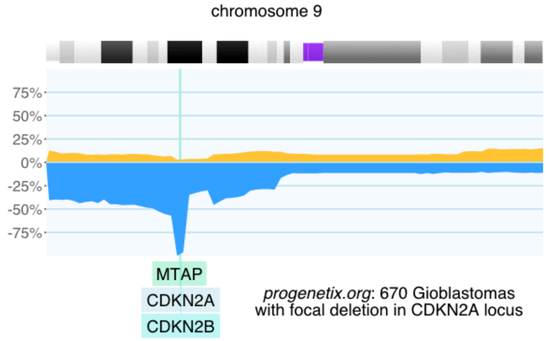
Progenetix & GA4GH

A cancer genomics resource built around and driving GA4GH standards

Michael Baudis | UZH BIO392 HS21 | 2021-10-12

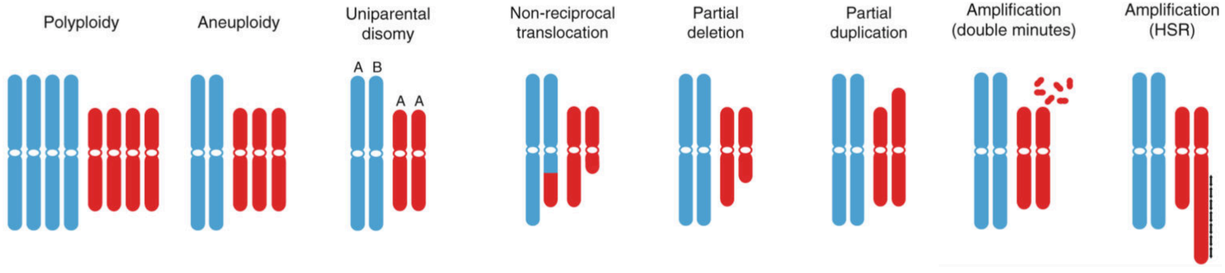
progenetix

Theoretical Cytogenetics and Oncogenomics Research | Methods | Standards

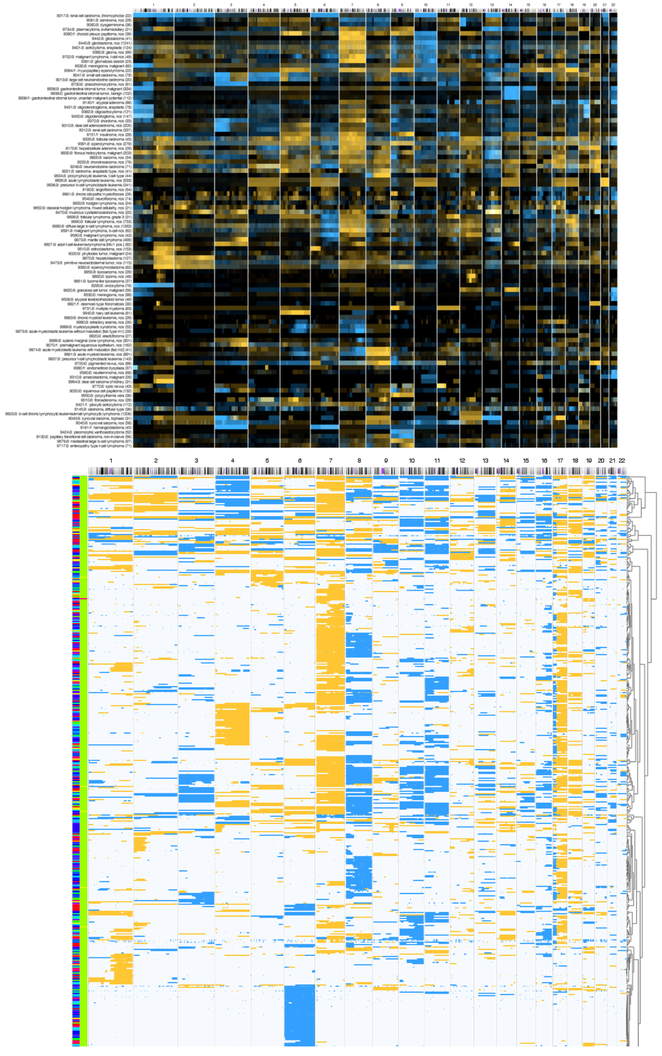


Genomic Imbalances in Cancer - Copy Number Variations (CNV)

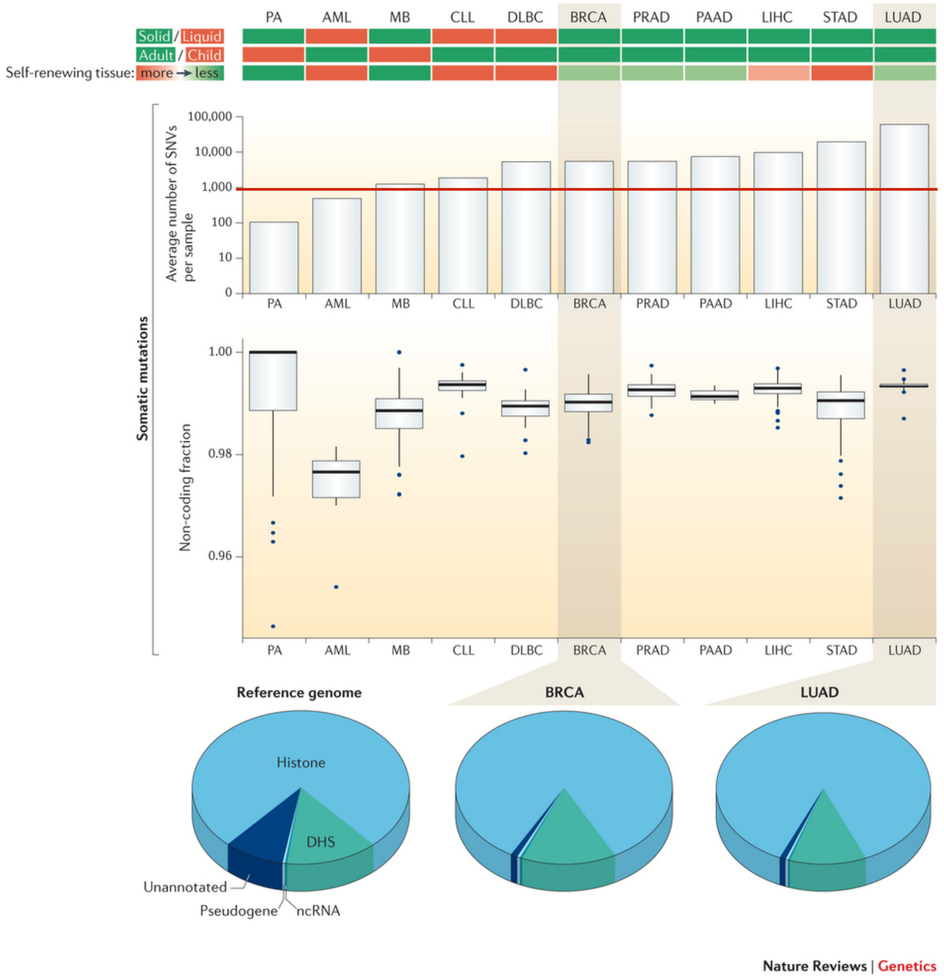
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



Grade et al., 2015 Recent Results Cancer Res

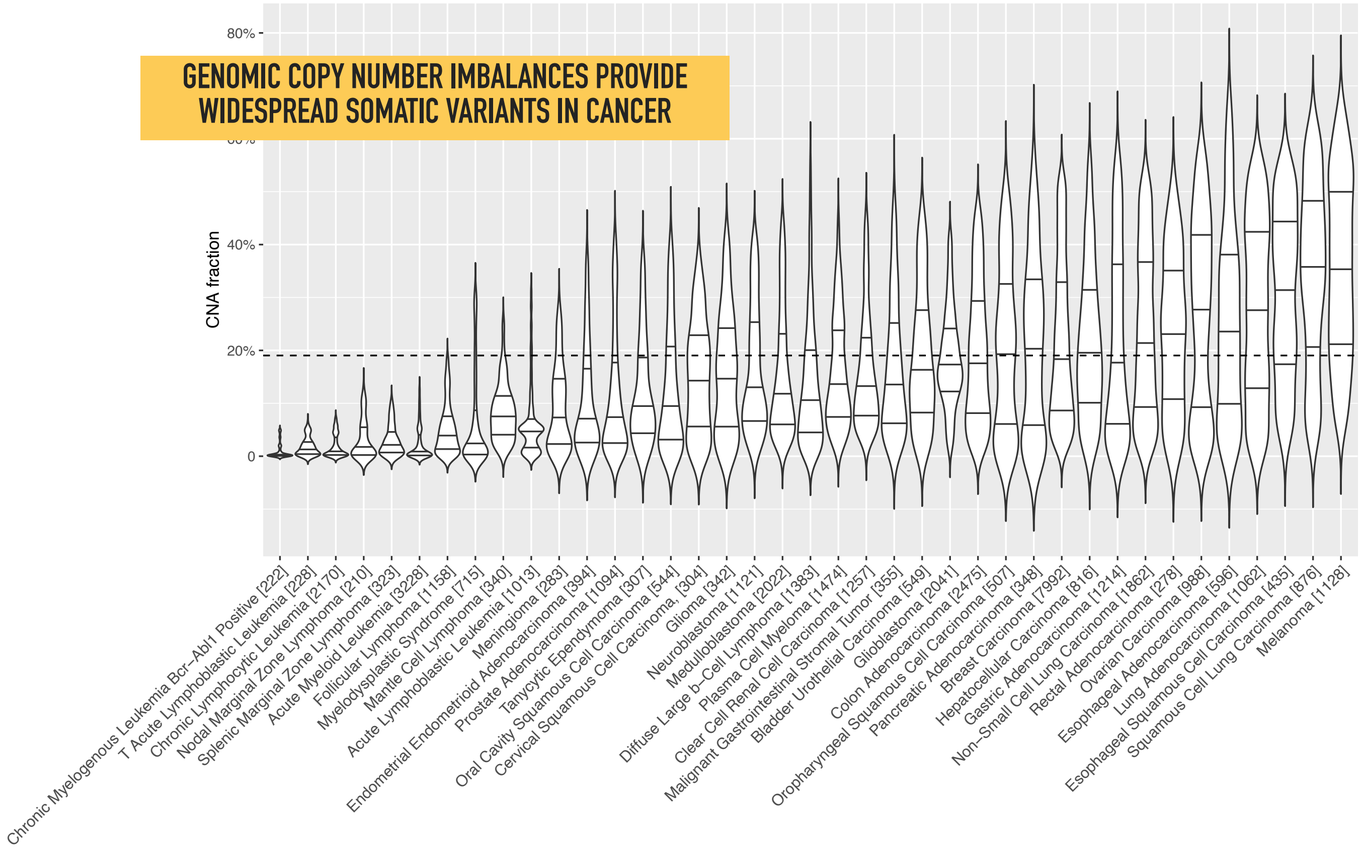


Quantifying Somatic Mutations In Cancer



CANCERS SHOW THOUSANDS OF SINGLE NUCLEOTIDE VARIANTS PER SAMPLE, MOSTLY IN NON-CODING REGIONS

Pan-Cancer Analysis of Whole Genomes (PCAWG) data show widespread mutations in non-coding regions of cancer genomes (Khurana et al., Nat. Rev. Genet. (2016)

