BIO390
Introduction to Bioinformatics

Personal Genome Data, Bioinformatics & Genomic Privacy

Michael Baudis **UZH SIB**
Computational Oncogenomics

University of Zurich[UZH]

BIO390: Course Schedule

- 2018-09-18: Michael Baudis - **What is Bioinformatics? Introduction and Resources**
- 2018-09-25: Christian von Mering - **Sequence Informatics**
- 2018-10-02: Mark Robinson - **Statistical Bioinformatics**
- 2018-10-09: Katja Bärenfaller - **Proteomics**
- 2018-10-16: Izaskun Mallona - **Regulatory Genomics and Epigenomics**
- 2018-10-23: Daniel Stekhoven - **Clinical Bioinformatics**
- 2018-10-30: Shinichi Sunagawa - **Metagenomics**
- 2018-11-06: Amedeo Caflisch - **Molecular Dynamics**
- 2018-11-13: Elif Ozkrimli Olmez - **Protein Interactions**
- ~~2018-11-20: Fabio Rinaldi - **Text Mining I**~~
- 2018-11-27: Fabio Rinaldi - **Text Mining II**
- 2018-12-04: Andreas Wagner - **Biological Networks**
- 2018-12-11: Michael Baudis - **Human Genome Data - Risks & Opportunities || Feedback**
- 2018-12-18: Michael Baudis - **Exam (multiple choice)**

The trouble with human genome variation

Finding Somatic Mutations In Cancer: Many Needles in a Large Haystack

- a typical human genome (~3 billion base pairs) has ~5 million variants
- most of them are "**rare**"; i.e. can only be identified as recurring when sequencing thousands of people
- cancer cells accumulate additional variants, only **few** of which ("**drivers**") are relevant for the disease

Graphic adapted from Mark Gerstein (GersteinLab.org; @markgerstein)

BRAF V600E (c.1799T>A) Mutation
Oncogene Activation by Single Nucleotide Alteration

- a single nucleotide exchange **T**hymidine > **A**denine leads to continuous RAF based activation of the MEK-ERK pathway
- BRAF V600E is a frequent mutation in >50% of malignant melanomas, but also CRC, lung ADC …
- pharmacologic block of B-Raf (e.g. through ***Vemurafenib***)

The BRAF V600E mutation leads to continuos phosphorylation of MEK, without the need for receptor based activation of the upstream pathway and loss of inhibitory feedback control.

Michael Baudis | UZH | 2016-12-19

Sequencing everywhere, and everything

- dramatic increase in genome sequencing, both for human / disease related projects as well as basic biology (species identification, metagenome analyses, evolutionary modeling ...)

Muir et al. Genome Biology (2016) 17:53

Genomes Everywhere

Large Genome Data Generation, Analysis & Sharing Initiatives

Organization / Initiative: Name	Organization / Initiative: Category	Cohort
100K Wellness Project	Research Project	107 unaffected individuals (scaling up to 100,000)
23andMe	Organization	>1 million customers (>80% consented to research)
Actionable Cancer Genome Initiative (ACGI)	Data-Sharing Project	Goal: 100,000 individuals
Ancestry.com	Organization	1.4 million customer DNA samples (what % consented to research?)
BioBank Japan	Repository	Specimens from >200,000 patients and unaffected controls
Cancer Moonshot2020	Consortium	Phase 1: 20,000 cancer patients
Children's Hospital of Philadelphia Biorepository	Repository	Capacity for 8.6 million samples
China Kadoorie Biobank	Repository	>512,000 participants (general population, China). Genotyping data available for ~100,000.
CIMBA	Consortium	>15,000 BRCA1 carriers, >8,000 BRCA2 carriers
Clinical Sequencing Exploratory Research (CSER)	Consortium	~4,000 patients and healthy controls
DECIPHER	Repository	19,014 patients (international)
deCode Genetics	Organization	500,000 participants (international)
East London Genes & Health	Research Project	100,000 unaffected individuals (East London, Pakistani or Bangladeshi heritage)
Electronic Medical Records and Genomics (eMERGE) Network	Repository, Consortium, Research Project	55,028 patients
European Network for Genetic and Genomic Epidemiology (ENGAGE)	Research Project	80,000 GWAS scans, and DNA and serum/plasma from >600,000 individuals
Exome Aggregation Consortium (ExAC)	Consortium	60,706 individuals
GENIE/AACR	Data-Sharing Project	>17,000 cancer patients (international)
Genome Asia 100K	Consortium	Goal: 100,000 individuals (Asia)
Genomics England	Organization	Goal: 100,000 genomes from 70,000 individuals (rare disease & cancer patients, and their relatives)
GoT2D	Consortium, Data-Sharing Project	Multiple case-control cohorts
International Cancer Genome Consortium (ICGC)	Consortium	currently data from >16'000 samples
International Genomics of Alzheimer's Project (IGAP)	Consortium	40,000 patients with Alzheimer's disease
International Multiple Sclerosis Genetics (IMSG) Consortium	Consortium	Goal: >50,000 patients with MS
Kaiser Permanente: Genes, Environment, and Health (RPGEH)	Repository, Research Project	200,000 DNA samples (scaling up to 500,000)
Leiden Open Variation Database (LOVD)	Repository	>170,000 individuals
Million Veteran Program	Research Project	Goal: 1 million individuals; first 200,000 is complete.
MyCode® Community Health Initiative	Repository, Research Project	Goal: >250,000 patients
Precision Medicine Initiative	Research Project	Goal: >1 million participants, starting in 2016 (US)
Psychiatric Genomics Consortium (PGC)	Consortium	>170,000 subjects
Resilience Project	Research Project	589,306 individuals
Saudi Human Genome Program	Research Project	Goal: ~100,000 patients and controls (Saudi Arabia)
Scottish Genomes Partnership (SGP)	Research Project	>3,000 individuals (Scotland)
T2D-GENES	Consortium, Data-Sharing Project	10,000 patients and controls (five ethnicities); 600 individuals (Mexican American)
TBResist	Consortium	>2,600 samples
UK Biobank	Repository, Consortium, Research Project	500,000 individuals (age 40-69 years; UK)
UK10K	Research Project	10,000 participants (6,000 patients and 4,000 controls)
Vanderbilt's BioVU	Repository	>215,000 samples

