch


Uncorrected (with batch)
3 Lab. colored


Uncorrected (with batch)
Cancer Type colored

https://www.biostars.org/p/133985/

HDLSS data analysis

- HDLSS: **n** (samples) << **K** (features)

- Biological data is HDLSS

- Gene expression analysis
  - few hundred samples and ~70000 genes

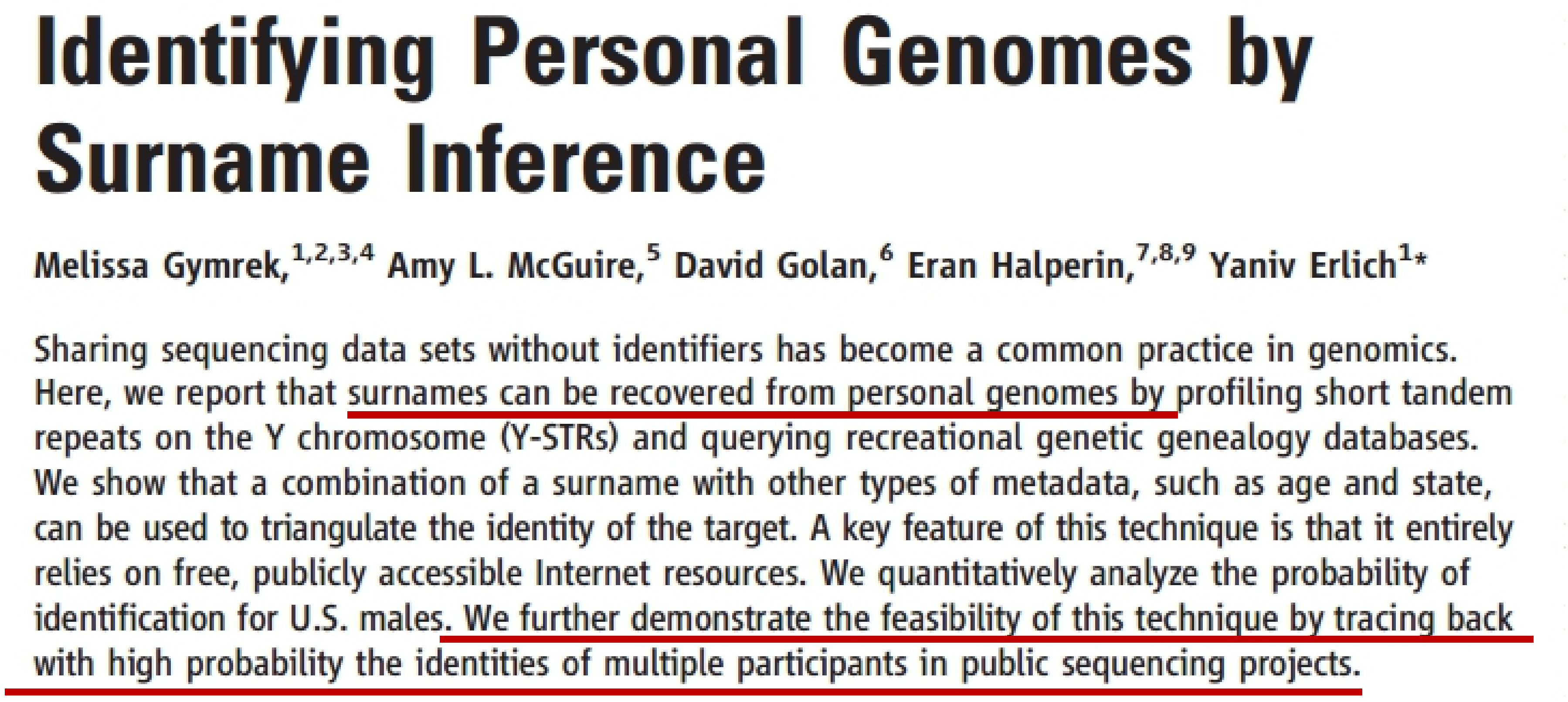
- SNP analysis
  - few thousand samples and ~ 4-10M SNPs

HDLSS data analysis

- Dimensionality reduction
  - PCA, MDS, SOM
- Multiple independent statistical tests (not a good idea)
  - t-tests, ANOVA, ….
- Machine learning
  - distance weighted discrimination, neural nets, SVM, association rule mining

Privacy-preserving computation

Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y.
Identifying personal genomes by surname inference.
Science. 2013 18;339(6117):321-4.