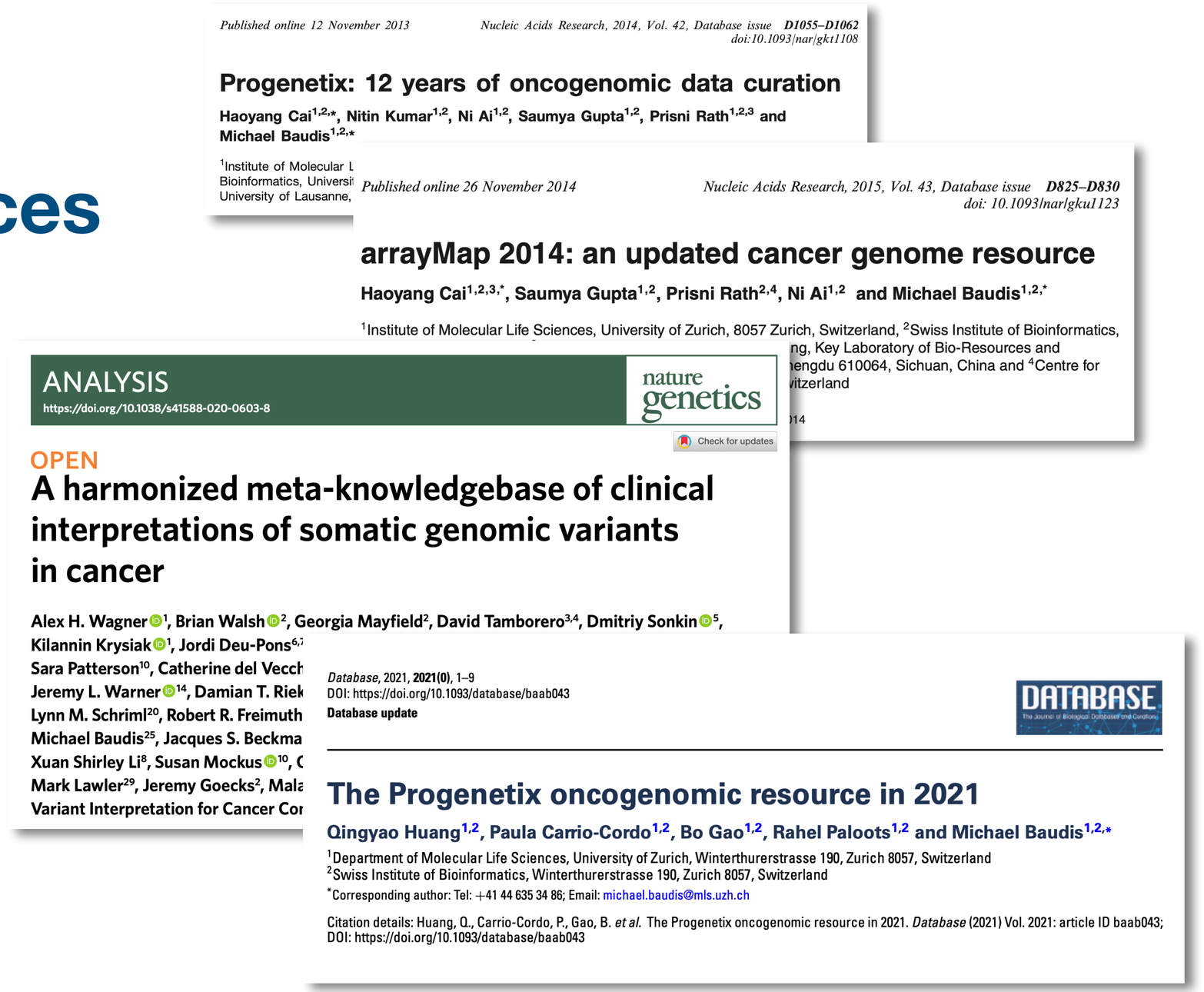


On average ~19% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on 43654 cancer genomes from progenetix.org

Recent Publications

Genomics Resources

- data resource publications describing content and technology updates of our own "resource ecosystem"
- contributions to international / large-scale data resources and federation efforts



Published online 12 November 2013 Nucleic Acids Research, 2014, Vol. 42, Database issue D1055–D1062
doi:10.1093/nar/gkt1108

Progenetix: 12 years of oncogenomic data curation

Haoyang Cai[1,2,*], Nitin Kumar[1,2], Ni Ai[1,2], Saumya Gupta[1,2], Prisni Rath[1,2,3] and Michael Baudis[1,2,*]

[1]Institute of Molecular L
Bioinformatics, Universit
University of Lausanne,

Published online 26 November 2014 Nucleic Acids Research, 2015, Vol. 43, Database issue D825–D830
doi: 10.1093/nar/gku1123

arrayMap 2014: an updated cancer genome resource

Haoyang Cai[1,2,3,*], Saumya Gupta[1,2], Prisni Rath[2,4], Ni Ai[1,2] and Michael Baudis[1,2,*]

[1]Institute of Molecular Life Sciences, University of Zurich, 8057 Zurich, Switzerland, [2]Swiss Institute of Bioinformatics,
ng, Key Laboratory of Bio-Resources and
engdu 610064, Sichuan, China and [4]Centre for
itzerland

ANALYSIS
https://doi.org/10.1038/s41588-020-0603-8
nature genetics
Check for updates

OPEN

A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer

Alex H. Wagner[1], Brian Walsh[2], Georgia Mayfield[2], David Tamborero[3,4], Dmitriy Sonkin[5],
Kilannin Krysiak[1], Jordi Deu-Pons[6,7]
Sara Patterson[10], Catherine del Vecch
Jeremy L. Warner[14], Damian T. Riek
Lynn M. Schriml[20], Robert R. Freimuth
Michael Baudis[25], Jacques S. Beckma
Xuan Shirley Li[8], Susan Mockus[10], C
Mark Lawler[29], Jeremy Goecks[2], Mala
Variant Interpretation for Cancer Cor

Database, 2021, **2021(0)**, 1–9
DOI: https://doi.org/10.1093/database/baab043
Database update
DATABASE
The Journal of Biological Databases and Curation

The Progenetix oncogenomic resource in 2021

Qingyao Huang[1,2], Paula Carrio-Cordo[1,2], Bo Gao[1,2], Rahel Paloots[1,2] and Michael Baudis[1,2,*]

[1]Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland
[2]Swiss Institute of Bioinformatics, Winterthurerstrasse 190, Zurich 8057, Switzerland
*Corresponding author: Tel: +41 44 635 34 86; Email: michael.baudis@mls.uzh.ch

Citation details: Huang, Q., Carrio-Cordo, P., Gao, B. *et al.* The Progenetix oncogenomic resource in 2021. *Database* (2021) Vol. 2021: article ID baab043; DOI: https://doi.org/10.1093/database/baab043

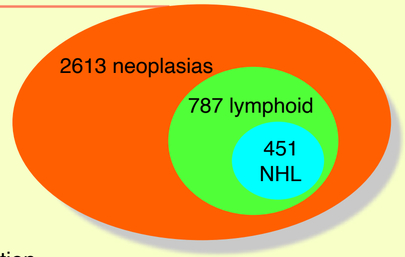
progenetix.net

[progenetix.net] online CGH database

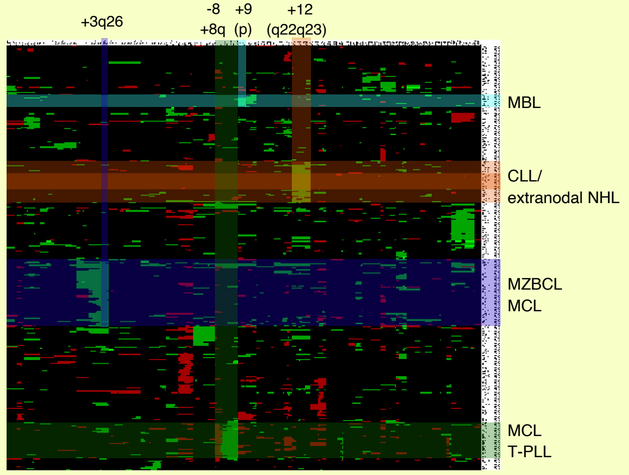
A compilation of published CGH analyses with reported case and band specific results

Currently includes
2613 cases from
92 publications

Automatic
conversion of
ISCN format to
aberration matrix
with 393 bands resolution



Clustering in 451 NHL: +12 in CLL and extranodal NHL



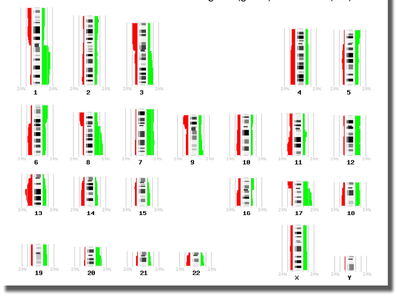
progenetix.net

Collection and Transformation of Chromosomal Imbalances in Human Neoplasias for Data Mining Procedures

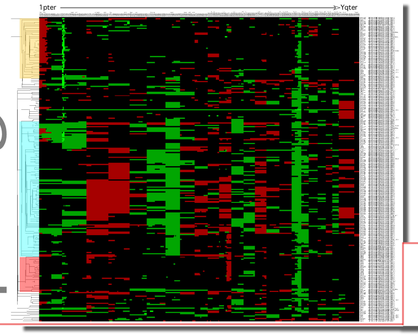
michael baudis, dept. of pathology, stanford university

Although the deciphering of the human genome has been pushed forward over the last years, little effort has been made to collect and integrate the treasure trove of clinical tumor cases analyzed by molecular-cytogenetic methods into current data schemes. Publicly announced at BCATS 2001, since then [progenetix.net] has been established as the largest public source of chromosomal imbalance data with band-specific resolution. Targets for the use of the data collection may be the description of prediction of oncogene and suppressor gene loci, identification of related loci for pathway creation, and especially the combination of the data with expression array experiments for filtering of relevant genes among the deregulated candidates.

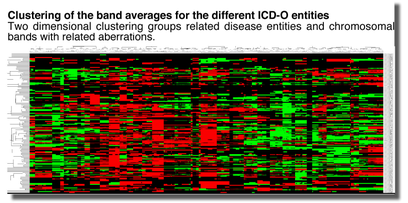
Chromosomal imbalances in 5478 clinical cases from 196 publications
Although not as prominent as in specific subgroups, this large collection shows the non-random distribution of chromosmal gains (green) and losses (red).



Material and Methods Chromosomal aberration data of more than 5478 cases from 196 publications describing results of Comparative Genomic Hybridization (CGH) experiments were collected. Minimal requirements were diagnosis of a malignant or benign neoplasia, analysis of clinical tumor samples and report of the analysis results on a case by case basis, resolved to the level of single chromosomal bands.Data was transformed from the diverse annotation formats to standardized ISCN “rev ish” nomenclature. For the transformation of the non-linear ISCN data to a two-dimensional matrix with code for the aberration status of each chromosomal band per case, a reverse pattern matching algorithm was developed in Perl. Graphical representations and cluster images are generated for all different subsets (Publications, ICD-O-3 entities, meta-groups) and presented on the progenetix.net website.

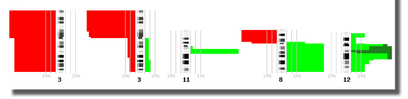


Clustering of the band averages for the different ICD-O entities
Two dimensional clustering groups related disease entities and chromosomal bands with related aberrations.



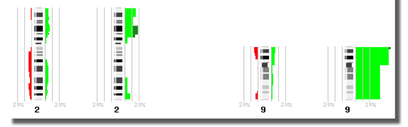
Results Out of 4896 tumor samples, 3862 (79%) showed chromosomal imbalances by CGH. The average per band probability was 4.5% for a loss (max. 12.9% at 13q21) and 6.5% for a gain (max. 15.6% at 8q23). Differences between neoplastic entities showed in the average frequency and distribution pattern of imbalanced chromosomal regions. Tumor subsets (10 or more cases) with the strongest hot spots for losses were small cell lung carcinomas (ave. 23.3% with max. 96.2% at 3p14p26) and pheochromocytomas (ave. 10.9% with max. 92.7% at 3p); prominent gain maxima were found in pure high grade infiltrating duct carcinomas of the breast (ave. 5.9% with max. 95.7% at 11q13), T-PLL (ave. 4.7% with max. 81.8% for whole 8q) and dedifferentiated liposarcomas (ave. 10.4% with max. 81.8% at 12q13), among others.
By cluster analysis, different combinations of chromosomal hot spot regions could be shown to occur in tumors subsummized in the same diagnostic entity; the example of neuroblastomas is shown.