Global Alliance for Genomics & Health

VARIANT RESOURCES FOR CANCER GENOMICS

Resource name	Primary institute	Constituent Knowledge base	Cancer focused	Therapeutic evidence	Predisp. evidence	Diagnostic evidence	Prognostic evidence	Variant emphasis	URL
Cancer Genome Interpreter (CGI)	Institute for Research in Biomedicine, Barcelona, Spain	x	x	x				Somatic	https://www.cancergenomeinterpreter.org/home
Clinical Interpretation of Variants in Cancer (CIViC)	Washington University School of Medicine (WashU)	x	x	x	x	x	x	All variants	www.civicdb.org
JAX Clinical Knowledgebase (CKB)	The Jackson Laboratory	x	x	x	x	x	x	Somatic	https://ckb.jax.org/
Molecular Match	Molecular Match	x	x	x			x	Somatic	https://app.molecularmatch.com/
OncoKB	Memorial Sloan Kettering Cancer Center	x	x	x				Somatic	http://oncokb.org/#/
Precision Medicine Knowledgebase (PMKB)	Weill Cornell Medical College	x	x	x	x	x	x	Somatic	https://pmkb.weill.cornell.edu/
BRCA exchange	GA4GH	x	x		x			Germline	http://brcaexchange.org/
Cancer Driver Log (CanDL)	Ohio State University (OSU) / James Cancer Hospital		x	x				Somatic	https://candl.osu.edu/
Gene Drug Knowledge Database	Synapse		x	x		x	x	Somatic	https://www.synapse.org/#!Synapse:syn2370773/wiki/62707
MatchMiner	Dana-Farber Cancer Institute		x					Somatic	http://bcb.dfci.harvard.edu/knowledge-systems/
COSMIC Drug Resistance Curation	Wellcome Trust Sanger Institute		x	x				Somatic	http://cancer.sanger.ac.uk/cosmic/drug_resistance
My Cancer Genome	Vanderbilt University		x	x		x	x	Somatic	https://www.mycancergenome.org/
NCI Clinical Trials	National Cancer Institute of the National Institutes of Health		x					Somatic	www.cancer.gov/about-cancer/treatment/clinical-trials
Personalized Cancer Therapy Database	MD Anderson Cancer Center		x	x	x	x	x	Somatic	https://pct.mdanderson.org/#/home
ClinGen Knowledge Base	ClinGen				x			Germline	https://www.clinicalgenome.org/resources-tools/
ClinVar	National Center for Biotechnology Information			x	x			All variants	http://www.ncbi.nlm.nih.gov/clinvar/
Pharmacogenomics Knowledgebase (PharmGKB)	Stanford University			x				Germline	https://www.pharmgkb.org/
The Human Gene Mutation Database (HGMD)	Institute of Medical Genetics in Cardiff				x			Germline	http://www.hgmd.cf.ac.uk

Wagner et al. A harmonized meta-knowledgebase of clinical interpretations of cancer genomic variants. Nature Genetics (under review 2018)

CANCER VARIANT KNOWLEDGE BASES

- cancer variant knowledge databases report evidences for disease association (causative, therapeutic targets...)
- data selection is driven by arbitrary observations and sample selections
- limited overlap of reported variant associations is evidence for large gaps in knowledge

Figure S2 - Knowledgebase overlap

(a) Upset plot of publications supporting clinical interpretations of variants. the overwhelming majority of publications are observed in only 1 of 6 resources. **(b)** Upset plot of genes described by clinical interpretations of variants. Compared to other interpretation elements, genes are much more commonly shared between resources.

Bias in Ascertainment / Background / Environment in Cancer Genome Studies

- the frequency of many genome variants depends on the genetic background
- cancer incidence & type can correlate to environmental factors
- geographic analysis can support interpretation and point to knowledge gaps

Geographic distribution of >140'000 cancer genome profiles reported in the literature. The numbers are derived from the 2947 publications registered in the Progenetix database.

progenetix

GA4GH to solve accessibility…

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

Search...

ABOUT US | HOW WE WORK | GA4GH TOOLKIT | NEWS & EVENTS | BECOME A MEMBER | CONTACT US

Enabling genomic data sharing for the benefit of human health

The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework

Genomic Data Toolkit →

Regulatory & Ethics Toolkit →

Data Security Toolkit →

VIEW OUR LEADERSHIP

MORE ABOUT US

BECOME A MEMBER

GA4GH API promotes sharing

GENOMICS

A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.

Global Alliance for Genomics & Health

GA4GH 2018: Interaction of Work Streams and Driver Projects

2018 GA4GH Driver Project Matrix

	BRCA	MME	ELIXIR Beacon	VICC	AGHA	GeL	ClinGen	ICGC ARGO	ENA EVA EGA	NCI GDC & DCF	Can DIG	Monarch Initiative	All of Us	HCA	TOP Med
Discovery		C	C	A	A	A	C		C			C			
LSG			A	A	A	A		A	C	X	X		A	C	A
DURI	A	A	C	A	C	C	A		C	X	C	C	C		
Cloud			C	A	C	C		C	A	C	C		A	C	C
GKS	C		C	C	C	C	C		C	C		C			X
Clin/Pheno	A	C	C	C	C	C	A	A	C		A	C	C		
REWS															
Security															

Legend

C = Contributing (active, developing tools)

A = Adopting (passive, observing progress)

X = TBC

ELIXIR Platforms

- Data
- Training
- Compute
- Interoperability
- Tools

ELIXIR Members

Belgium, Czech Republic, Denmark, EMBL, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Luxembourg, Israel, Netherlands, Norway, Slovenia, Portugal, Spain, Sweden, Switzerland, United Kingdom

ELIXIR Observers

Cyprus

www.elixir-europe.org

@ELIXIREurope

/company/elixir-europe

Beacon Project

An open web service that tests the willingness of international sites to share genetic data.

> 50 Beacons throughout the world

Global Alliance for Genomics & Health

ELIXIR Beacon Project

- Driver project on GA4GH roadmap
- aligns with Discovery Work Stream
- strong impact on GA4GH developments as a concrete, funded project

Global Alliance for Genomics & Health

Beacon *forward*

☑ **structural variations** (DUP, DEL) in addition to SNV

- … more structural queries (translocations/fusions…)
- (bio-) **metadata** queries

☑ layered authentication system using **ELIXIR AAI**

- quantitative responses
- Beacon queries as entry for **data delivery** (outside Beacon protocol)
- Ubiquitous **deployment** (e.g. throughout ELIXIR network)

ELIXIR - Making Beacons Biomedical

- Authentication to enable non-aggregate, patient derived datasets
  - ELIXIR AAI with compatibility to other providers (OAuth...)
- Scoping queries through "biodata" parameters
- Extending the queries towards clinically ubiquitous variant formats
  - cytogenetic annotations, named variants, variant effects
- Beacons as part of local, secure environments
  - local EGA ...
- Beacon queries as entry for **data delivery**
  - handover to stream and download using htsget, VCF, EHRs
- Interacting with EHR standards
  - FHIR translations for queries and handover ...