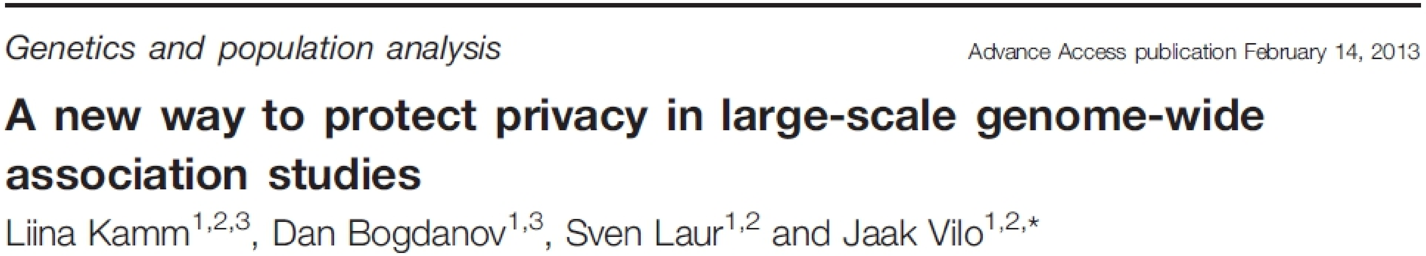


Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

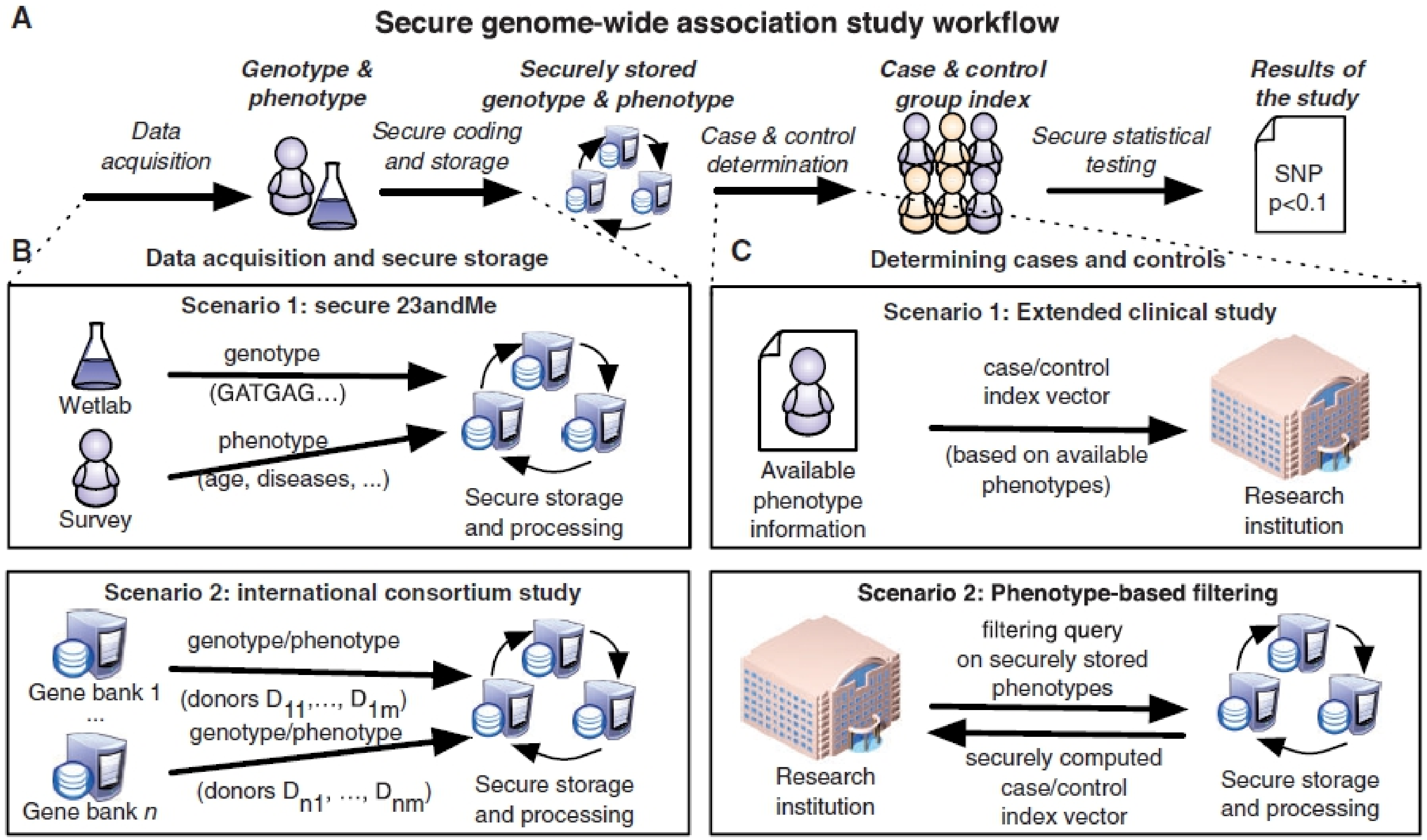
Privacy-preserving computation



Genetics and population analysis Advance Access publication February 14, 2013

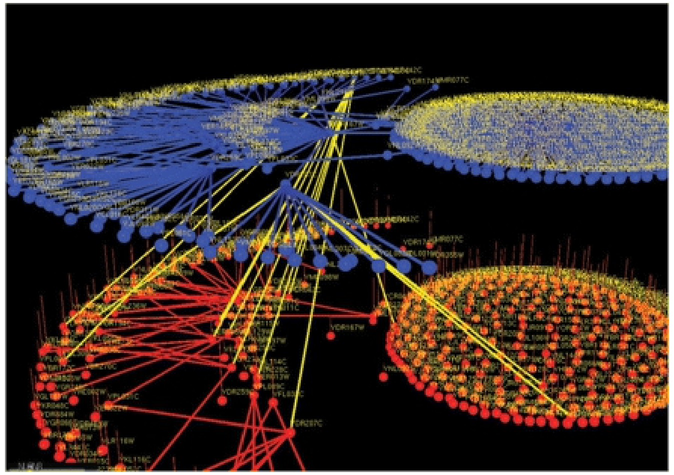
A new way to protect privacy in large-scale genome-wide association studies

Liina Kamm[1,2,3], Dan Bogdanov[1,3], Sven Laur[1,2] and Jaak Vilo[1,2,*]

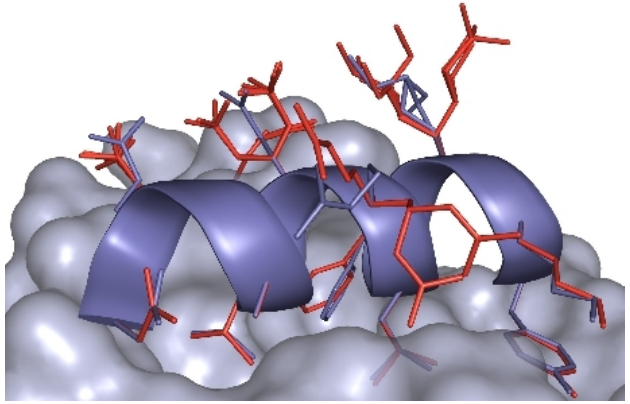

A
Secure genome-wide association study workflow
Data acquisition
Genotype & phenotype
Secure coding and storage
Securely stored genotype & phenotype
Case & control determination
Case & control group index
Secure statistical testing
Results of the study
SNP p<0.1
B
Data acquisition and secure storage
C
Determining cases and controls
Scenario 1: secure 23andMe
Wetlab
genotype (GATGAG...)
Survey
phenotype (age, diseases, ...)
Secure storage and processing
Scenario 2: international consortium study
Gene bank 1
...
Gene bank n
genotype/phenotype (donors $D_{11}, \ldots, D_{1m}$)
genotype/phenotype (donors $D_{n1}, \ldots, D_{nm}$)
Secure storage and processing
Scenario 1: Extended clinical study
Available phenotype information
case/control index vector
(based on available phenotypes)
Research institution
Scenario 2: Phenotype-based filtering
Research institution
filtering query on securely stored phenotypes
securely computed case/control index vector
Secure storage and processing