Examples of hotspots of genomic imbalance
SCLC, pheochromocytoma, high grade DCIS, T-PLL, dedifferentiated liposarcoma



Conclusion So far, progenetix.net project was able to:
1. collect a large dataset of genomic aberration data generated through a molecular-cytogenetic screening technique (CGH)
2. develop the software tools to transform those data to a meta format compatible to commonly used genomic interval descriptions
3. produce graphical and numerical output from those data for hot spot detection and statistical analysis.
For future approaches, the data collection will be valuable for filtering data from expression array experiments for relevant genes, and possibly for the description of common and divergent genetic pathways in the oncogenetic process of different tumor entities.The transformed raw data of the progenetix.net collection is avaible for research purposes over the website.

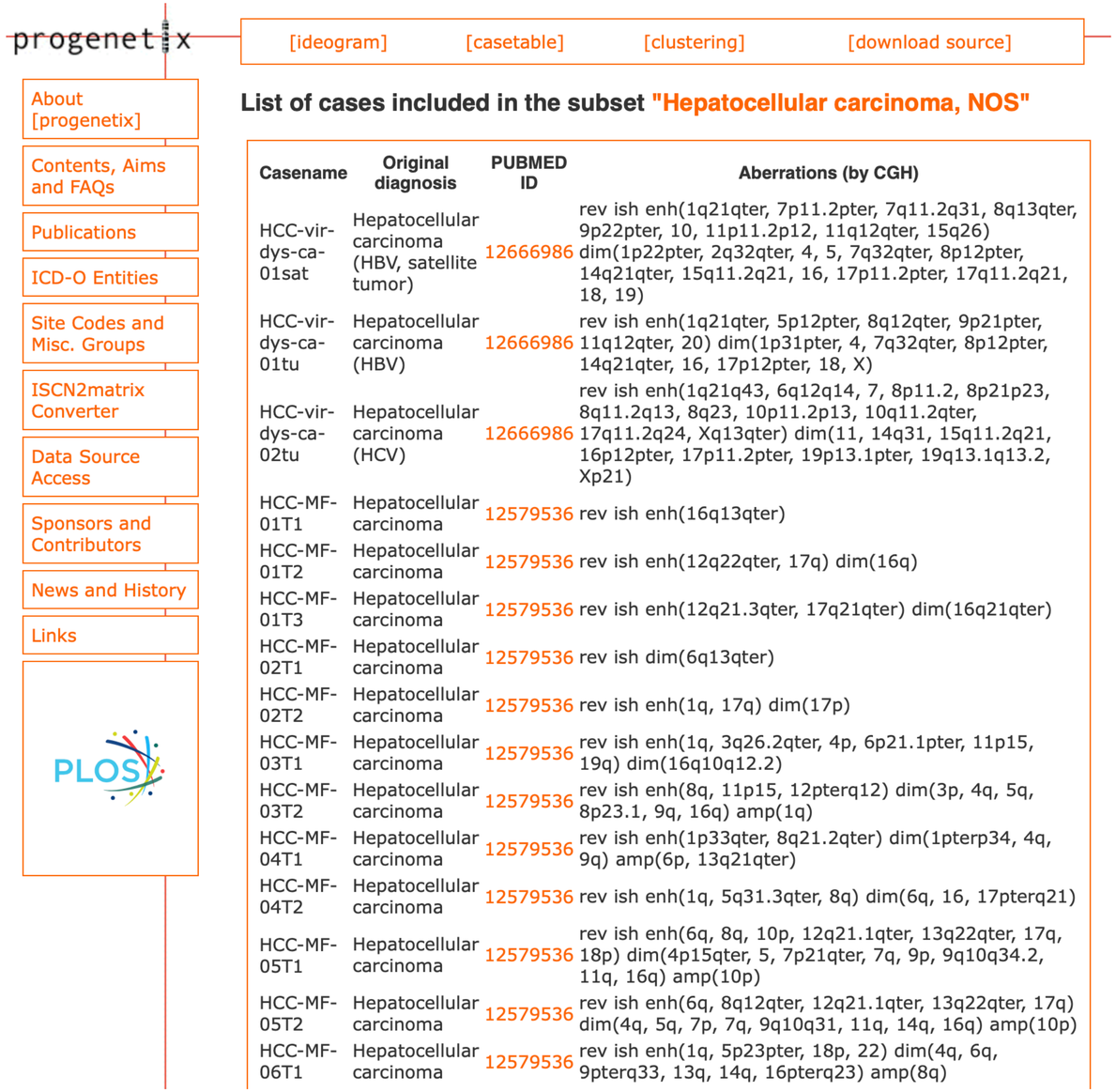
Distinction of histologically related through their chromosomal aberration pattern
Amplification of the REL locus on 2p16 and gain of 9p(ter) distinguishes primary mediastinal B-cell lymphomas (PMBL, right) from diffuse large cell lymphomas (DLCL, left). The distinction may have clinical implications



Identification of different aberration patterns in Neuroblastoma (289 cases)
N-Myc (2p25) amplification is the hallmark of a subgroup, showing only consistant loss of the terminal portion 1p. Other groups are defined by the loss of parts of 11q, or a “chromosomal instability” phenotype. Gains on 17q are a common feature of all groups.
Those patterns may be combined with gene-level information to reconstruct the different pathways leading to malignant transformation.

Progenetix Database in 2003
Text conversion for CNVs

- based on listed CGH results from publications
  - literature detection using optimized PubMed queries
  - extraction (copy/paste, typing) of rev ish ISCN karyotypes from articles and supplementary material
  - annotation cleanup using scripting with regular expressions (Perl)
  - custom script to convert cleaned ISCN annotations to cytoband status maps
  - custom graphics libraries to create graphical representations of CNV frequencies



Progenetix in 2021

Cancer Genomics Reference Resource

- largest open resource for curated cancer genome profiling data, with focus on copy number variations (CNV)
- >116'000 cancer CNV profiles, mapped to >800 NCIt codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCIt, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services

progenetix

progenetix

Cancer CNV Profiles

Search Samples

Studies & Cohorts

- arrayMap
- TCGA Samples
- DIPG Samples
- Gao & Baudis, 2021
- Cancer Cell Lines

Publication DB

Services

- NCIt Mappings
- UBERON Mappings

Upload & Plot

Download Data

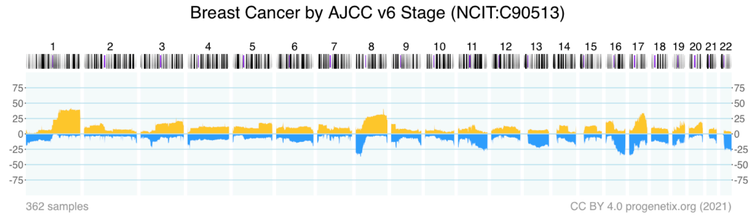
Beacon+

Progenetix Info

- About Progenetix
- Use Cases
- Documentation
- Baudisgroup @ UZH

Cancer genome data @ progenetix.org

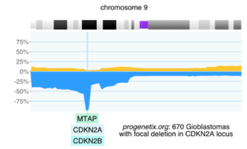
The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **139448** samples.



Download SVG | Go to NCIT:C90513 | Download CNV Frequencies

Example for aggregated CNV data in 362 samples in Breast Cancer by AJCC v6 Stage.
Here the frequency of regional **copy number gains** and **losses** are displayed for all 22 autosomes.

Progenetix Use Cases



Local CNV Frequencies

A typical use case on Progenetix is the search for local copy number aberrations - e.g. involving a gene - and the exploration of cancer types with these CNVs. The [Search Page] provides example use cases for designing queries. Results contain basic statistics as well as visualization and download options.

Cancer CNV Profiles

The progenetix resource contains data of **810** different cancer types (NCIt neoplasm classification), mapped to a variety of biological and technical categories. Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the [Cancer Types] page with direct visualization and options for sample retrieval and plotting options.

Cancer Genomics Publications

Through the [Publications] page Progenetix provides **4025** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

Database, 2021, **2021(0)**, 1–9
DOI: https://doi.org/10.1093/database/baab043
Database update

The Progenetix oncogenomic resource in 2021

Qingyao Huang[1,2], Paula Carrio-Cordo[1,2], Bo Gao[1,2], Rahel Paloots[1,2] and Michael Baudis[1,2,*]

[1]Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland
[2]Swiss Institute of Bioinformatics, Winterthurerstrasse 190, Zurich 8057, Switzerland

*Corresponding author: Tel: +41 44 635 34 86; Email: michael.baudis@mls.uzh.ch

Citation details: Huang, Q., Carrio-Cordo, P., Gao, B. *et al.* The Progenetix oncogenomic resource in 2021. *Database* (2021) Vol. 2021: article ID baab043; DOI: https://doi.org/10.1093/database/baab043

Abstract

In cancer, copy number aberrations (CNAs) represent a type of nearly ubiquitous and frequently extensive structural genome variations. To disentangle the molecular mechanisms underlying tumorigenesis as well as identify and characterize molecular subtypes, the comparative and meta-analysis of large genomic variant collections can be of immense importance. Over the last decades, cancer genomic profiling projects have resulted in a large amount of somatic genome variation profiles, however segregated in a multitude of individual studies and datasets. The Progenetix project, initiated in 2001, curates individual cancer CNA profiles and associated metadata from published oncogenomic studies and data repositories with the aim to empower integrative analyses spanning all different cancer biologies. During the last few years, the fields of genomics and cancer research have seen significant advancement in terms of molecular genetics technology, disease concepts, data standard harmonization as well as data availability, in an increasingly structured and systematic manner. For the Progenetix resource, continuous data integration, curation and maintenance have resulted in the most comprehensive representation of cancer genome CNA profiling data with 138 663 (including 115 357 tumor) copy number variation (CNV) profiles. In this article, we report a 4.5-fold increase in sample number since 2013, improvements in data quality, ontology representation with a CNV landscape summary over 51 distinctive National Cancer Institute Thesaurus cancer terms as well as updates in database schemas, and data access including new web front-end and programmatic data access.

Database URL: progenetix.org

Table 1. Statistics of samples from various data resources

Data source	GEO	ArrayExpress	cBioPortal	TCGA	Total
No. of studies	898	51	38	33	1939
No. of samples	63 568	4351	19 712	22 142	138 663
Tumor	52 090	3887	19 712	11 090	115 357
Normal	11 478	464	0	11 052	23 306
Classifications					
ICD-O (Topography)	100	54	88	157	209
ICD-O (Morphology)	246	908	265	140	491
NCIt	346	148	422	182	788
Collections					
Individuals	63 568	4351	19 712	10 995	127 549
Biosamples	63 568	4351	19 712	22 142	138 663
Callsets[a]	63 568	4351	19 712	22 376	138 930
Variants	5 514 126	118 4170	1 778 096	2 654 065	10 716 093

[a]set of variants from one genotyping experiment; ICD-O, International Classification of Diseases for Oncology; NCIt, National Cancer Institute Thesaurus.