University of Zurich UZH

Beacon+ - Extending the protocol ...

beacon.progenetix.org

Beacon+

This forward looking Beacon interface implements additional, planned features beyond the current GA4GH specifications. Info

Query

SNV Example | DGV Example | CNV Example

Dataset	tcga
Reference name*	9
Genome Assembly*	GRCh38 / hg
Start *min* Position*	19,500,000
Start *max* Position	21,975,098
End *min* Position	21,967,753
End *max* Position	24,500,000
Alt. Base(s)*	DEL
Bio-ontology	icdot:c50.9: (4065)

Beacon Implementations

- implementing existing resources with Beacon protocol
- e.g. TCGA cancer variants (structural and SNV)

Beacon Response

- quantitative (counts for variants, callsets and samples)
- *Handover* to authentication system for data retrieval
- **no exposure** of data beyond standard Beacon response and additional pointer to matched data

Prototyping Query Extensions

- testing e.g. bio-metadata queries using ontology terms

Dataset	Assembly	Chro	Start Range	End Range	Pos	Ref Alt	Bio Query	Variants Calls Samples	$f_{alleles}$	Response Context
tcga	hg38	9	19,500,000 21,975,098	21,967,753 24,500,000		DEL	icdot:c50.9	54 54 54	0.0243	JSON UCSC Handover

arrayMap progenetix This Beacon implementation is developed by the Computational Oncogenomics Group at the University of Zurich, with support from the SIB Technology group and ELIXIR.

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

elixir

Querying Copy Number Variants

- copy number variants (CNV) are a typical type of "imprecise" structural changes
- "fuzzy" queries for start and end positions

chr12:94,306,043-98,466,437:DEL

start_min: 94,000,000
start_max: 94,500,000

reference_name: “9”

variant_type: “DEL”

end_min: 98,200,000
end_max: 98,700,000

```
{
   "allele_request" : {
      "$and" : [
         { "reference_name" : “9" },
         { "variant_type" : “DEL" },
         { "start" : { "$gte" : 19500000 } },
         { "start" : { "$lte" : 21984490 } },
         { "end" : { "$gte" : 21957751 } },
         { "end" : { "$lte" : 24500000 } }
      ]
   },
   "api_version" : "0.4",
   "beacon_id" : "org.progenetix:progenetix-beacon",
   "exists" : true,
   "info" : {
      "query_string" :
"datasetId=arraymap&referenceName=chr9&assemblyId=GRCh38&variantType=DE
L&startMax=19000000&startMin=21984490&endMin=21900000&endMax=25000000&b
iosamples.bio_characteristics.ontology_terms.term_id=icdom:9440_3",
      "version" : "Beacon+ implementation based on a development branch
of the beacon-team project: https://github.com/ga4gh/beacon-team/pull/
94"
   },
   "url" : “http://progenetix.org/beacon/info/“,
   "dataset_allele_responses" : [
      {
         “datasetId" : "arraymap",
         "error" : null,
         "exists" : true,
         "external_url" : “http://arraymap.org",
         "sample_count" : 584,
         "call_count" : 3781,
         "variant_count" : 3244,
         "frequency" : 0.0094,
         "info" : {
            "description" : "The query was against database
\"arraymap_ga4gh\", variant collection \"variants_cnv_grch36\". 3781 /
59428 matched callsets for 3602919 variants. Out of 62105 biosamples in
the database, 2047 matched the biosample query; of those, 584 had the
variant.",
            "ontology_ids" : [
               "ncit:C3058",
               "pgx:icdom:9440_3",
               "pgx:icdot:C71.9",
               “pgx:icdot:C71.0"
            ]
         },
      }
   ],
```

start_min
start_max
end_min
end_max

Translation for Store
(here MongoDB)

Match using query ranges “at least one base in interval affected”

Region of Interest, e.g. CDR of Gene (here: CDKN2A)

Example “focal” matches (overlap w/ size limit)

Mismatches
- too large
- end outside
- start outside

- Beacon+ **range queries** allow the definition of a genome region of interest, containing a specified variant (or other mappable feature)
- “fuzzy” matching of region ends is essential for features without base specific positions
- current Beacon implementation addresses CNV (DUP, DEL), as are specified in VCF && GA4GH variant schema

Bio-Metadata: Query ontology classes

- Emerging consensus about representation of "bio-characteristics" (phenotypes, diseases, observations...) through ontology classes as robust classifiers
- Convergence around *Phenopackets* (PXF) as implementation for data exchange
- Use of (one or multiple) namespace mapped ontology classes as first implementation of a "bio-query" for Beacon?

```
// A class (aka term, concept) in an ontology
message OntologyClass {
    // a CURIE-style identifier e.g. HP:0100024, MP:0001284, UBERON:0001690.
    // This is the primary key for the ontology class
    // REQUIRED?
    string id = 1;

    // class label, aka name. E.g. "Abnormality of cardiovascular system"
    string label = 2;
}
```

Phenopackets

```
// An ontology term describing an attribute. (e.g. the phenotype attribute
// 'polydactyly' from HPO)
message OntologyTerm {
  // Ontology term identifier - the CURIE f
  // differs from the standard GA4GH schema
  // in that it is a CURIE pointing to an information resource outside of the
  // scope of the schema or its resource implementation.
  string term_id = 1;

  // Ontology term - the label of the ontology term the termId is pointing to.
  string term_label = 2;
}
```

GA4GH::Metadata

```
"bio_character
  {
    "descripti
    "ontology_
      {
        "term_
        "term_id" : "ncit:C3039"
      },
      {
        "term_label" : "Glioma NOS",
        "term_id" : "pgx:icdom:9380_3"
      },
      {
        "term_label" : "Brain NOS",
        "term_id" : "pgx:icdot:C71.9"
      }
    ],
    "negated_ontology_terms" : [ ]
  }
],
```

...wrapping it up

Beacon & Handover

Future Beacons to support advanced types of queries and connection of **data delivery services**

Beacon I/O

Handover

Authentication

Global Alliance for Genomics & Health

SIB

University of Zurich UZH

elixir

ELIXIR - Making Beacons Biomedical

- Authentication to enable non-aggregate, patient derived datasets
  - ELIXIR AAI with compatibility to other providers (OAuth...)
- Scoping queries through "biodata" parameters
- Extending the queries towards clinically ubiquitous variant formats
  - cytogenetic annotations, named variants, variant effects
- Beacons as part of local, secure environments
  - local EGA ...
- Beacon queries as entry for **data delivery**
  - handover to stream and download using htsget, VCF, EHRs
- Interacting with EHR standards
  - FHIR translations for queries and handover ...