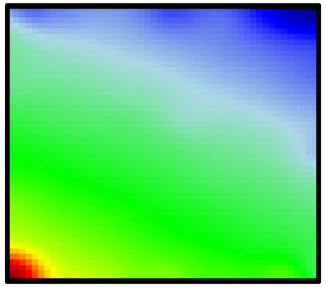
Visualization

Network visualization

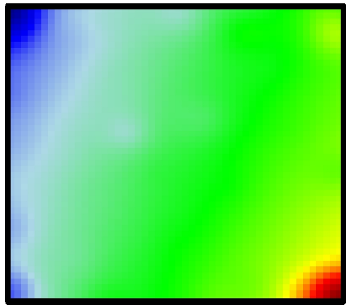


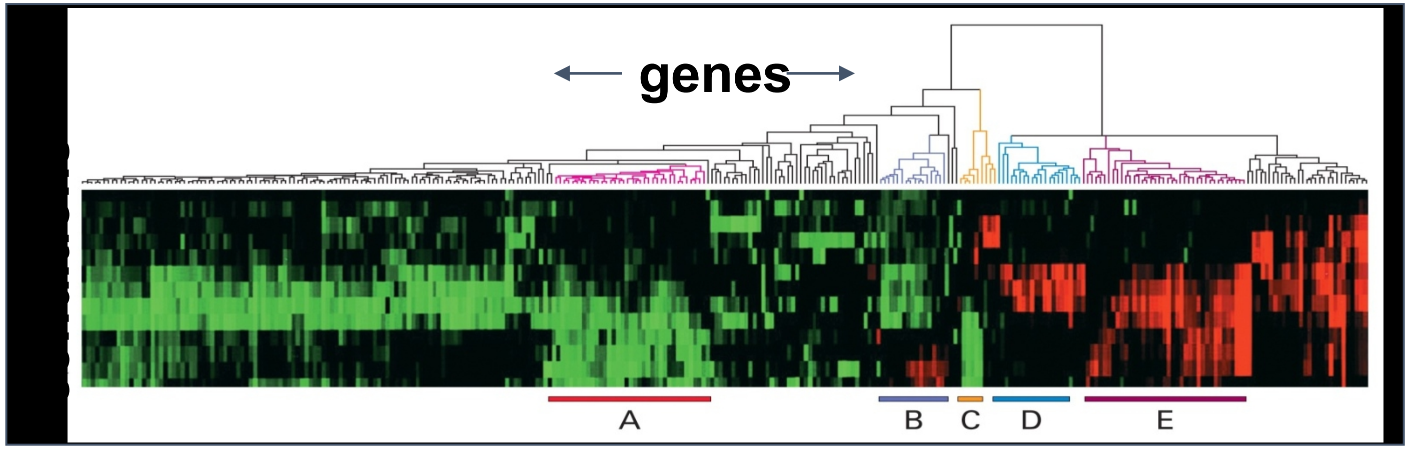
3D structures





Samples
Tuberculosis vs
lung cancer

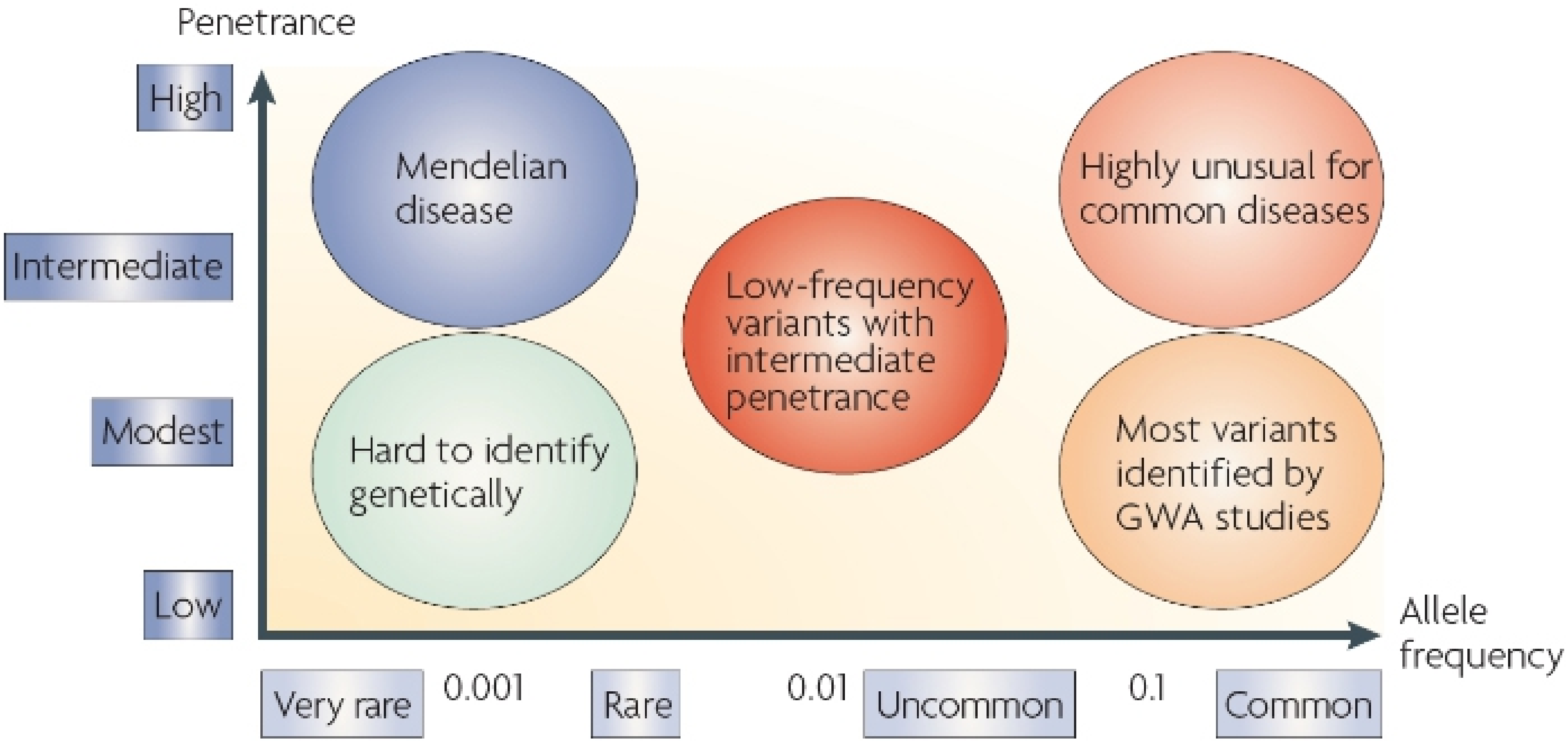




Clustering

Combining EHR, Genomics and Bioinformatics

Mutation data is the most frequent source of data for public health/precision medicine

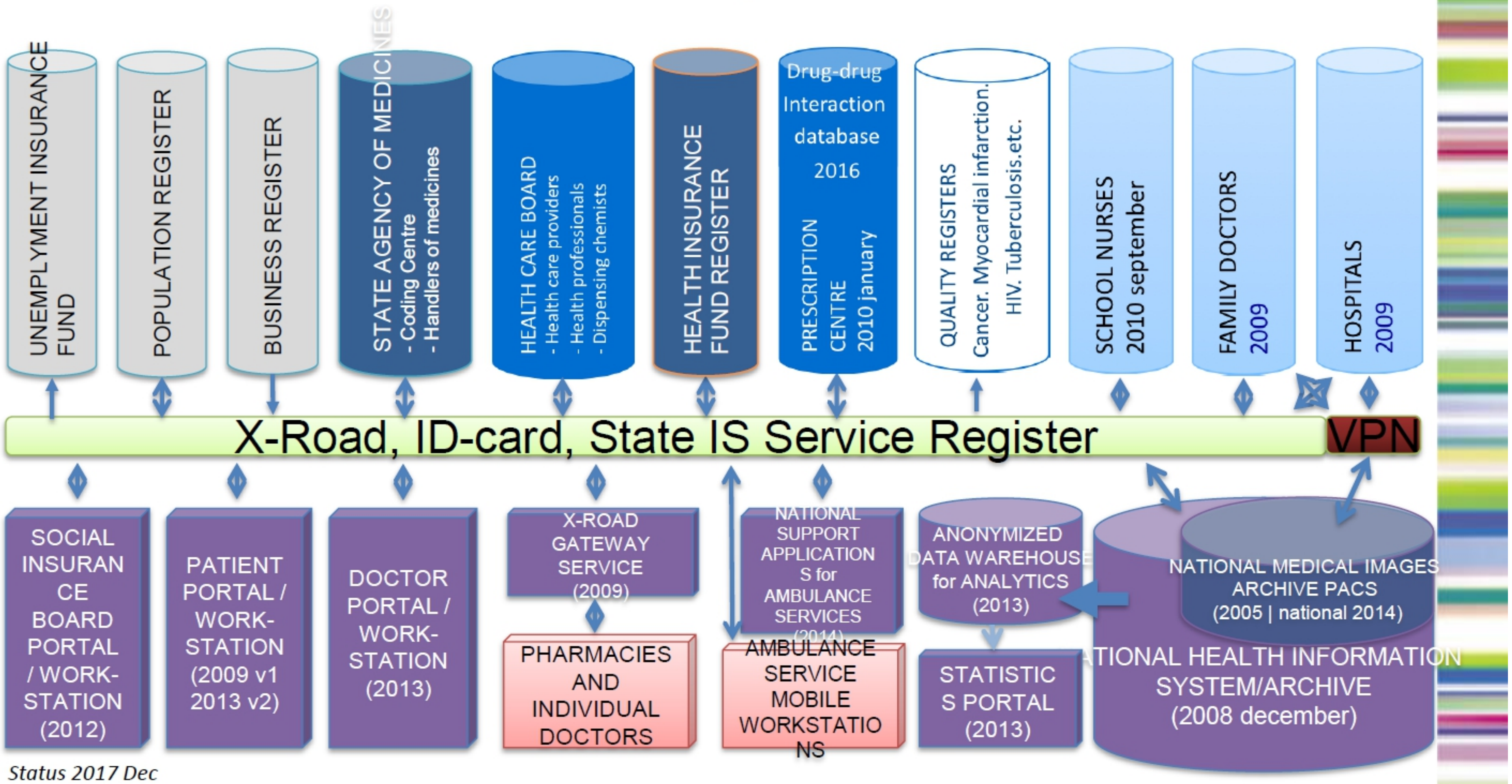

Penetrance
High
Intermediate
Modest
Low
Mendelian disease
Hard to identify genetically
Low-frequency variants with intermediate penetrance
Highly unusual for common diseases
Most variants identified by GWA studies
Allele frequency
Very rare
0.001
Rare
0.01
Uncommon
0.1
Common

Nat Rev Genet. 2008;9: 356–369. doi:10.1038/nrg2344

Estonian Genome Project

From biobanking to personalized medicine

Secure exchange of health data – cornerstone of Estonian digital health architecture



UNEMPLOYMENT INSURANCE FUND

POPULATION REGISTER

BUSINESS REGISTER

STATE AGENCY OF MEDICINES
- Coding Centre
- Handlers of medicines

HEALTH CARE BOARD
- Health care providers
- Health professionals
- Dispensing chemists

HEALTH INSURANCE FUND REGISTER

Drug-drug Interaction database 2016

PRESCRIPTION CENTRE 2010 january

QUALITY REGISTERS
Cancer. Myocardial infarction.
HIV. Tuberculosis.etc.