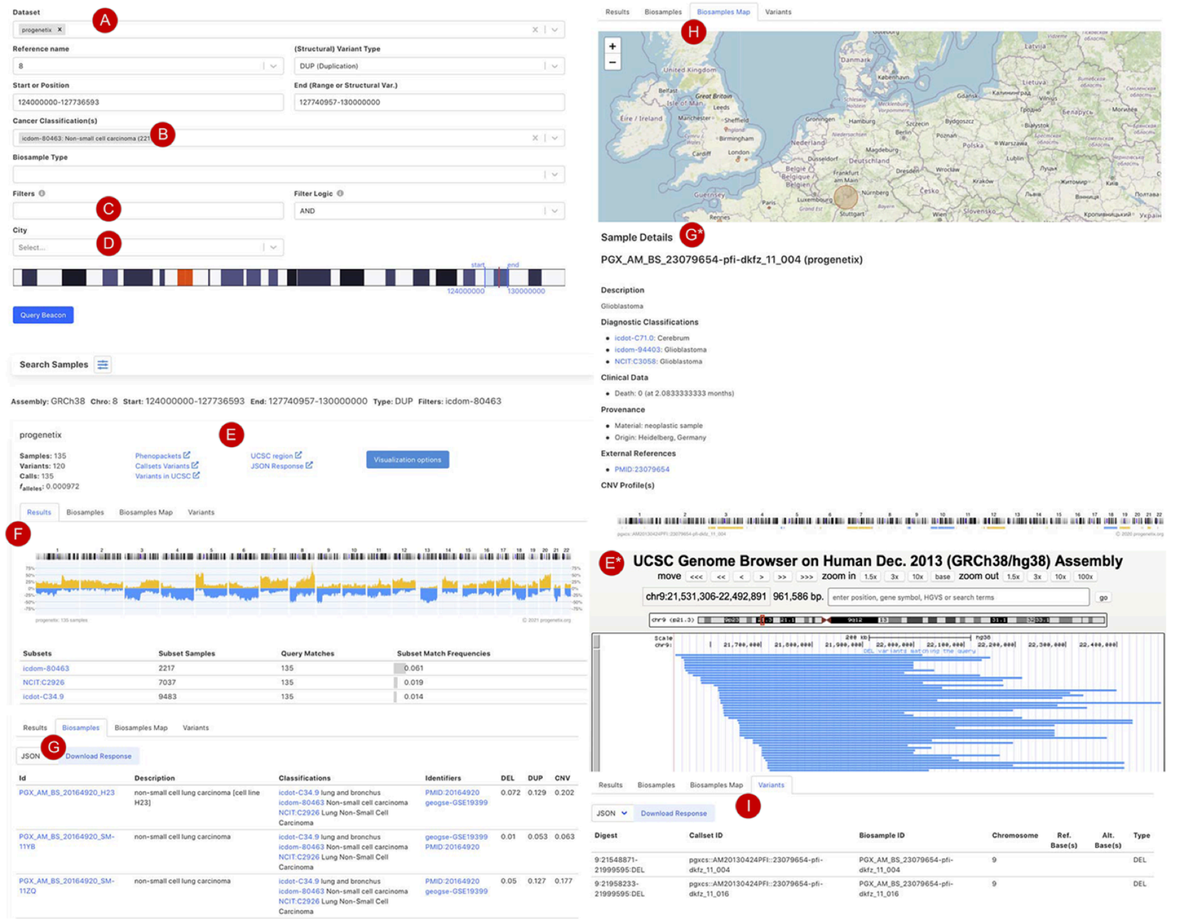
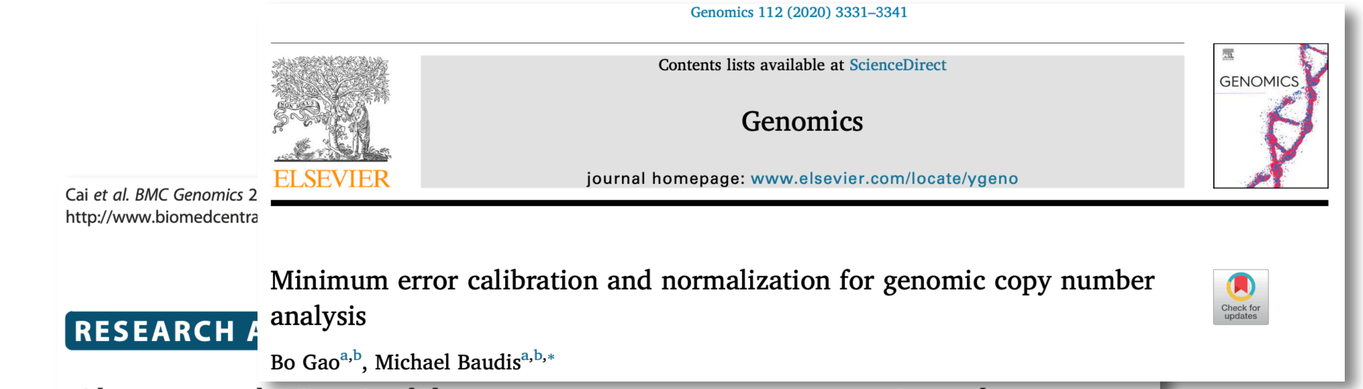


Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with < = 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and ICD-O-3 ontology terms available); C: additional filter, e.g. Cellosaurus; D: additional filter for geographic location; E: external link to UCSC browser to view the alignment of matched variants; F: cancer type classification sorted by frequency of the matched biosamples present in the subset; G: list of matched biosamples with description, statistics and reference. More detailed biosample information can be viewed through 'id' link to the sample detail page; H: matched variants with reference to biosamples can be downloaded in json or csv format.

Recent Publications
CNV Data
Analyses & Methods

- method development with focus on cross-platform analysis of CNV and "omics" data
- CNV patterns and signatures across cancer types



frontiers in Genetics

ORIGINAL RESEARCH
published: 13 May 2021
doi: 10.3389/fgene.2021.654887

Signatures of Discriminative Copy Number Aberrations in 31 Cancer Subtypes

Bo Gao [1,2] *and Michael Baudis* [1,2*]

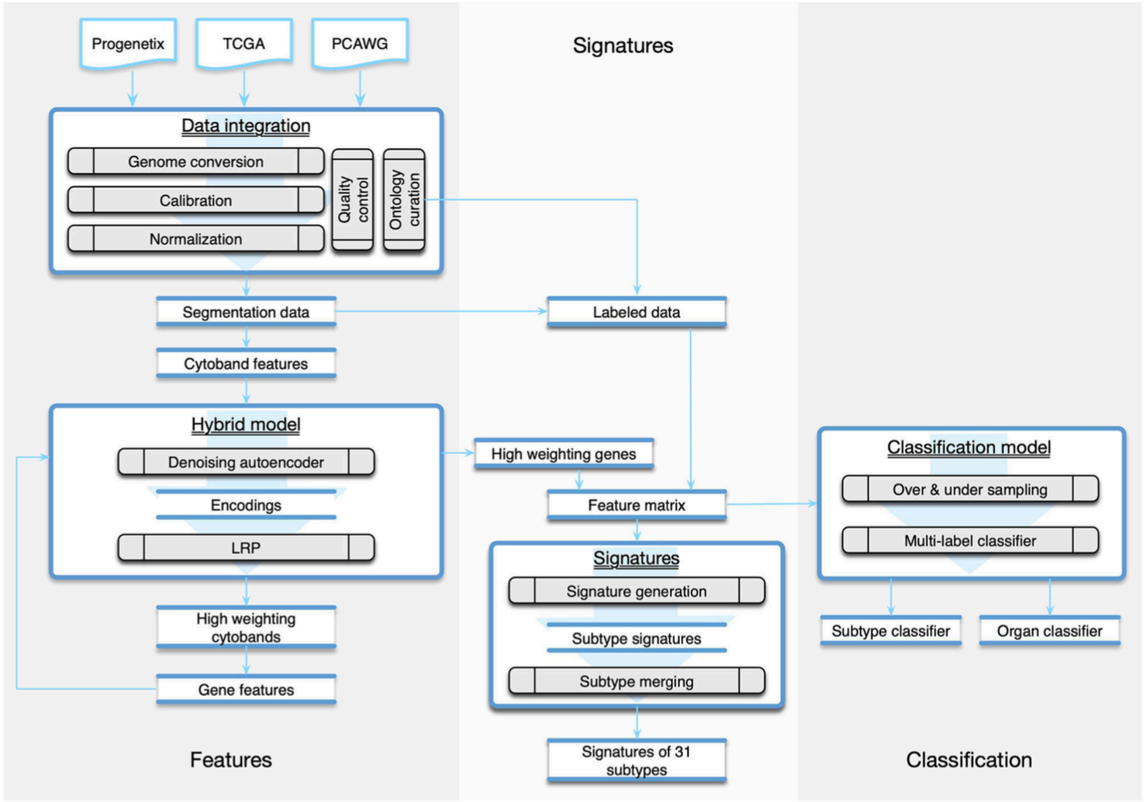


FIGURE 1 | The workflow of the study was composed of three parts. The *Features* part consisted of methods of data integration and feature generation. The *Signature* part focused on creating CNA signatures for cancer subtypes and the categorization of subtypes. The *Classification* part recruited machine learning techniques to predict the organ and the subtype from a given copy number profile.

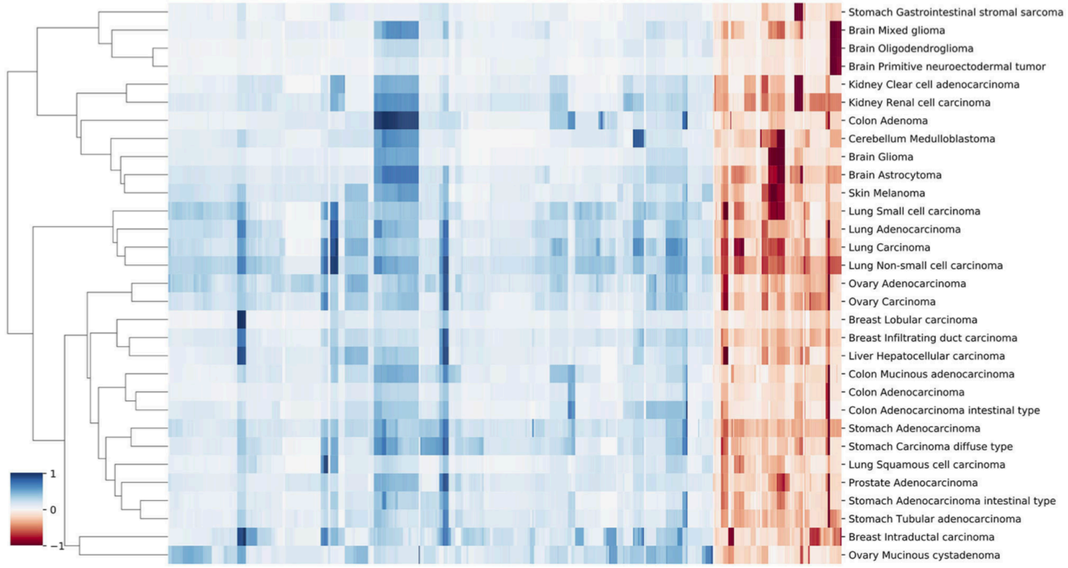


FIGURE 5 | A clustering heatmap of features in 31 signatures. Columns are normalized average CNV intensities of feature genes, where the blue colors are duplication features and red colors are deletion features. Duplication and deletion frequencies are normalized separately.