GA4GH::GKS
GA4GH::CP

DUO
ADA-M

SPHN - The Swiss Personalized Health Network

sPHN swiss PERSONALIZED HEALTH NETWORK

Steps #1

July 2016

SPHN: Setting up the Organisation

- Prof. Peter Meier-Abt, President

Dear Personalized Health Community, we welcome you to this first newsletter – Steps #1 – about the Swiss Personalized Health Network. In the SAMS bulletin 1/2016[1] we have informed you about the goals, the contents, the organisational structure and the implementation steps of this initiative. Meanwhile we have started to set up the SPHN organisation. The National Steering Board (NSB), the Executive Board (EB) and the Data Expert Group (DEG) are established, and a Managing Director is elected. In addition, the Rules of Procedures have been defined and the principles of Funding Procedures are in preparation. The latter is also true for the composition of the Ethical-Legal-Social-Issues Advisory Group (ELSIag), the International Advisory Board (IAB), the Data Organization and the definition of the Data Sharing principles.

Phase I: Organizational structure

Spring/Summer 2016
- National Steering Board (NSB)
- Data Expert Group (DEG)
- Executive Board (EB)

Autumn/Winter 2016
- Management Office (MO)
- ELSI Advisory Group (ELSIag)
- Scientific Advisory Board (SAB)

December 2016: Decision in the federal parliament

2016 | 2017 | 2018 | 2019 | 2020 | 2021+

Infrastructure and Platform Implementation

SNSF Research Funding for Personalized Health

Phase II: Implementation and operations
- Establish official SPHN project organization
- Definition of SPHN fund allocation procedures
- Implementation of informed consent and data sharing policies
- Implementation of a Data Coordination Center
- Implementation of data quality and interoperability standards
- Research project funding by SNSF

Figure 1: Schedule according to the report «Implementation of the SPHN Initiative» (November 2015)

The SPHN initiative aims for a nationwide harmonisation of molecular and clinical data semantics and of health information technology systems in order to achieve data interoperability among all relevant stakeholders such as University Hospitals, Universities, Swiss Federal Institutes of Technology (ETHZ, EPFL), research funding institutions (e.g. SNSF), other research performing hospitals, political authorities (e.g. FOPH) and relevant biomedical organizations (e.g. Swiss Biobanking Platform). It is planned to run the SPHN initiative over more than just one Federal ERI period of four years. As specified in the adopted implementation report of November 2015[2], initial funding shall focus on the development of essential infrastructures. Subsequently, funding of research projects will become more and more important (see figure 1, left).

With combined efforts we want to set the stage for Switzerland becoming a leading nation in personalized health research. We will inform you regularly with newsletters about the actual stage of development and important news about the SPHN. The next edition – Steps #2 – describes the Data Organization and the Data Coordination Center (DCC) as planned by the Swiss Institute of Bioinformatics. We are looking forward to working together with you and to making the SPHN initiative a success for the benefit of medical research, the health system and for the whole of Swiss society.

1 www.samw.ch/de/Publikationen/Bulletins/Aktuell.html

2 http://www.samw.ch/en/Projects/Personalized-Health.html

1

The next edition – Steps #2 – describes

the Data Organization and the Data

Coordination Center (DCC)

....

National level
mandates
Swiss Confederation
SNSF
Swiss Biobanking Platform project
SAMS
appoints
operates
National Steering Board (NSB)
Management Office (MO)
International Advisory Board (IAB)
Research Projects
NSB Ausschuss
Institutional level
Personalized Health Platforms
– Biomedical platforms
– Private hospitals
– Private companies
– Universities/ETH, EPF
– University hospitals
– Public hospitals
– Public Health
Technical level
Biobanking, biomonitoring related activities
-omics data
Clinical data
Patient data
Citizen data
Interoperability standards
Data Coordination Center (DCC)
Data Expert Group (DEG)
BioMedIT RI Network
SIB Swiss Institute of Bioinformatics
Scientific Expert Board (SEB)
Ethical Legal Social Issues (ELSIag)

BioDA Groundwork Topics

- Assessment of tools & solutions both from current Swiss practice and international systems & providers
  - Use case catalogue & requirements analyses
  - "Wants" vs. known solutions
- Identifying reference datasets with data access information
- Active point of interaction with international organisations (GA4GH, EBI/EGA, ELIXIR ...), including tracking of (upcoming) **standards** & solutions

Genomes & Privacy

Generalkonsent

BENEFIT

BLOCKCHAIN

HEALTH

PRIVACY

SECURITY

CONSENT

ACCESS

Right to Research

HACKERS

LAWS

Genetic
Information
Nondiscrimination
Act

Health
Insurance
Portability and
Accountability
Act

SAFETY

CRYPTOGRAPHY

NETFLIX

Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update

Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the $1M Grand Prize to team “BellKor’s Pragmatic Chaos”. Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

FAQ | Forum | Netflix Home

© 1997-2009 Netflix, Inc. All rights reserved.

Linking Attacks: Case of Netflix Prize

NETFLIX

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
…	…	…	…
…	…	…	…

IMDb

Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
…	…	…	…
…	…	…	…
…	…	…	…

- **Many users are shared**
- **The grades of same users are correlated**
- **A user grades one movie around the same date in two databases**

Anonymized Netflix Prize Training Dataset made available to contestants

2 Sides of the Coin for RNA-seq:
Ensuring Individual Privacy v. Allowing Easy Mining
Mark Gerstein, Yale
Slides from lectures.GersteinLab.org (2018)

Linking Attacks: Case of Netflix Prize

NETFLIX ⟷ IMDb

Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
…	…	…	…
…	…	…	…

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
…	…	…	…
…	…	…	…
…	…	…	…

- Many users are shared
- The grades of same users are correlated
- A user grades one movie around the same date in two databases

- IMDB users are public

- NetFLIX and IMdB moves are public

Linking Attacks: Case of Netflix Prize

NETFLIX ⟷ IMDb

Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
…	…	…	…
…	…	…	…

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
…	…	…	…
…	…	…	…
…	…	…	…

- Many users are shared
- The grades of same users are correlated
- A user grades one movie around the same date in two databases

2 Sides of the Coin for RNA-seq:
Ensuring Individual Privacy v. Allowing Easy Mining
Mark Gerstein, Yale
Slides from lectures.GersteinLab.org (2018)

Linking Attack Scenario

[Harmanciet al. Nat. Meth. ('16)]

2 Sides of the Coin for RNA-seq:
Ensuring Individual Privacy v. Allowing Easy Mining
Mark Gerstein, Yale
Slides from lectures.GersteinLab.org (2018)

Summary of a Linking Attack

Ancestry signatures in common genotyping platforms

Figure S1 The fraction or contribution of theoretical ancestors (k=9) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed neighboring to each other: AFR (1-7), SAS (8-12), EAS (13-17), EUR (18-22), AMR (23-26).

Beacon Project

An open web service that tests the willingness of international sites to share genetic data.