SCHOOL NURSES 2010 september

FAMILY DOCTORS 2009

HOSPITALS 2009

X-Road, ID-card, State IS Service Register

VPN

SOCIAL INSURANCE BOARD PORTAL / WORK-STATION (2012)

PATIENT PORTAL / WORK-STATION (2009 v1 2013 v2)

DOCTOR PORTAL / WORK-STATION (2013)

X-ROAD GATEWAY SERVICE (2009)

PHARMACIES AND INDIVIDUAL DOCTORS

NATIONAL SUPPORT APPLICATIONS for AMBULANCE SERVICES (2014)

AMBULANCE SERVICE MOBILE WORKSTATIONS

ANONYMIZED DATA WAREHOUSE for ANALYTICS (2013)

STATISTICS PORTAL (2013)

NATIONAL MEDICAL IMAGES ARCHIVE PACS (2005 | national 2014)

NATIONAL HEALTH INFORMATION SYSTEM/ARCHIVE (2008 december)

Status 2017 Dec

estonian genome center
university of tartu

Estonian biobank: omics profiling

Method	Sample size
Whole genome sequencing (30X)	3,000
Whole exome sequencing	2,500
Genome-wide genotyping arrays	130,000
Genome-wide methylation arrays	700
Genome-wide expression arrays	1,100
mRNA sequencing	600
Total RNA sequencing	50
Metabolomics (NMR)	11 000
Metabolomics (MS/MS)	1,100
Telomere length	5,200
Clinical biochemistry	2,700
IgG glycosylation	1,000

estonian genome center
university of tartu

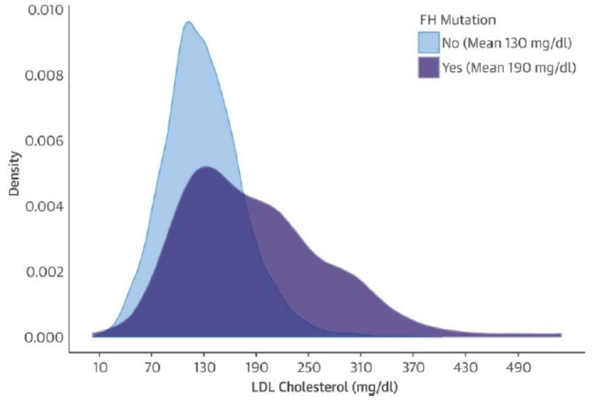
Return the genetic data back to people from the Estonian Biobank

- Estonian biobank is returning the **research** data back to the people who want and agree to get it.
- We are inviting back approx. 3000 people, around 2000 have received by now the polygenetic risk scores (PRS) and 30 min counseling.

estonian genome center
university of tartu

Familiar hypercholesterolemia - FH

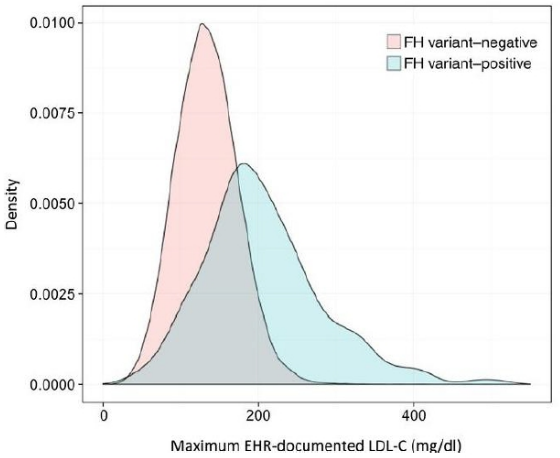
FH-linked variant (*LDLR*, *APOB*, *PCSK9* gene) carriers display **50 mg/dl** (1.3 mmol/L) and **greater** and a **wide spectrum** of LDL-C level



Diagnostic LDL-C level cut-off for FH cases >4.9 mmol/L

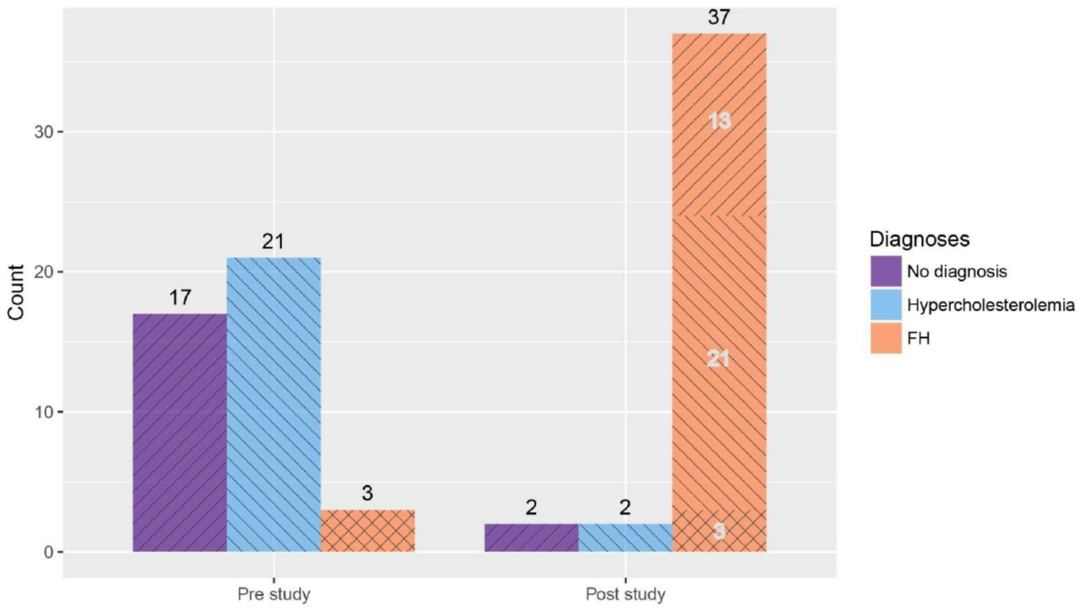
Khera et al. J Am Coll Cardiol. 2016

Abul-Husn et al. Science 2016



estonian genome center
university of tartu

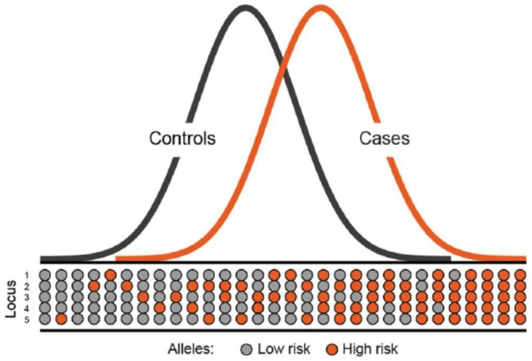
Alver et al. (2018) **Genetics in Medicine**



estonian genome center
university of tartu

Polygenic risk scores

- Most of the associated loci identified in GWAS have very small effects
- Polygenic risk score can be constructed by combining the effects of all associated loci
  - unweighted: sum of all risk alleles
  - weighted: sum of all risk alleles weighted by their effect size



estonian genome center
university of tartu

- **PRS – this is what we are born with!**
- **Biomarkers** (elevated LDL-C, systolic blood pressure, glycose tolerance test etc.) will change when disease process is already ongoing

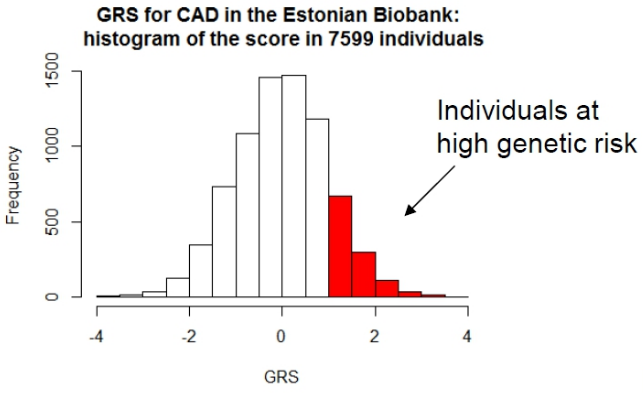
estonian genome center
university of tartu

Polygenic risk scores (PRS) weighted: sum of all risk alleles weighted by their effect size

Calculated as $S = w_1X_1 + w_2X_2 + \ldots + w_kX_k$,

$X_1, \ldots, X_k$ - allele dosages for k independent markers (SNP-s),

$w_1, w_2, \ldots, w_k$ – weights



Methodological questions:

A) How to select the SNPs – how many and what are the selection criteria?