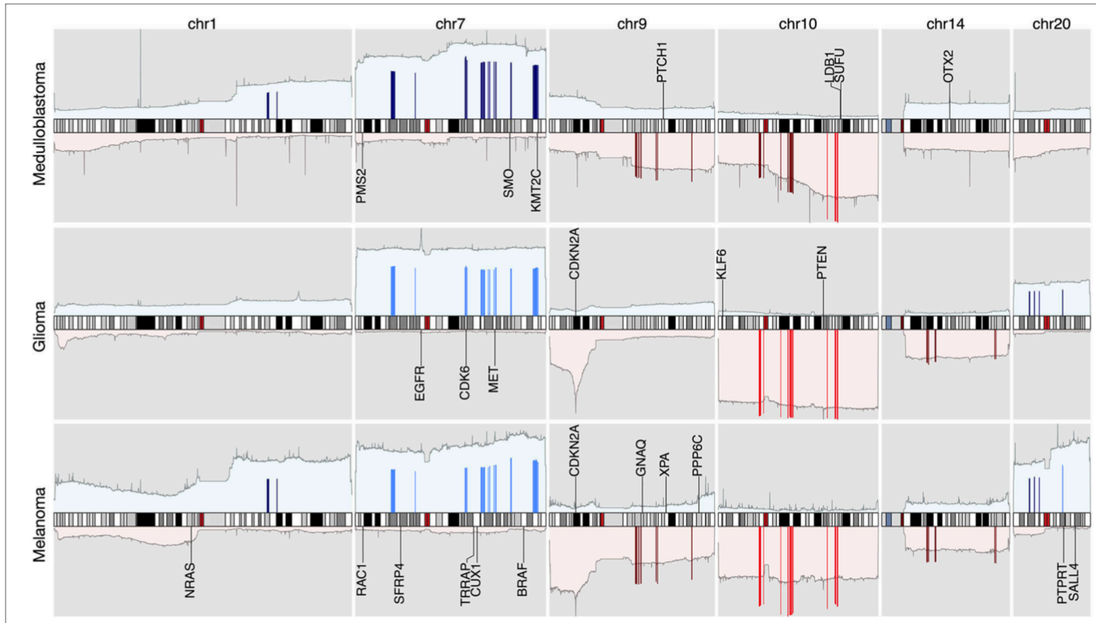
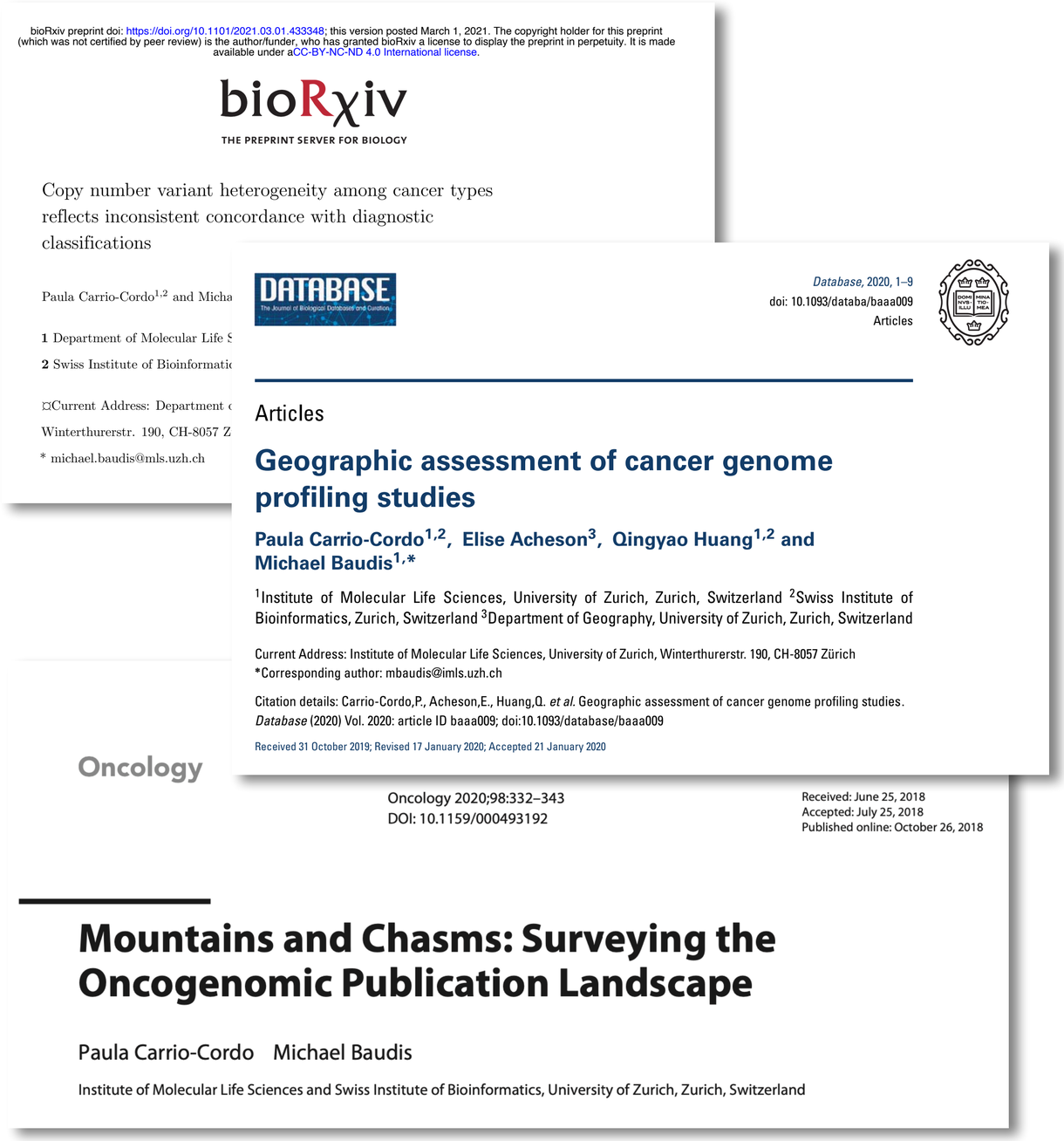


FIGURE 6 | The integrated view of the original data and the selected features, in the neural crest originating entities medulloblastoma, glioma, and melanoma. The shaded background area color illustrates the original data. Color bars illustrate the feature genes, where brighter colors indicate stronger signal intensity. The blue colors above the chromosome axis represent the average amplifications, and the red colors below the chromosome axis represent the average deletions. The amplitude of amplifications and deletions are normalized to [0,1] separately. The adjacent known driver genes are also included for each tumor type.

Recent Publications

Metadata & "Metascience"

- huge Progenetix data collection with systematic annotation of different types of "metadata" allows its utilization e.g. for
  - ➡ "research epistemology", e.g. biases in sample origins or disease representation
  - ➡ conceptual coherence, e.g. how genomic tumor heterogeneity corresponds to precision levels in disease classifications



bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Copy number variant heterogeneity among cancer types reflects inconsistent concordance with diagnostic classifications

Paula Carrio-Cordo[1,2] and Micha

1 Department of Molecular Life S

2 Swiss Institute of Bioinformatic

¤Current Address: Department o

Winterthurerstr. 190, CH-8057 Z

* michael.baudis@mls.uzh.ch

DATABASE
The Journal of Biological Databases and Curation

Database, 2020, 1–9
doi: 10.1093/databa/baaa009
Articles

Articles

Geographic assessment of cancer genome profiling studies

Paula Carrio-Cordo[1,2], Elise Acheson[3], Qingyao Huang[1,2] and Michael Baudis[1,*]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland [2]Swiss Institute of Bioinformatics, Zurich, Switzerland [3]Department of Geography, University of Zurich, Zurich, Switzerland

Current Address: Institute of Molecular Life Sciences, University of Zurich, Winterthurerstr. 190, CH-8057 Zürich
*Corresponding author: mbaudis@imls.uzh.ch

Citation details: Carrio-Cordo,P., Acheson,E., Huang,Q. *et al.* Geographic assessment of cancer genome profiling studies. *Database* (2020) Vol. 2020: article ID baaa009; doi:10.1093/database/baaa009

Received 31 October 2019; Revised 17 January 2020; Accepted 21 January 2020

Oncology

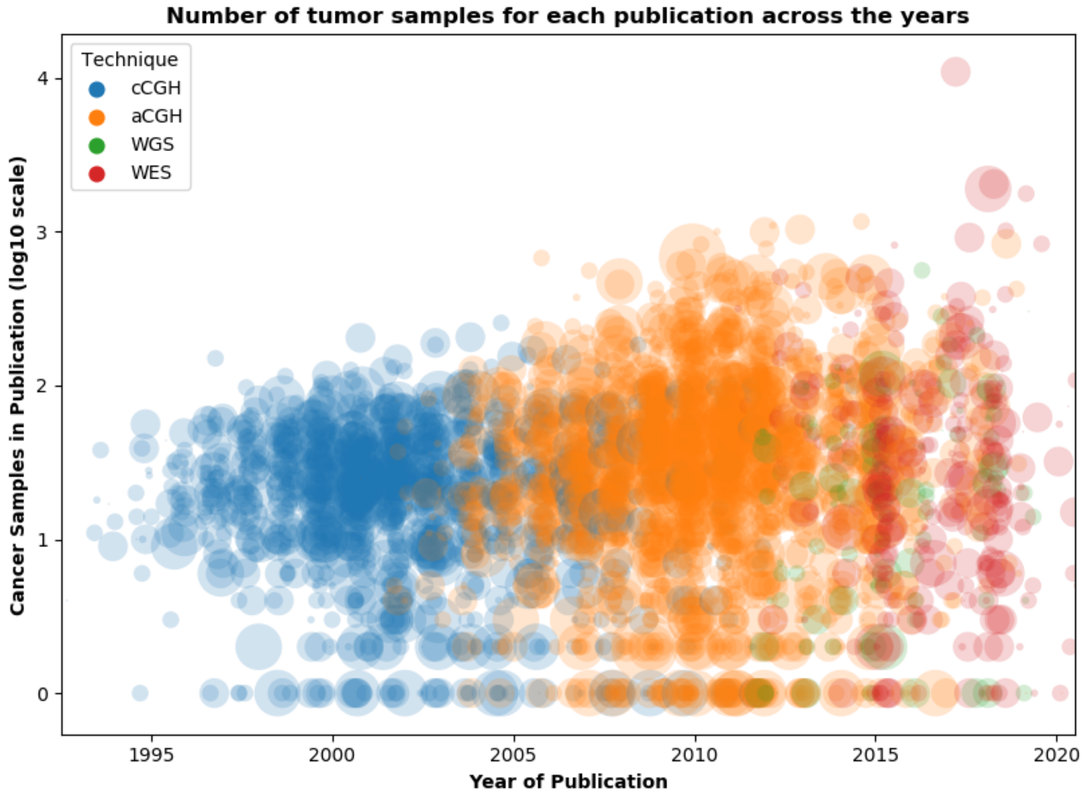
Oncology 2020;98:332–343
DOI: 10.1159/000493192

Received: June 25, 2018
Accepted: July 25, 2018
Published online: October 26, 2018

Mountains and Chasms: Surveying the Oncogenomic Publication Landscape

Paula Carrio-Cordo Michael Baudis

Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland



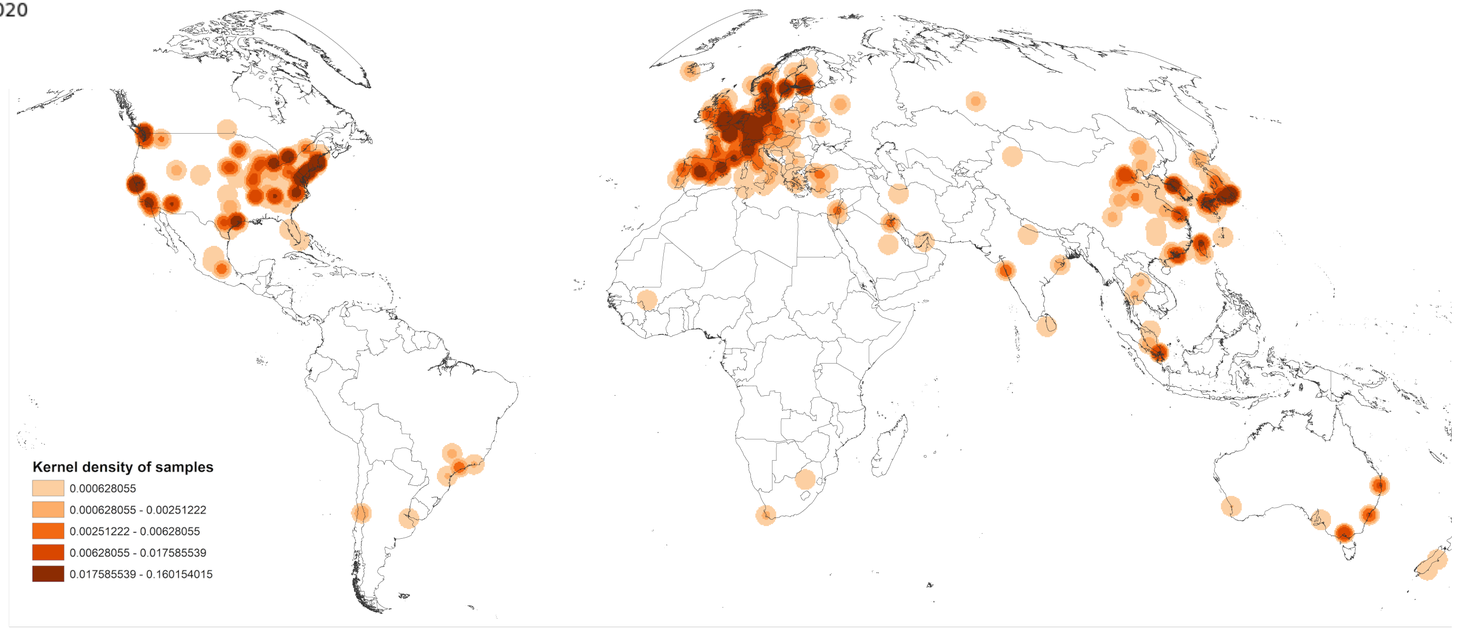
Publication Landscape of Cancer CNV Profiling

Publication statistics for cancer genome screening studies. The graphic shows our as- sessment of publications reporting whole-genome screening of cancer samples, using molecular detection methods (chromosomal CGH, genomic array technologies, whole exome and genome sequencing).