> 50 Beacons throughout the world

Global Alliance for Genomics & Health

Finding Somatic Mutations In Cancer: Many Needles in a Large Haystack

- a typical human genome (~3 billion base pairs) has ~5 million variants
- most of them are "**rare**"; i.e. can only be identified as recurring when sequencing thousands of people
- cancer cells accumulate additional variants, only **few** of which ("**drivers**") are relevant for the disease

Graphic adapted from Mark Gerstein (GersteinLab.org; @markgerstein)

Genome ***Beacons*** Compromise Security?

Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals

Stanford researchers identify potential security hole in genomic data-sharing network

Hackers with access to a person's genome might find out if that genome is in an international network of disease databases.

OCT 29
2015

Sharing genomic information among researchers is critical to the advance of biomedical research. Yet genomic data contains identifiable information and, in the wrong hands, poses a risk to individual privacy. If someone had access to your genome sequence — either directly from your saliva or other tissues, or from a popular genomic information service — they could check to see if you appear in a database of people with certain medical conditions, such as heart disease, lung cancer or autism.

Stanford researchers are working with the Global Alliance for Genomics and Health to make genomic information in the Beacon Project more secure.
Science photo/Shutterstock

Work by a pair of researchers at the Stanford University School of Medicine makes that genomic data more secure. Suyash Shringarpure, PhD, a postdoctoral scholar in genetics, and Carlos Bustamante, PhD, a professor of genetics, have demonstrated a technique for hacking a network of global genomic databases and how to prevent it. They are working with investigators from the Global Alliance for Genomics and Health on implementing preventive measures.

The work, published Oct. 29 in *The American Journal of Human Genetics,* also bears importantly on the larger question of how to analyze mixtures of genomes, such as those from different people at a crime scene.

IDENTIFICATION OF INDIVIDUALS FROM MIXED COLLECTIONS USING RARE ALLELES

The American Journal of Human Genetics 97, 631–646, November 5, 2015

Privacy Risks from Genomic Data-Sharing Beacons

Suyash S. Shringarpure[1,*] and Carlos D. Bustamante[1,*]

The human genetics community needs robust protocols that enable secure sharing of genomic data from participants in genetic research. Beacons are web servers that answer allele-presence queries—such as "Do you have a genome that has a specific nucleotide (e.g., A) at a specific genomic position (e.g., position 11,272 on chromosome 1)?"—with either "yes" or "no." Here, we show that individuals in a beacon are susceptible to re-identification even if the only data shared include presence or absence information about alleles in a beacon. Specifically, we propose a likelihood-ratio test of whether a given individual is present in a given genetic beacon. Our test is not dependent on allele frequencies and is the most powerful test for a specified false-positive rate. Through simulations, we showed that in a beacon with 1,000 individuals, re-identification is possible with just 5,000 queries. Relatives can also be identified in the beacon. Re-identification is possible even in the presence of sequencing errors and variant-calling differences. In a beacon constructed with 65 European individuals from the 1000 Genomes Project, we demonstrated that it is possible to detect membership in the beacon with just 250 SNPs. With just 1,000 SNP queries, we were able to detect the presence of an individual genome from the Personal Genome Project in an existing beacon. Our results show that beacons can disclose membership and implied phenotypic information about participants and do not protect privacy a priori. We discuss risk mitigation through policies and standards such as not allowing anonymous pings of genetic beacons and requiring minimum beacon sizes.

Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

- rare allelic variants can be used to identify an individual (or her relatives) in a genome collection without having access to individual datasets
- however, such an approach requires previous knowledge about the individual's SNPs

IDENTIFICATION OF INDIVIDUALS BASED ON "GENOMIC FINGERPRINTS"

18 JANUARY 2013 VOL 339 SCIENCE www.sciencemag.org

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.

Fig. 2. The top five records retrieved after searching Ysearch with the Y-STR haplotypes of Michael Snyder, John West, and Craig Venter. The expected number of generations to the MRCA is given in parentheses for each record. Searching with Craig Venter returned a "Venter" record (closed bar) as the top match.

- Genomic data of many types can be used to re-identify individuals in data collections

Routes for breaching and protecting genetic privacy

The map contrasts different scenarios, such as identifying de-identified genetic data sets, revealing an attribute from genetic data and unmasking of data. It also shows the interdependencies between the techniques and suggests potential routes to exploit further information after the completion of one attack. There are several simplifying assumptions (black circles). In certain scenarios (such as insurance decisions), uncertainty about the target's identity within a small group of people could still be considered a success (assumption 1). For certain privacy harms (such as surveillance), identity tracing can be considered a success and the end point of the process (assumption 2). The complete DNA sequence is not always necessary (assumption 3).

Yaniv Erlich & Arvind Narayanan. *Nature Reviews Genetics* 15, 409–421 (2014)

Long-Range Familial Searches

MyHeritage DNA

Hi Michael,

Good news! We've discovered new DNA Matches for you.

- Commercial, "Direct to Customer" DNA analyses are provided through independend sites and such affiliated to genealogy services (MyHeritage, Ancestry.com, 23andMe...)
- Genealogy sites identify individuals with matching haplotype blocks & provide a prediction about degree of genetic relation
- Law enforcement agencies (and who else?!) can send individual SNP profiles (e.g. recovered from evidence many years after a crime) using a Jane Doe identity, to identify relatives of the suspect - **long range familial search**

Daily Journal

Helping Northeast Mississippi Grow!

We're donating a portion of every 1-year or 6-month subscription to Tupelo High Band Boosters!

842-2613 or djournal.com/subscribe

New home delivery subscriptions only | Offer ends June 30

ALL SEC

E-EDITION

SUBSCRIBE

Devaughn had never been a suspect until genetic genealogy put police on his trail several months ago. Earlier this year, police sent the DNA profile to Parabon, a private genetics company, to compare the suspect's DNA sample to a public genealogy DNA database looking for people with similar DNA profiles who might be kin to the suspect. That eventually led authorities to look at Devaughn.

Rienzi man charged with 1990 Starkville murder

By William Moore Daily Journal 15 hrs ago Comments

© Copyright 2018 Daily Journal, 1242 S Green St Tupelo, MS

The New York Times

How a Genealogy Site Led to the Front Door of the Golden State Killer Suspect

Investigators used DNA from crime scenes that had been stored all these years and plugged the genetic profile of the suspected assailant into an online genealogy database. One such service, GEDmatch, said in a statement on Friday that law enforcement officials had used its database to crack the case. Officers found distant relatives of Mr. DeAngelo's and, despite his years of eluding the authorities, traced their DNA to his front door.

The New York Times, April 26, 2018

Attacks Associated With the Golden State Killer

Science

REPORTS

Cite as: Y. Erlich *et al.*, *Science* 10.1126/science.aau4832 (2018).

Identity inference of genomic data using long-range familial searches

Yaniv Erlich[1,2,3,4]*, Tal Shor[1], Itsik Pe'er[2,3], Shai Carmi[5]

[1]MyHeritage, Or Yehuda 6037606, Israel. [2]Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY, USA. [3]Center for Computational Biology and Bioinformatics (C2B2), Department of Systems Biology, Columbia University, New York, NY, USA. [4]New York Genome Center, New York, NY, USA. [5]Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel.