B) How to select the optimal weights?

K. Läll …. & K. Fischer, GM, 2016

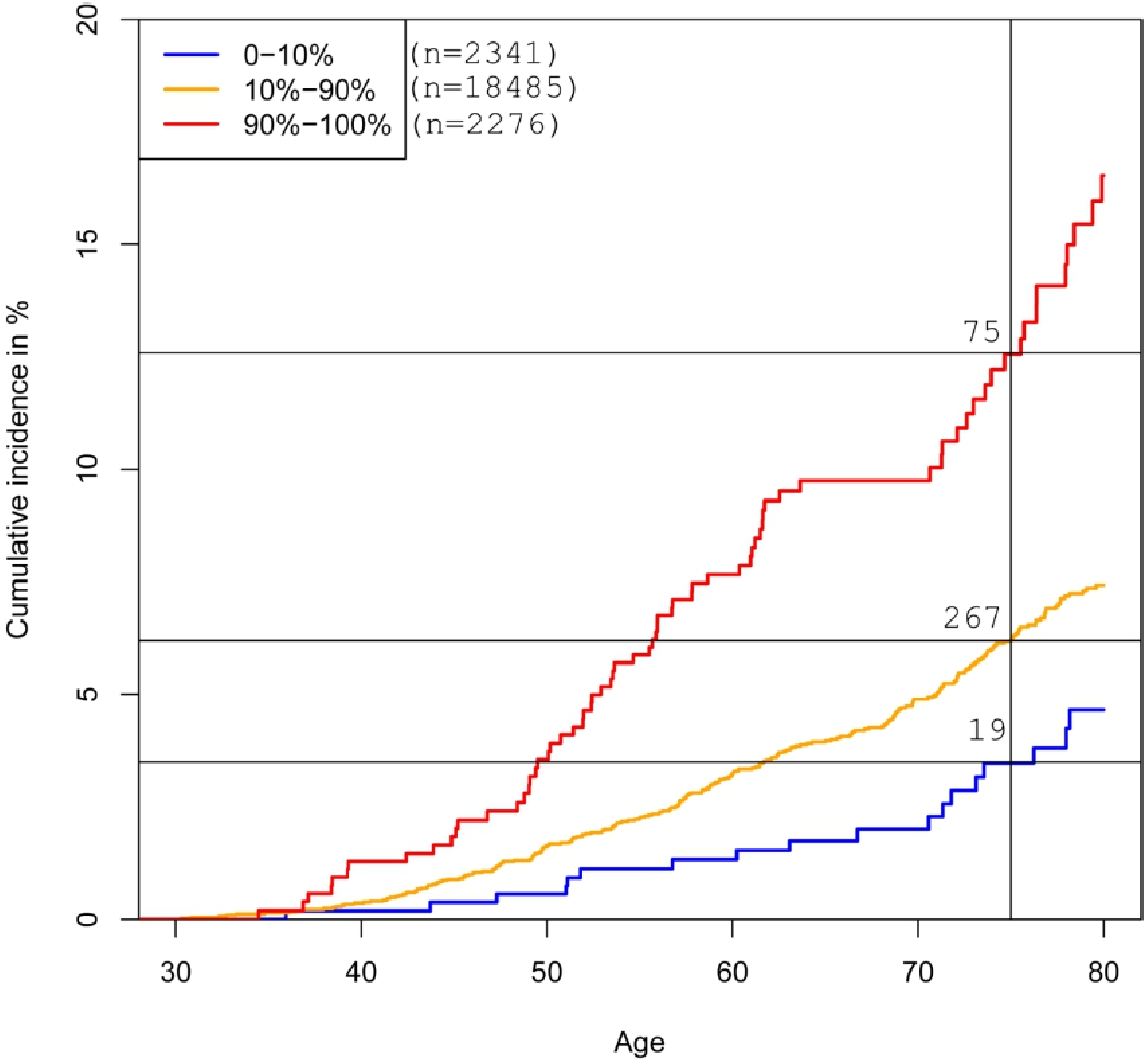
estonian genome center
university of tartu

PRS of Breast Cancer

- No BRACA1 & BRCA2, but ca 900 SNP variants

Läll et al (2019) *BMC Cancer 19, 557*

estonian genome center
university of tartu



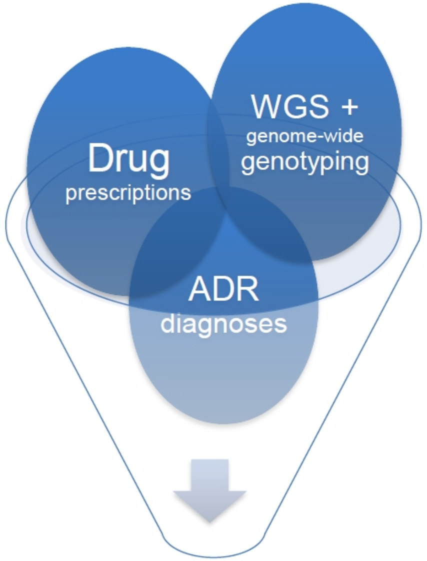
Cumulative incidence **by the age of 75** in GRS top 10% category was **12.6%**. In middle category **6.2%** and in the lowest 10% GRS category, **3.5%.**

Median follow up 8.6 years, total number of cases 361.