For the years 1993-2018, we found 3'229 publications reporting 174'530 individual samples in single series from 1 to more than 1000 samples. Y-axis and size of the dots correspond to the sample number; the color codes indicate the technology used.

Map of the geographic distribution (by first author affiliation) of the 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets.

The numbers are derived from the 3'240 publications registered in the Progenetix database.



progenetix

Recent Publications

Data Standards and "White Papers"

- over the last years increasing participation in standards, open science, large initiatives and research communities
  - ➡ GA4GH
  - ➡ ELIXIR
  - ➡ SPHN

bioRxiv

The GA4GH Variation Representation Specification (VRS): a Computational Framework for the Precise Representation and Federated Identification of Molecular Variation

Authors

Alex H Wagner[1,2*], Lawrence Babb[3*], Gil Alterovitz[4,5], Michael Baudis[6], Matthew Brush[7], Daniel L Cameron[8,9], Melissa Cline[10], Malachi Griffith[11], Obi L Griffith[11], Sarah Hunt[12], David Kreda[13], Jennifer Lee[14], Javier Lopez[15], Eric Moyer[16], Tristan Nelson[17], Ronak Y Patel[18], Kevin Riehle[18], Peter N Robinson[19], Shawn Rynea… Konopko[21], Heidi Rehm[3,22]

Federated discovery and sharing of genomic data using Beacons

NATURE BIOTECHNOLOGY | VOL 37 | MARCH 2019 | 215–226

European Journal of Human Genetics (2018) 26:1721–1731
https://doi.org/10.1038/s41431-018-0219-y

ESHG

ARTICLE

Registered access: authorizing data access

Stephanie O. M. Dyke[1,35] · Mikael Linden[2,3] · Ilkka Lappalainen[2,3,4] · Jordi Rambla De Argila[5,6] · Knox Carey · David Lloyd[4,7] · J. Dylan Spalding[4] · Moran N. Cabili[8] · Giselle Kerry[4] · Julia Foreman[9] · Tim Cutts[9] · Mahsa Shabani[10] · Laura L. Rodriguez[11] · Maximilian Haeussler[12] · Brian Walsh[13] · Xiaoqian Jiang[14] · Shuang Wang[14] · Daniel Perrett[9] · Tiffany Boughtwood[15] · Andreas Matern[16] · Anthony J. Brookes[17] · Miro Cupak[18] · Marc Fiume[18] · Ravi Pandya[19] · Ilia Tulchinsky[20] · Serena Scollen[3] · Juha Törnroos[2] · Samir Das[21] · Alan C. Evans[21] · Bradley A. Malin[22] · Stephan Beck[23] · Steven E. Brenner[24] · Tommi Nyrönen[2,25] · Niklas Blomberg[3] · Helen V. Firth[9] · Matthew Hurles[9] · Anthony A. Philippakis[8] · Gunnar Rätsch[26] · Michael Baudis[31] · Stephen T. Sherry[32] · Flicek[4]

Marc Fiume[1*], Miroslav Cupak[1], Stephen Keenan[2,3], Jordi Rambla[4], Sabela de la Torre[4], Stephanie O. M. Dyke[5], Anthony J. Brookes[6], Knox Carey[7], David Lloyd[8], Peter Goodhand[2,9], Maximilian Haeussler[10], Michael Baudis[11,12], Heinz Stockinger[12], Lena Dolman[2,9], Ilkka Lappalainen[3,13], Juha Törnroos[13], Mikael Linden[13], J. Dylan Spalding[3], Saif Ur-Rehman[3], Angela Page[2,14]

ROADMAP

F1000Research 2020, 9(ELIXIR):1229 Last updated: 20 JUL 2021

Check for updates

The ELIXIR Human Copy Number Variations Community: building bioinformatics infrastructure for research [version 1; peer review: 2 approved]

David Salgado[1], Irina M. Armean[2], Michael Baudis[3], Sergi Beltran[4,5], Salvador Capella-Gutierrez[6,7], Denise Carvalho-Silva[2,8], Victoria Dominguez Del Angel[9], Joaquin Dopazo[10], Laura I. Furlong[11], Bo Gao[3], Leyla Garcia[2,12,13], Dietlind Gerloff[14], Ivo Gut[4,5], Attila Gyenesei[15], Nina Habermann[16], John M. Hancock[13], Marc Hanauer[17], Eivind Hovig[18,19], Lennart F. Johansson[20], Thomas Keane[2], Jan Korbel[16], Katharina B. Lauer[13], Steve Laurie[4], Brane Leskošek[21], David Lloyd[13], Tomas Marques-Bonet[22], Hailiang Mei[23], Katalin Monostory[24], Janet Piñero[11], Krzysztof Poterlowicz[25], Ana Rath[17], Pubudu Samarakoon[26], Ferran Sanz[11], Gary Saunders[13], Daoud Sie[27], Morris A. Swertz[20], Kirill Tsukanov[2], Alfonso Valencia[6,7,28], Marko Vidak[21], Cristina Yenyxe González[2], Bauke Ylstra[29], Christophe Béroud[1,30]

Leveraging European infrastructures to access 1 million human genomes by 2022

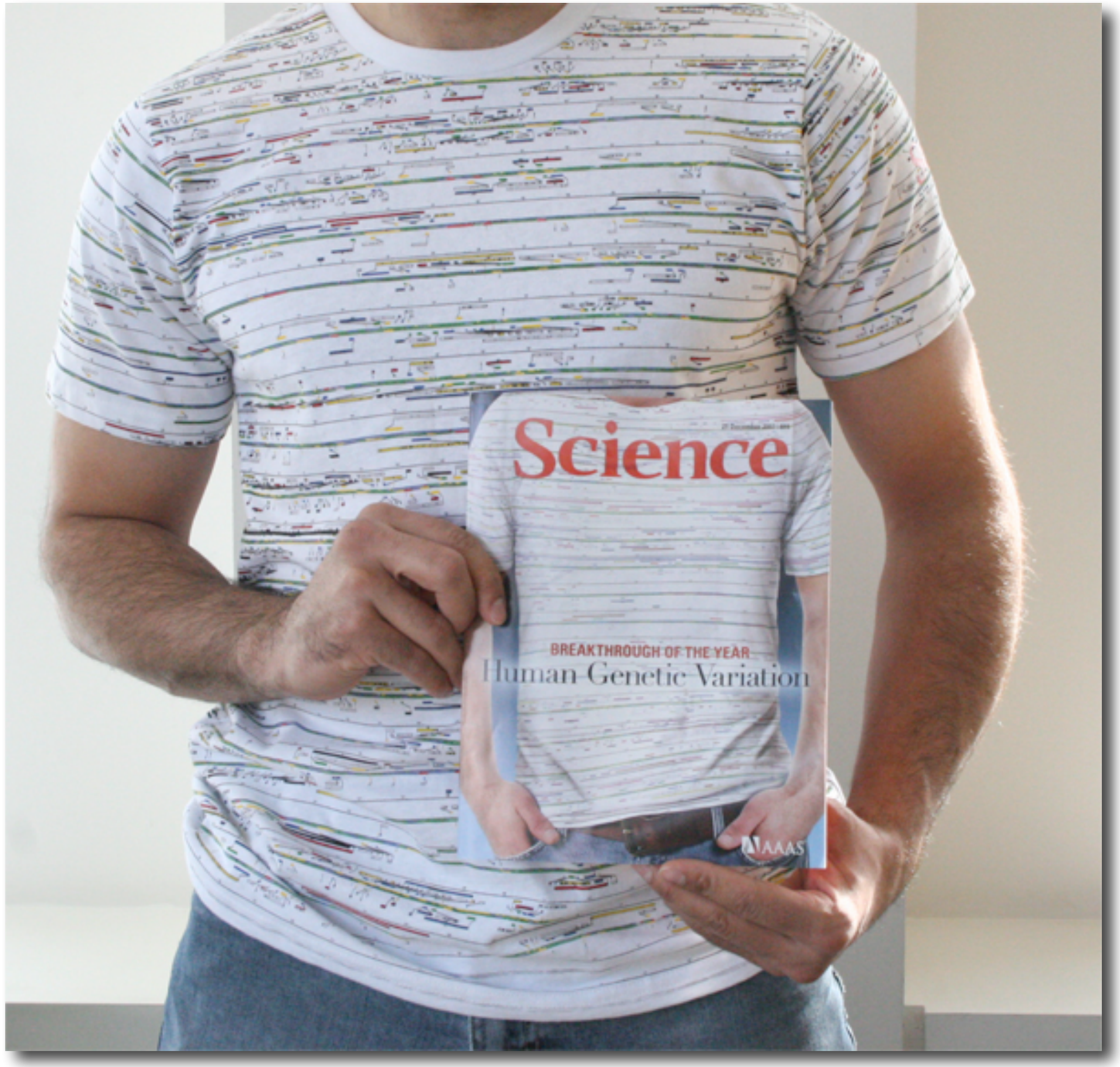
*Gary Saunders[1], Michael Baudis[2], Regina Becker[3], Sergi Beltran[4,5], Christophe Béroud[6,7], Ewan Birney[8], Cath Brooksbank[8], Søren Brunak[9,10], Marc Van den Bulcke[11], Rachel Drysdale[1], Salvador Capella-Gutierrez[12], Paul Flicek[8], Francesco Florindi[13], Peter Goodhand[14,15], Ivo Gut[4,5], Jaap Heringa[16], Petr Holub[13], Jef Hooyberghs[17], Nick Juty[18], Thomas M. Keane[8], Jan O. Korbel[19], Ilkka Lappalainen[20], Brane Leskosek[21], Gert Matthijs[22], Michaela Th. Mayrhofer[13], Andres Metspalu[23], Arcadi Navarro[24,25,26], Steven Newhouse[8], Tommi Nyrönen[20], Angela Page[15,27], Bengt Persson[28], Aarno Palotie[29], Helen Parkinson[8], Jordi Rambla[26], David Salgado[6], Erik Steinfelder[13], Morris A. Swertz[30], Alfonso Valencia[12,31], Susheel Varma[8], Niklas Blomberg[1] and Serena Scollen[1]**