Consumer genomics databases have reached the scale of millions of individuals. Recently, law enforcement authorities have exploited some of these databases to identify suspects via distant familial relatives. Using genomic data of 1.28 million individuals tested with consumer genomics, we investigated the power of this technique. We project that about 60% of the searches for individuals of European-descent will result in a third cousin or closer match, which can allow their identification using demographic identifiers. Moreover, the technique could implicate nearly any US-individual of European-descent in the near future. We demonstrate that the technique can also identify research participants of a public sequencing project. Based on these results, we propose a potential mitigation strategy and policy implications to human subject research.

Fig. 3. Tracing a 1000Genomes sample using long range familial search. In black: the CEU pedigree. To respect the privacy of the family, we omitted the sample identifiers and the exact pedigree structure. A GEDmatch search of the person of interest (black circle) returned two males (squares with gray dots) with a total IBD sharing of 180cM and 171cM to the target, respectively, and 62cM between themselves. Using public genealogical records, we identified the ancestral couple (asterisk) of the matches and the person of interest.

Genetic genealogy databases can identify 60 per cent of people — even if they haven't been tested

Case	Announcement	By	Closest match	Comments
Buckskin Girl	April 9, 2018	DNA Doe Project	First cousin one removed	
Golden State Killer	April 24, 2018	Barbara Rae-Venter	Third cousin	
Lyle Stevik	May 8, 2018	DNA Doe Project	Second cousin	Inbreeding complicated the estimation of the match.
William Earl Talbott II	May 21, 2018	Parabon	Half first cousin once removed	Second cousins were identified as well
Joseph Newton Chandler III	June 21 2018	DNA Doe Project	Second cousin once removed	
Gary Hartman	June 22, 2018	Parabon	Half first-cousin	Genealogists were able to overcome non-paternity event in the tree of the suspect.
Raymond "DJ Freez" Rowe	June 25, 2018	Parabon	-	
James Otto Earhart	June 26, 2018	Parabon	Second cousin	
John D. Miller	July 15, 2018	Parabon	-	
Matthew Dusseault/Tyler Grenon	July 28, 2018	Parabon	-	
Spencer Glen Monnett	July 29, 2018	Parabon	-	Active case for a crime that occurred in April 2018
Darold Wayne Bowden	August 23, 2018	Parabon	-	
Michael F. Henslick	August 29, 2018	Parabon	-	

Erlich, Y., Shor, T., Pe'er, I., & Carmi, S. (2018). Identity inference of genomic data using long-range familial searches. Science, eaau4832.

Rapid re-identification of human samples

...

We developed a rapid, inexpensive, and portable strategy to re-identify human DNA using the MinION. Our strategy requires only ~60 min preparation and 5-30 minutes of MinION sequencing, works with low input DNA, and enables familial searches using Direct-to-Consumer genomic reference datasets. This method can be implemented in a variety of fields:

Forensics

Identification of abandoned meterial using DNA fingerprinting is a common practice. The main challange currently being: time. Our method allows rapid sample preparation at the crime scene (see movie). We envision that the method can be adopted in the field for rapid checks, after a mass disaster, and can be adopted in border control to fight human traffacking.

Clinic

Clinics procces many samples, either for analysis or, for example, organ donations. These samples are DNA fingerprinted to prevent sample mix-up mistakes. Our method can be implemnted in the clinic for rapid sanitiy-check of all incoming samples.

Cell line identification

Cross contamination of cell lines in science is a major problem. It results in unreproducible data, and clinical trails based on inaccurate findings. This problem costs billions of dollars per year. We envision labs can adopt our identification method to ensure the purity of the cell line, and detect contamination.

“DEMOCRATIZING DNA FINGERPRINTING”

Sophie Zaaijer, Assaf Gordon, Robert Piccone, Daniel Speyer, Yaniv Erlich, 2016

ddf.teamerlich.org

- DNA sequencing for identification/fingerprinting soon “commodity” technology (in contrast with technological/data challenges in “precision medicine”)

Rapid re-identification of human samples using portable DNA sequencing

Sophie Zaaijer[1,2†*], Assaf Gordon[2], Daniel Speyer[1,2], Robert Piccone[3], Simon Cornelis Groen[4], Yaniv Erlich[1,2,5*]

Re-identification of three DNA samples against a database with 31,000 individuals.

However ...

The Right to Scientific Knowledge

- In 1948, the General assembly of the United nations adopted the Universal Declaration of Human Rights (UDHr) to guarantee the rights of every individual in the world. Included were twin rights “to share in scienti c advancement and its bene ts” and “to the protection of the moral and material interests resulting from any scien- ti c...production of which [a person] is the author” (art. 27, United nations 1948).

from *Knoppers et al, 2014*

Hum Genet (2014) 133:895–903
DOI 10.1007/s00439-014-1432-6

ORIGINAL INVESTIGATION

A human rights approach to an international code of conduct for genomic and clinical data sharing

Bartha M. Knoppers · Jennifer R. Harris · Isabelle Budin-Ljøsne · Edward S. Dove

Received: 9 December 2013 / Accepted: 16 February 2014 / Published online: 27 February 2014
© The Author(s) 2014. This article is published with open access at Springerlink.com

Abstract Fostering data sharing is a scientific and ethical imperative. Health gains can be achieved more comprehensively and quickly by combining large, information-rich datasets from across conventionally siloed disciplines and geographic areas. While collaboration for data sharing is increasingly embraced by policymakers and the international biomedical community, we lack a common ethical and legal framework to connect regulators, funders, consortia, and research projects so as to facilitate genomic and clinical data linkage, global science collaboration, and responsible research conduct. Governance tools can be used to responsibly steer the sharing of data for proper stewardship of research discovery, genomics research resources, and their clinical applications. In this article, we propose that an international code of conduct be designed to enable global genomic and clinical data sharing for biomedical research. To give this proposed code universal application and accountability, however, we propose to position it within a human rights framework. This proposition is not without precedent: international treaties have long recognized that everyone has a right to the benefits of scientific progress and its applications, and a right to the protection of the moral and material interests resulting from scientific productions. It is time to apply these twin rights to internationally collaborative genomic and clinical data sharing.