estonian genome center
university of tartu

Importance of pharmacogenomics

98% of Europeans carry ≥ 1 mutation of pharmacogenetic relevance



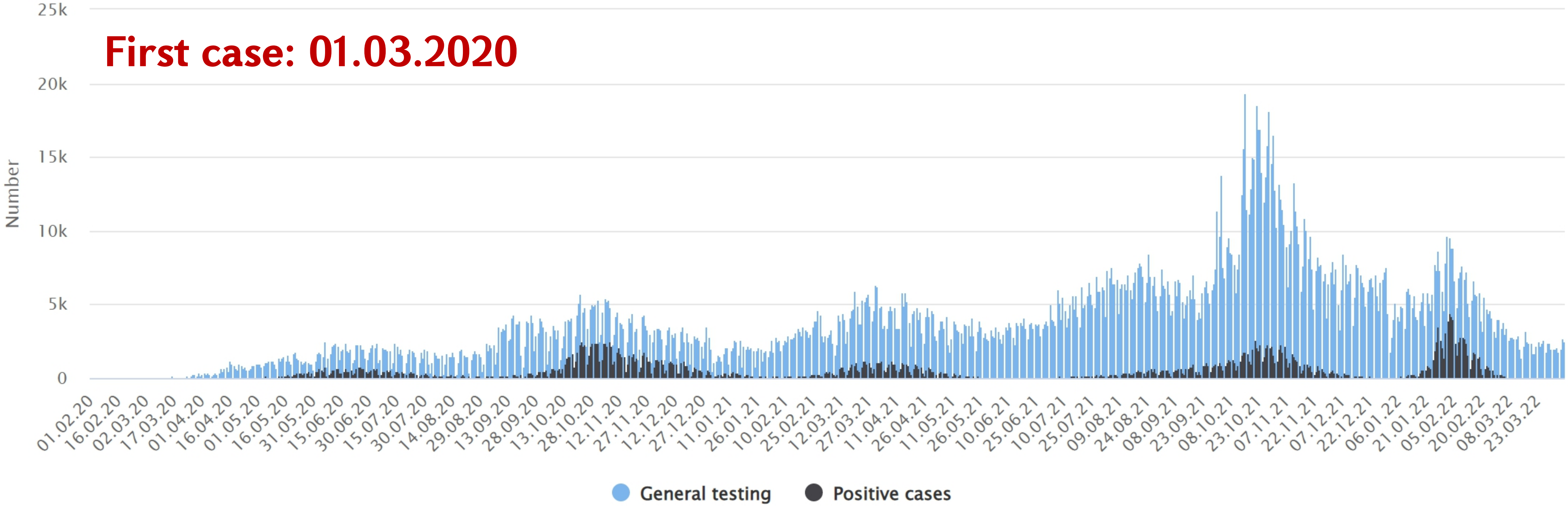
Pharmacogenetic study

On average 5.5% of individuals in the population use at least one of the 32 drugs associated with the studied genes on a daily basis.

estonian genome center
university of tartu

Epidemiological surveillance of SARS-COV-2

Covid-19 statistics in Armenia



First case: 01.03.2020

By April 6, 2020:

- Positive – 422,610 Recovered – 410,272 Deaths – 8,619
  Tests total – 2,988,475

Situation in Armenia

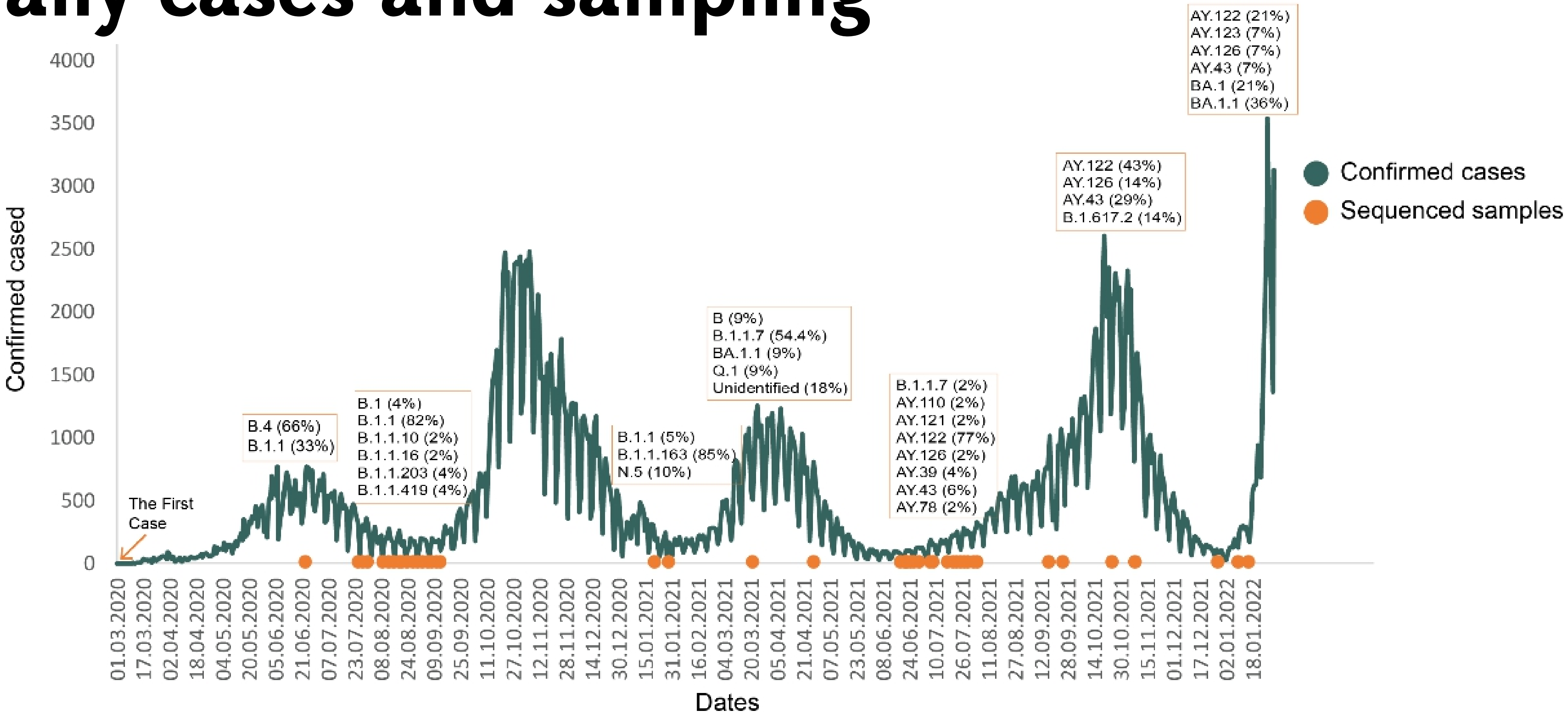
Sampling timeline	Number of samples	Results availability	Sequencing Institution
March-August 2020	3	December 2020	Institute of Virology Charité Universitätsmedizin Berlin (Germany)
September-November 2020	53	May 2021	Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center (USA) in collaboration with IMB

This delay seriously impeded the ability of objective analysis of molecular epidemiologic information and hampered the informed decision-making by health authorities.

Sample counts and sequencing scheme

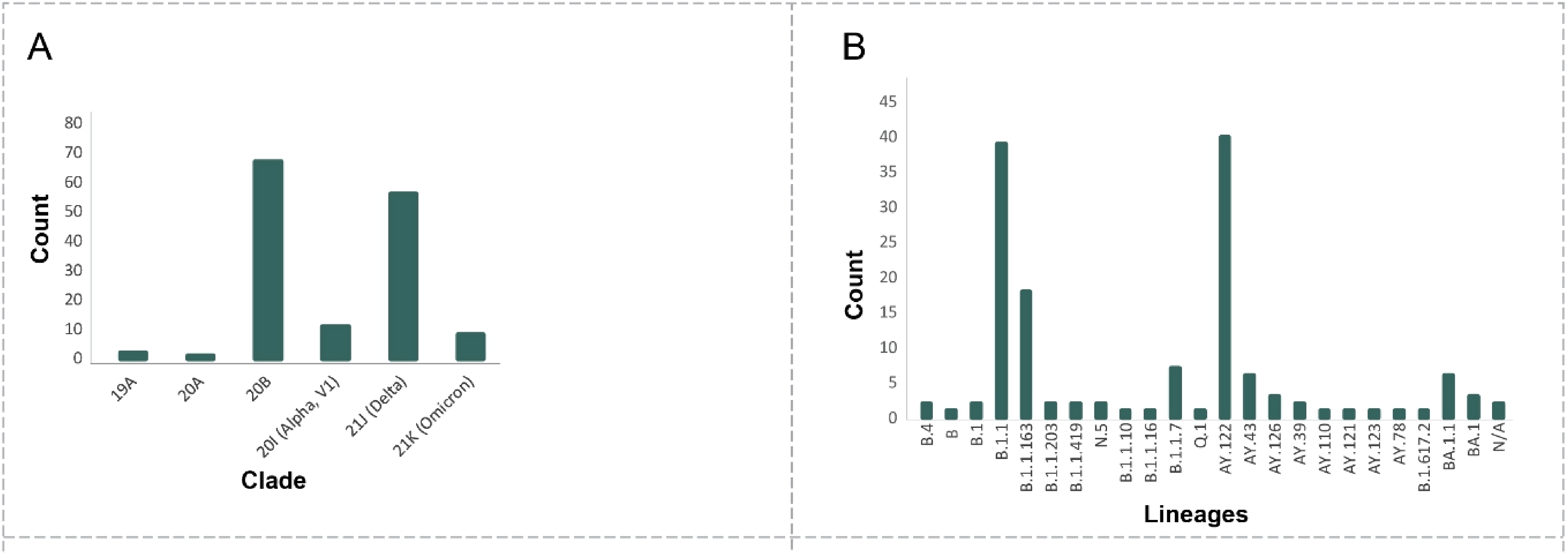
	Nanopore sequencing	Illumina sequencing	Obtained from GISAID
141 Samples	V	-	-
45 Samples	-	V	-
5 Samples	V	V	-
3 Samples	-	-	V
Total:	194 samples		