Genomic Data & Privacy

Risks & opportunities

Michael Baudis | UZH BIO392 | October 2021

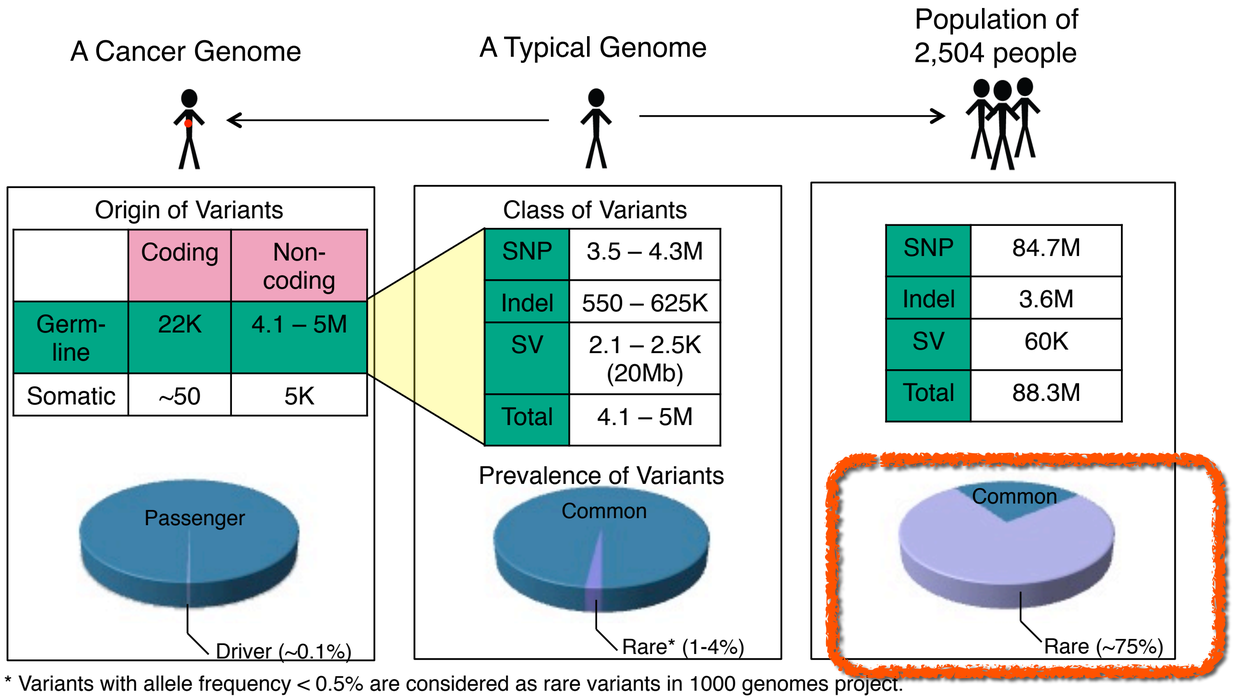
The trouble with human genome variation



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



* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

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

Limited Population Diversity in Cancer Studies

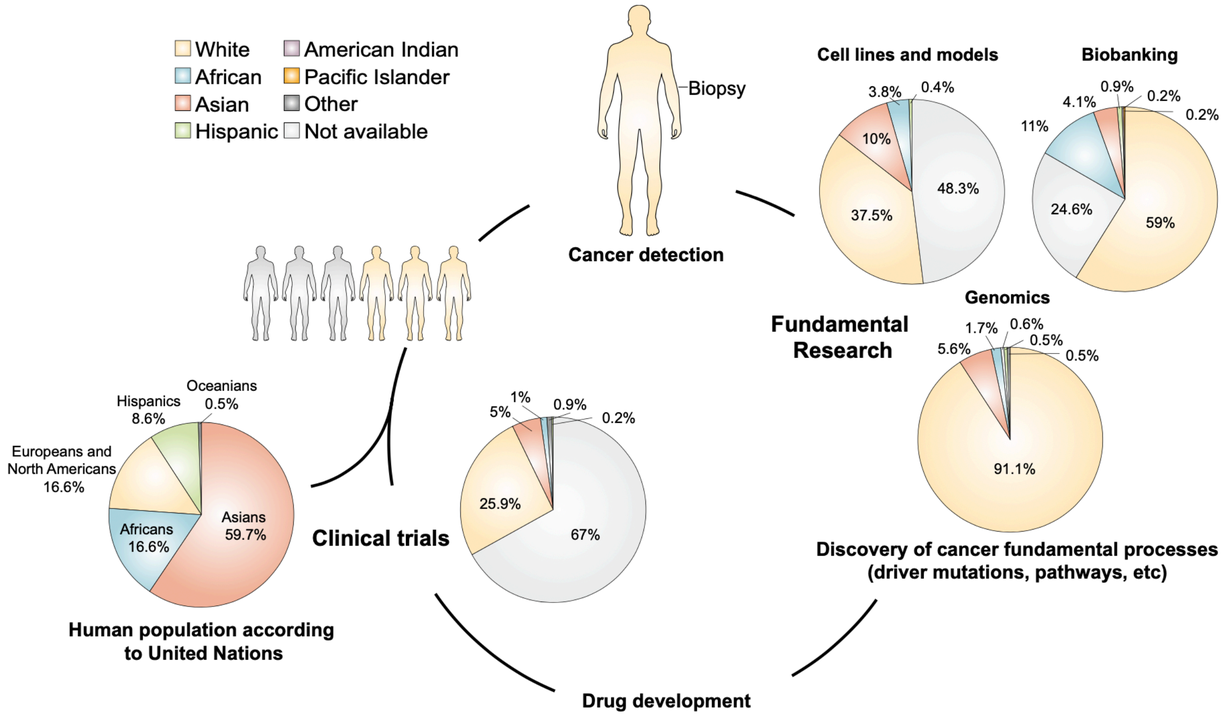


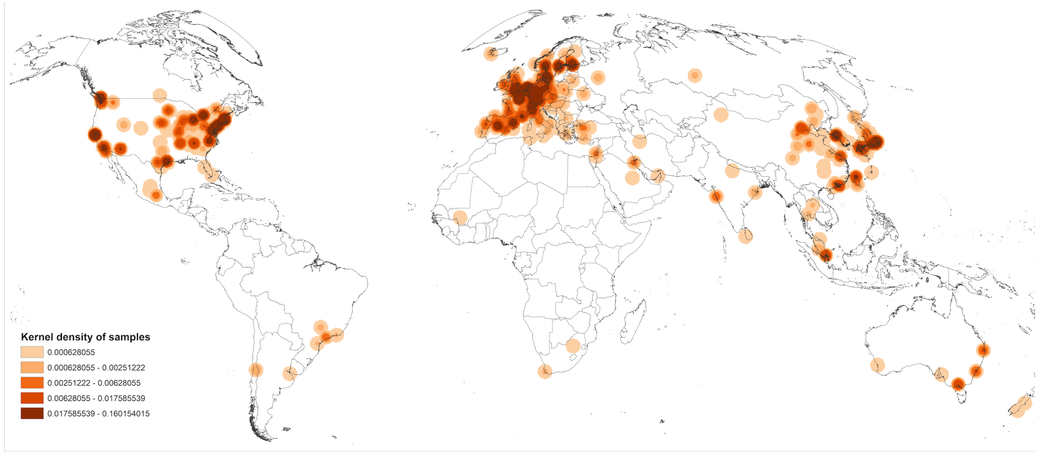
Figure 1. Racial/Ethnic disparities in cancer research. Racial/ethnic inclusion was studied in several aspects of oncological research, from cell lines and patient-derived xenografts to biobanking, genomics and clinical trials.

Guerrero S, López-Cortés A, Indacochea A, et al. Analysis of Racial/Ethnic Representation in Select Basic and Applied Cancer Research Studies. *Sci Rep*. 2018;8(1):13978.

Publication Landscape of Cancer CNV Profiling

Publication statistics for cancer genome screening studies. The graphic shows our as- sessment of publications reporting whole-genome screening of cancer samples, using molecular detection methods (chromosomal CGH, genomic array technologies, whole exome and genome sequencing).

For the years 1993-2018, we found 3'229 publications reporting 174'530 individual samples in single series from 1 to more than 1000 samples. Y-axis and size of the dots correspond to the sample number; the color codes indicate the technology used.



progenetix

The vision: Federation of data

GENOMICS

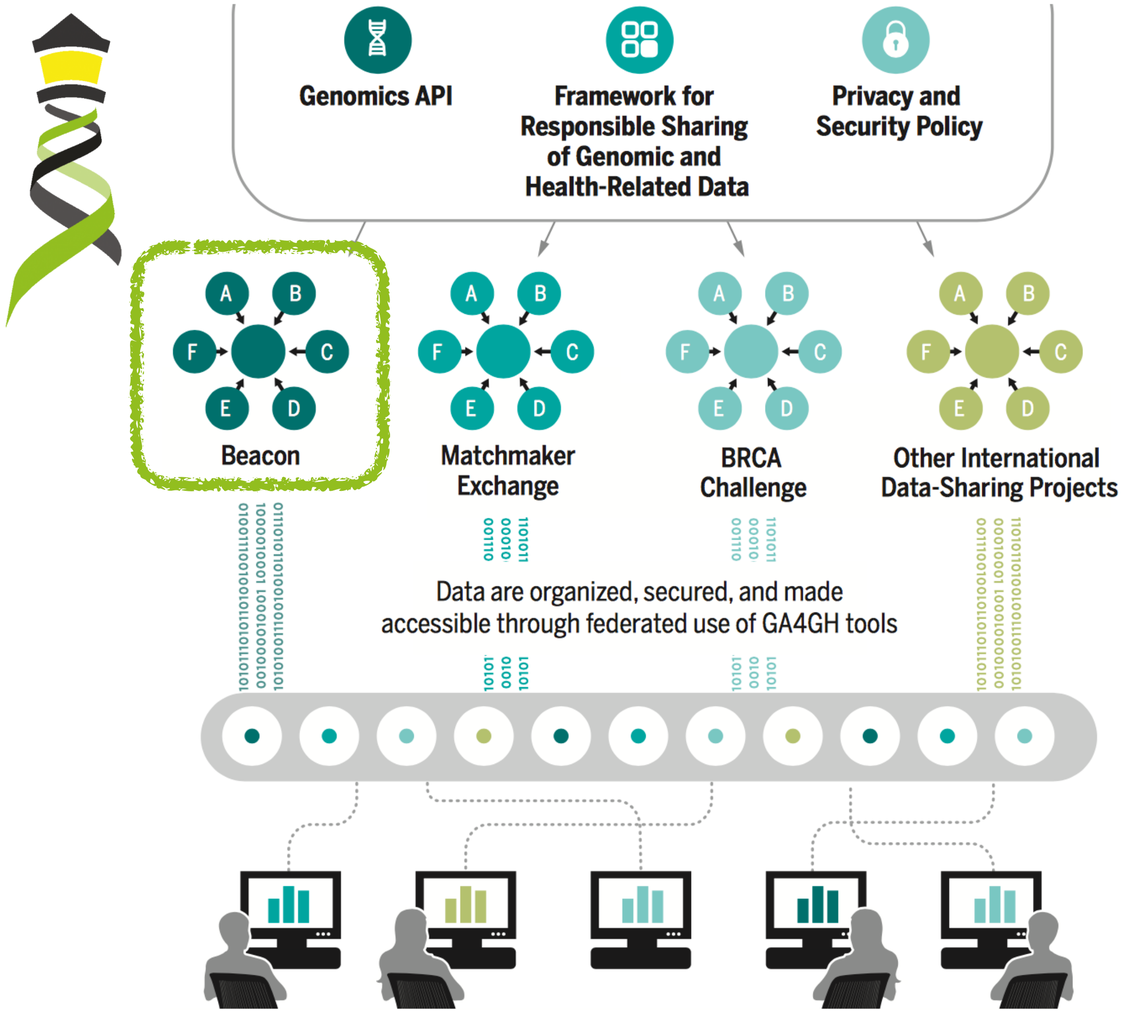
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