B. M. Knoppers (✉) · E. S. Dove
Centre of Genomics and Policy, McGill University, 740 Dr. Penfield Avenue, Suite 5200, Montreal H3A 0G1, Canada
e-mail: bartha.knoppers@mcgill.ca

E. S. Dove
e-mail: edward.dove@mcgill.ca

J. R. Harris · I. Budin-Ljøsne
Division of Epidemiology, Department of Genes and Environment, Norwegian Institute of Public Health, PO Box 4404, Nydalen 0403, Oslo, Norway
e-mail: Jennifer.Harris@fhi.no

I. Budin-Ljøsne
e-mail: Isabelle.Budin.Ljosne@fhi.no

Introduction

In 1948, the General Assembly of the United Nations adopted the *Universal Declaration of Human Rights* (UDHR) to guarantee the rights of every individual in the world. Included were twin rights “to share in scientific advancement and its benefits” and “to the protection of the moral and material interests resulting from any scientific…production of which [a person] is the author” (Art. 27, United Nations 1948). In the 21st century, where are we in realizing the sharing of scientific advancement and its benefits, and the importance of protecting a scientific producer’s moral and material interests? In this article, we argue that these little-developed twin rights, what we call the right “to benefit from” and “to be recognized for”, have direct application to internationally collaborative genomic and clinical data sharing, and can be activated through an international code of conduct.

Sharing genomic and clinical data is critical to achieve precision medicine (National Research Council 2011), that is, more accurate disease classification based on molecular profiles to enable tailored effective treatments, interventions, and models for prevention. Better communication flow across borders and research teams, encompassing data from clinical and population research, enables researchers to connect the diverse types of datasets and expertise needed to elucidate the genomic basis and complexities of disease etiology. Such data integration can make it possible to reveal the genetic basis of cancer, inherited diseases,

Modernizing Patient Consent

- forward looking, transparent and technically feasible regulations for enabling access to research material and data while empowering patients

Generalkonsent: Eine einheitliche Vorlage soll schweizweite Forschung erleichtern

Personenbezug \ Art des Forschungs-materials	Biologisches Material und genetische Daten	Nicht-genetische Daten
Unverschlüsselt (identifizierend)	Information + Einwilligung in jedes einzelne Forschungsprojekt	Information über Weiterverwendung für zukünftige noch unbestimmte Forschungsprojekte + Generalkonsent für Forschungszwecke
Verschlüsselt	Information über Weiterverwendung für zukünftige noch unbestimmte Forschungs-projekte + Generalkonsent für Forschungs-zwecke	Information über Weiterverwendung für zukünftige noch unbestimmte Forschungs-projekte + über Möglichkeit Weiterverwen-dung abzulehnen > Widerspruchsrecht
Anonymisiert	**Genetische Daten:** Information über Weiterverwendung für zukünftige noch unbestimmte Forschungszwecke + über Möglichkeit Weiterverwendung abzulehnen > Widerspruchsrecht **Proben:** Information zur Anonymisierung > Widerspruchsrecht	Ausserhalb des Geltungsbereichs des HFG

Switzerland: Definition of a unified "Generalkonsent", to provide a single framework to manage permissions for access to patient derived material and related data

Consent Codes: Upholding Standard Data Use Conditions

Stephanie O. M. Dyke[1]*, Anthony A. Philippakis[2], Jordi Rambla De Argila[3,4], Dina N. Paltoo[5], Erin S. Luetkemeier[5], Bartha M. Knoppers[1], Anthony J. Brookes[6], J. Dylan Spalding[7], Mark Thompson[8], Marco Roos[8], Kym M. Boycott[9], Michael Brudno[10,11], Matthew Hurles[12], Heidi L. Rehm[2,13], Andreas Matern[14], Marc Fiume[15], Stephen T. Sherry[16]

Global Alliance for Genomics & Health

Consent Codes		
Name	**Abbreviation**	**Description**
Primary Categories (I^{Ty})		
no restrictions	NRES	No restrictions on data use.
general research use and clinical care	GRU(CC)	For health/medical/biomedical purposes and other biological research, including the study of population origins or ancestry.
health/medical/biomedical research and clinical care	HMB(CC)	Use of the data is limited to health/medical/biomedical purposes, does not include the study of population origins or ancestry.
disease-specific research and clinical care	DS-[XX](CC)	Use of the data must be related to [disease].
population origins/ancestry research	POA	Use of the data is limited to the study of population origins or ancestry.
Secondary Categories (II^{Ty}) (can be one or more extra conditions, in addition to I^{Ty} category)		
other research-specific restrictions	RS-[XX]	Use of the data is limited to studies of [research type] (e.g., pediatric research).
research use only	RUO	Use of data is limited to research purposes (e.g., does not include its use in clinical care).
no "general methods" research	NMDS	Use of the data includes methods development research (e.g., development of software or algorithms) ONLY within the bounds of other data use limitations.
genetic studies only	GSO	Use of the data is limited to genetic studies only (i.e., no research using only the phenotype data).
Requirements		
not-for-profit use only	NPU	Use of the data is limited to not-for-profit organizations.
publication required	PUB	Requestor agrees to make results of studies using the data available to the larger scientific community.
collaboration required	COL-[XX]	Requestor must agree to collaboration with the primary study investigator(s).
return data to database/resource	RTN	Requestor must return derived/enriched data to the database/resource.
ethics approval required	IRB	Requestor must provide documentation of local IRB/REC approval.
geographical restrictions	GS-[XX]	Use of the data is limited to within [geographic region].
publication moratorium/embargo	MOR-[XX]	Requestor agrees not to publish results of studies until [date].
time limits on use	TS-[XX]	Use of data is approved for [x months].
user-specific restrictions	US	Use of data is limited to use by approved users.
project-specific restrictions	PS	Use of data is limited to use within an approved project.
institution-specific restrictions	IS	Use of data is limited to use within an approved institution.

SOM Dyke, *et al*. Consent Codes: Upholding Standard Data Use Conditions. *PLoS Genetics* 12(1): e1005772.
http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005772

Contact: Dr. Stephanie Dyke (stephanie.dyke@mcgill.ca)

Health Related Data & Privacy

- Is the genetic condition **outwardly visible**?
- How **severe** is it? (serious disease, **penetrance**, age of onset)
- Is it associated with what could be considered to be **stigmatizing** health information (e.g., associated with ***mental*** health, ***reproductive*** care, ***disability***)?
- Is it **familial** (i.e., potential carrier status/ reproductive implications for family/relatives)?
- Does it provide information about the likely **geographical location** of individuals?
- Does it provide information about **ethnicity** that may be considered potentially stigmatizing information?

Sharing health-related data: a privacy test?