Daily cases and sampling



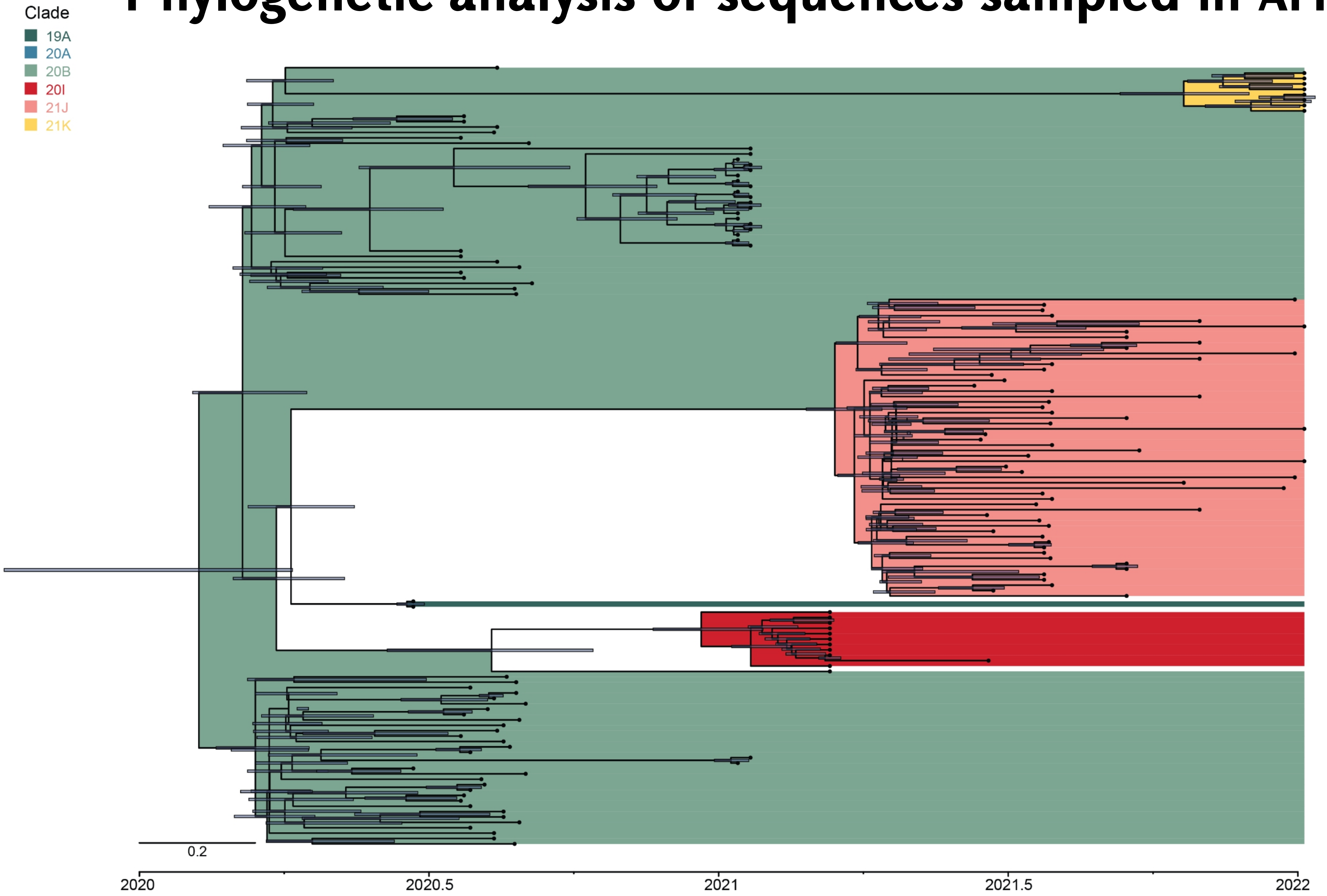
The total of 194 sequences represents 0.04% of 399,727 reported cases in Armenia (as of February 11th, 2022)

Clade and lineage diversity in Armenian sequences

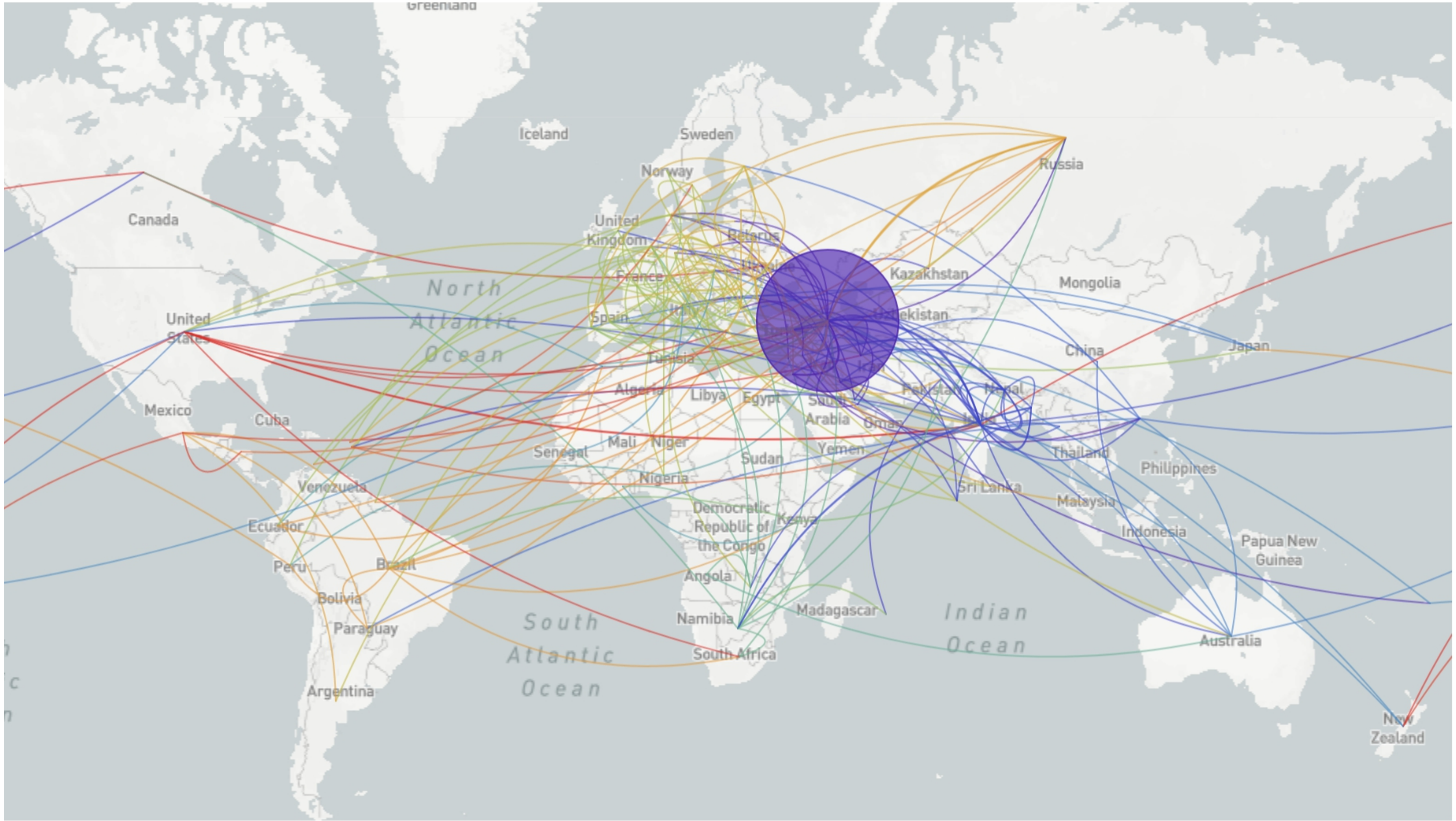


The highest genomic diversity was noticed for clades 21J (Delta) (9 PANGO lineages) and 20B (8 PANGO lineages).