Stephanie OM Dyke[1], Edward S Dove[2] and Bartha M Knoppers[1]

Greater sharing of potentially sensitive data raises important ethical, legal and social issues (ELSI), which risk hindering and even preventing useful data sharing if not properly addressed. One such important issue is respecting the privacy-related interests of individuals whose data are used in genomic research and clinical care. As part of the Global Alliance for Genomics and Health (GA4GH), we examined the ELSI status of health-related data that are typically considered 'sensitive' in international policy and data protection laws. We propose that 'tiered protection' of such data could be implemented in contexts such as that of the GA4GH Beacon Project to facilitate responsible data sharing. To this end, we discuss a Data Sharing Privacy Test developed to distinguish degrees of sensitivity within categories of data recognised as 'sensitive'. Based on this, we propose guidance for determining the level of protection when sharing genomic and health-related data for the Beacon Project and in other international data sharing initiatives.

npj Genomic Medicine (2016) **1,** 16024; doi:10.1038/npjgenmed.2016.24; published online 17 August 2016

Figure 1. The three steps of a Data Sharing Privacy Test to distinguish degrees of data sensitivity within categories of data recognised as 'sensitive'.

Share (your) Genome data?

- The Beacon concept - balanced approach for accessing genome variant data from internationally distributed resources
- However: Genome data has the inherent “risk” of being identified and linked to a person

Solutions from Technology or Society? Discourse!

Share (your) Genome data?

- The Beacon concept - balanced approach for accessing genome variant data from internationally distributed resources
- However: Genome data has the inherent “risk” of being identified and linked to a person

Solutions from Technology or Society? Discourse!

THE Sun | BIZ | NEWS | FABULOUS | MONEY | MOTORS | TRAVEL | TECH | DEAR

SAFE SPACE **Facebook data breach explained – what happened with Cambridge Analytica and are you safe?**

If you're completely baffled by the latest Facebook here's a very simple guide explaining what's gone

REVEALED By Sean Keach, Digital Technology and Science Editor
21st March 2018, 11:40 am | Updated: 21st March 2018, 11:52 am

DPPH

Data Protection in Personalized Health

Personalized Medicine, Personalized Health Research Project funded by the Strategic Focus Area Personalized Health and Related Technologies (PHRT) of the ETH Board.

By Richard Harris | Satellite imagery by U.S.D.A. N.A.I.P.

Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret

Dozens of companies use smartphone locations to help advertisers and even hedge funds. They say it's anonymous, but the data shows how personal it is.

JENNIFER VALENTINO-DeVRIES, NATASHA SINGER, MICHAEL H. KELLER and AARON KROLIK
The New York Times, DEC. 10, 2018

```
"geo_data" : {
        "geo_label" : "Zurich, Switzerland, Europe",
        "geo_json" : {
                "type" : "Point",
                "coordinates" : [
                        8.55,
                        47.37
                ]
        },
        "geo_precision" : "city",
        "info" : {
                "continent" : "Europe",
                "country" : "Switzerland",
                "city" : "Zurich"
        }
},
```

A visit to a doctor's office is also included. The data is so specific that The Times could determine how long she was there.

At least 75 companies receive anonymous, precise location data from apps whose users enable location services to get local news and weather or other information, The Times found. Several of those businesses claim to track up to 200 million mobile devices in the United States — about half those in use last year. The database reviewed by The Times — a sample of information gathered in 2017 and held by one company — reveals people's travels in startling detail, accurate to within a few yards and in some cases updated more than 14,000 times a day.

More than 1,000 popular apps contain location-sharing code from such companies, according to 2018 data from MightySignal, a mobile analysis firm. Google's Android system was found to have about 1,200 apps with such code, compared with about 200 on Apple's iOS.

JENNIFER VALENTINO-DeVRIES, NATASHA SINGER, MICHAEL H. KELLER and AARON KROLIK
The New York Times, DEC. 10, 2018

Generalkonsent

BENEFIT

BLOCKCHAIN

HEALTH

PRIVACY

SECURITY

CONSENT

ACCESS

Right to Research

HACKERS

LAWS

Genetic
Information
Nondiscrimination
Act

Health
Insurance
Portability and
Accountability
Act

SAFETY

CRYPTOGRAPHY

John Yuyi, NYT 2018-02-09

BIO390: Course Schedule

- 2018-09-18: Michael Baudis - **What is Bioinformatics? Introduction and Resources**
- 2018-09-25: Christian von Mering - **Sequence Informatics**
- 2018-10-02: Mark Robinson - **Statistical Bioinformatics**
- 2018-10-09: Katja Bärenfaller - **Proteomics**
- 2018-10-16: Izaskun Mallona - **Regulatory Genomics and Epigenomics**
- 2018-10-23: Daniel Stekhoven - **Clinical Bioinformatics**
- 2018-10-30: Shinichi Sunagawa - **Metagenomics**
- 2018-11-06: Amedeo Caflisch - **Molecular Dynamics**
- 2018-11-13: Elif Ozkrimli Olmez - **Protein Interactions**
- ~~2018-11-20: Fabio Rinaldi - **Text Mining I**~~
- 2018-11-27: Fabio Rinaldi - **Text Mining II**
- 2018-12-04: Andreas Wagner - **Biological Networks**
- 2018-12-11: Michael Baudis - **Human Genome Data - Risks & Opportunities || Feedback**
- 2018-12-18: Michael Baudis - **Exam (multiple choice)**

BIO390 Exam

- 8:10 - 9:40
- UZH ID
- Multiple choice questions (single or multiple answers; indicated)

22 **Mark ALL true statements: Data from which of the following method(s) would you use to infer *direct* protein-protein interactions?**

A	AP-MS	☐
B	BioID-MS	☐
C	APEX	☐
D	*in vitro* reconstitution	☐
E	size exclusion chromatography	☐
F	Yeast two hybrid screen	☐

MG — multiple options

12 **Tidy data is a way of arranging data where each variable is in one column and each observation is in one row. The table below shows some data which has been imported into R as a tibble data.frame. Is this data in tidy format?**

```
## # A tibble: 10 x 5
##     control treatment  time    sex individual
##       <dbl>     <dbl> <chr> <fctr>     <fctr>
##  1 3.711482  4.476728   day female          1
##  2 4.652311  3.680602   day   male          2
##  3 4.478371  4.832210 night     NA       <NA>
##  4 6.273473  5.440702 night   male          2
##  5 6.824521  6.380166 night female          1
##  6 3.488692  5.768513   day female          3
##  7 5.110508  5.480421   day female          4
##  8 4.239204  4.975053   day   male          5
##  9 4.330103  4.090669 night   male          5
## 10 5.274520  3.666011 night   male          6
```

A	No, because there are missing values.	☐
B	No, because the columns “control” and “treatment” are observations of the same variable and for tidy format should be in the same column.	☐
C	No, because the column “time” has type “character” but should be categorial (i.e. with type “factor”).	☐
D	Yes, because “control” and “treatment” should be in separate columns as they are contain the values that will be compared.	☐

HL — 1 option

University of Zurich[UZH]

SIB

Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB | Swiss Institute of Bioinformatics
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

arraymap.org

progenetix.org

info.baudisgroup.org

sib.swiss/baudis-michael

imls.uzh.ch/en/research/baudis

Global Alliance
for Genomics & Health

elixir