Phylogenetic analysis of sequences sampled in Armenia



Inbound and outgoing transmission routes of SARS-CoV-2



The majority of importations inferred by phylogeographic analyses were through air-way travels, while ground transportation played very little or no role

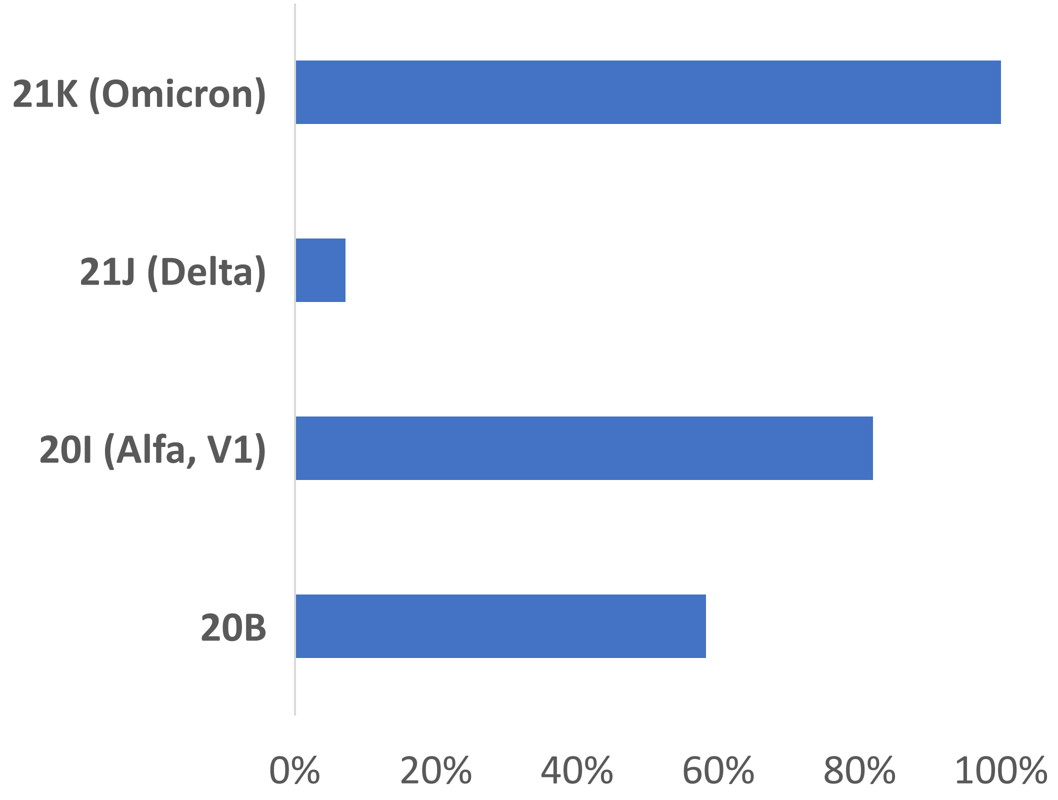
The majority of early importations were from countries with considerably large Armenian diaspora (such as Russia, Kazakhstan) as well as touristic destinations (Italy)

The geography of of later VOC lineages was much wider

Clade associated mutations associated with HLA epitopes

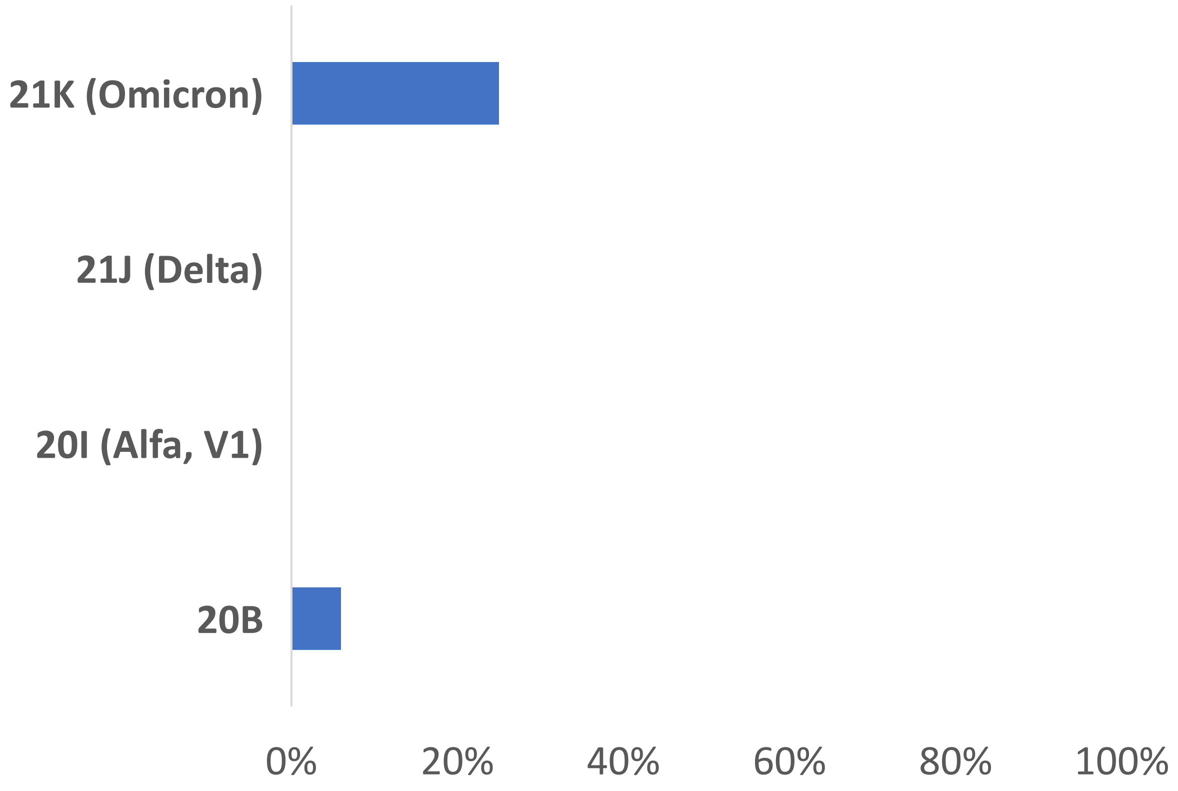
Protective HLA alleles

HLA-A*02:01, HLA-A*24:02



Risk HLA alleles

HLA-A*01:01, HLA-A*03:01, HLA-B*51:01



Collaborative agenda for bioinformatics and medical informatics

MI in support of functional genomics

- ‘Phenotype’ databases for clinical annotation of biological samples and for clinical validation of biological research results.
- Disease reclassification.
- Informatics for supporting rational drug design and development.

Collaborative agenda for bioinformatics and medical informatics

BI in support of individualized health care

- Including genetic data in the electronic health record.
- Methods for personalized health care: guidelines and decision-making support systems.
- Stratifying patients by their genetic profiles: molecular diagnosis, clinical trials, and pharmacogenomics.
- Point-of-care data collection and access.

Collaborative agenda for bioinformatics and medical informatics

Biomedical informatics in support of genomic medicine

- Molecular and functional imaging.
- Modelling and simulation for an approach that integrates physiology and pathology.
- Epidemiology: biobanks and population repositories.
- New methods for e-learning in genomic-based medicine.

Thank you!