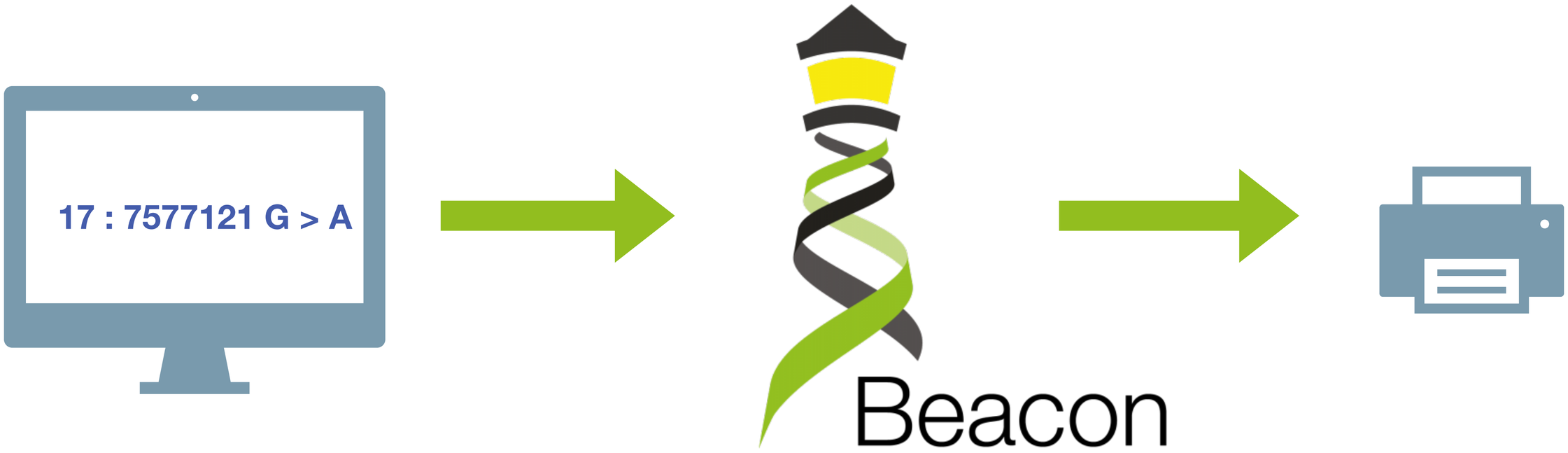
SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

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



DNASTACK

Global Alliance for Genomics & Health



A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections

YES | **NO** | **\0**

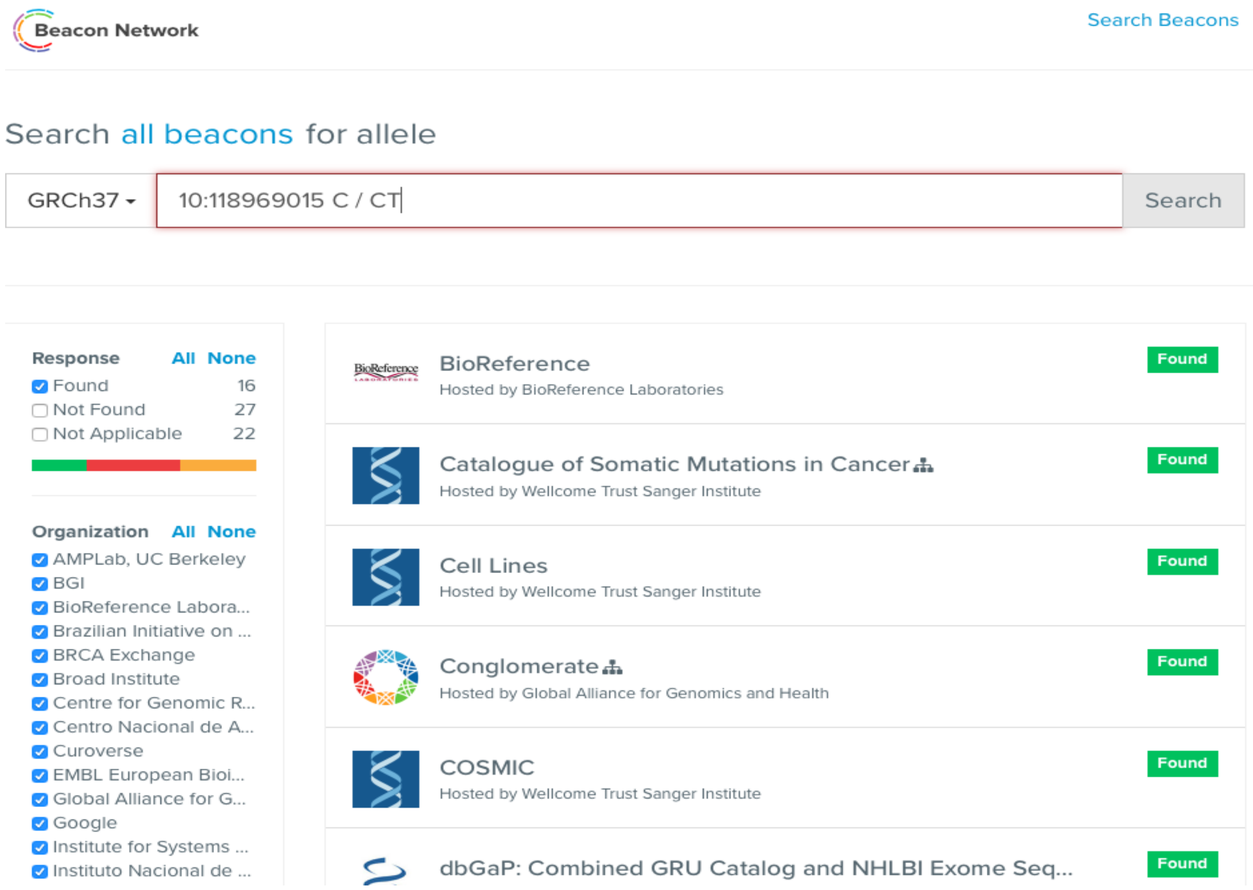
Have you seen this variant? It came up in my patient and we don't know if this is a common SNP or worth following up.

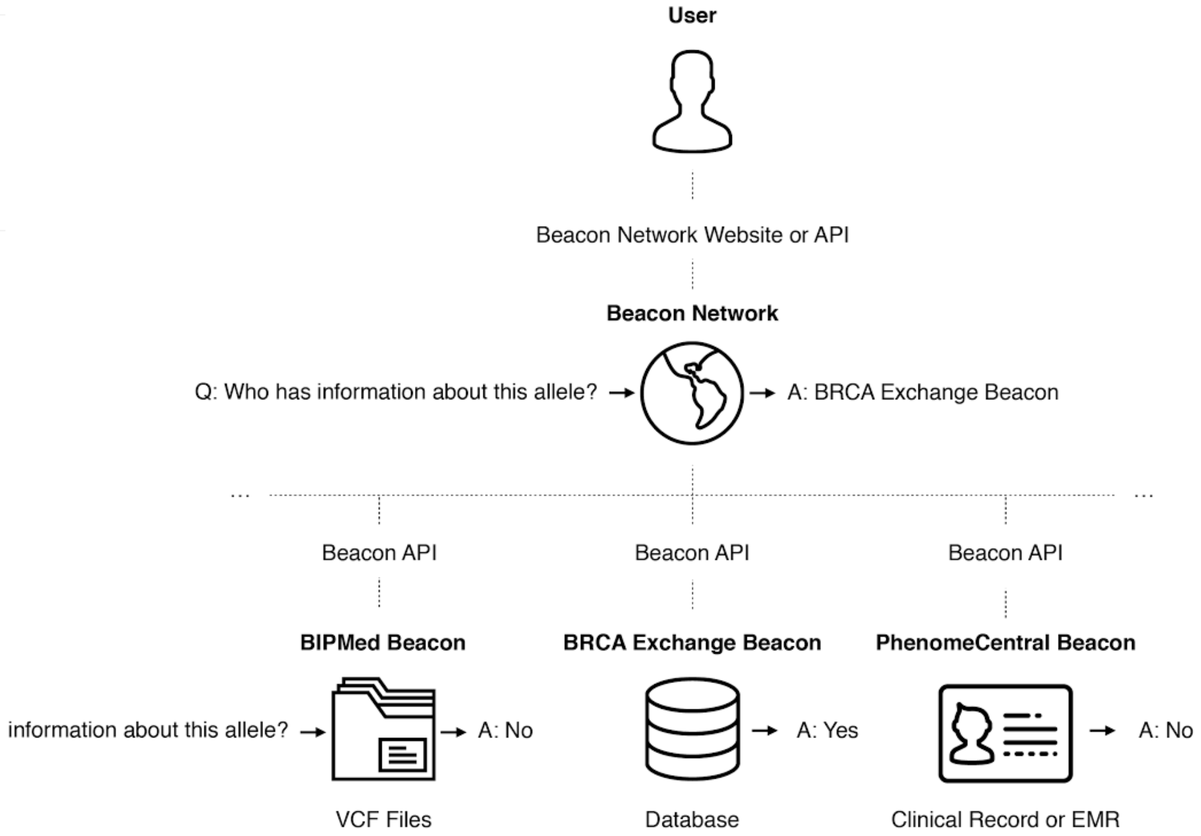
A Beacon network federates *genome variant queries* across databases that support the ***Beacon API***

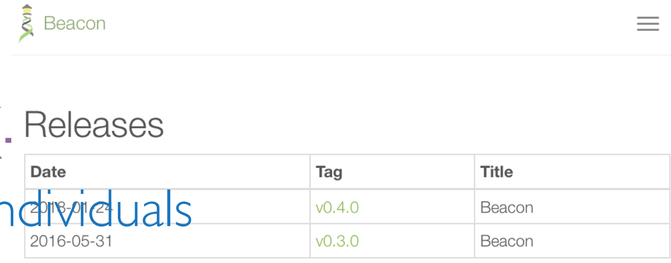
Here: The variant has been found in **few** resources, and those are from **disease** specific **collections**.

Beacon Project in 2016

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







35+ Organizations

90+ Beacons

200+ Datasets

100K+ Individuals

Global Alliance for Genomics & Health

Genome *Beacons* Compromise Security?

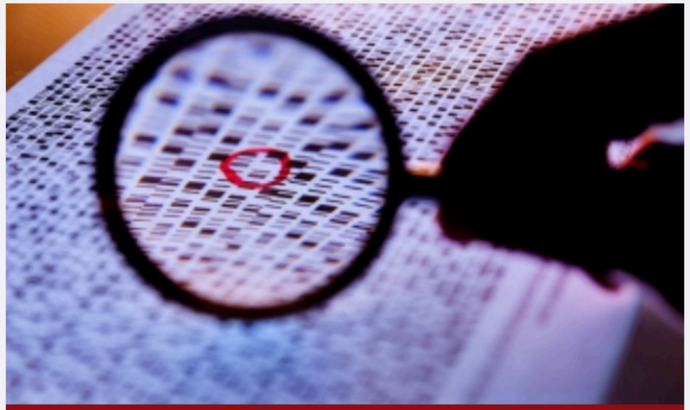
Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals in an anonymized genomic data collection

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.

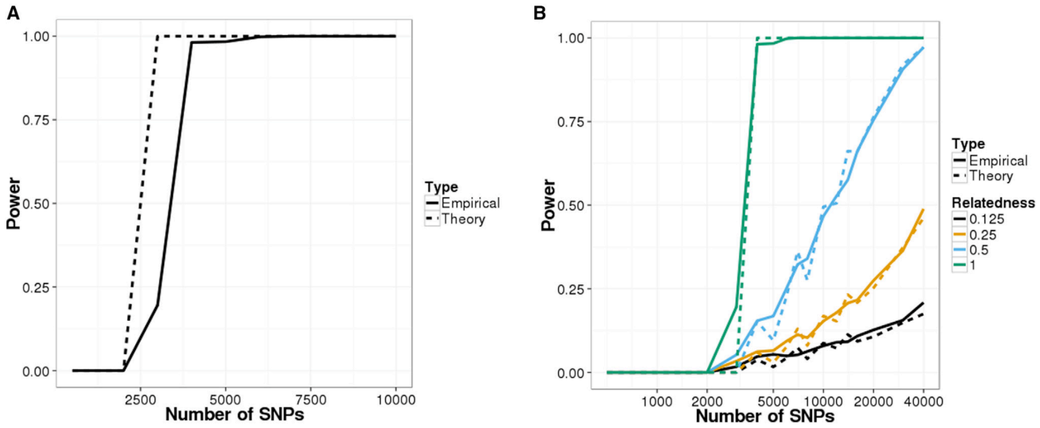


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

- 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

Information Leakage from Functional Genomics Data

- many research studies contain "functional" genomics data, e.g. from expression analyses
- such (anonymized) data may have lower protection levels than data from dedicated genotyping studies
- with a non-noisy genome of interest, attackers can generate linkage scores to identify the best match to the genomic profile

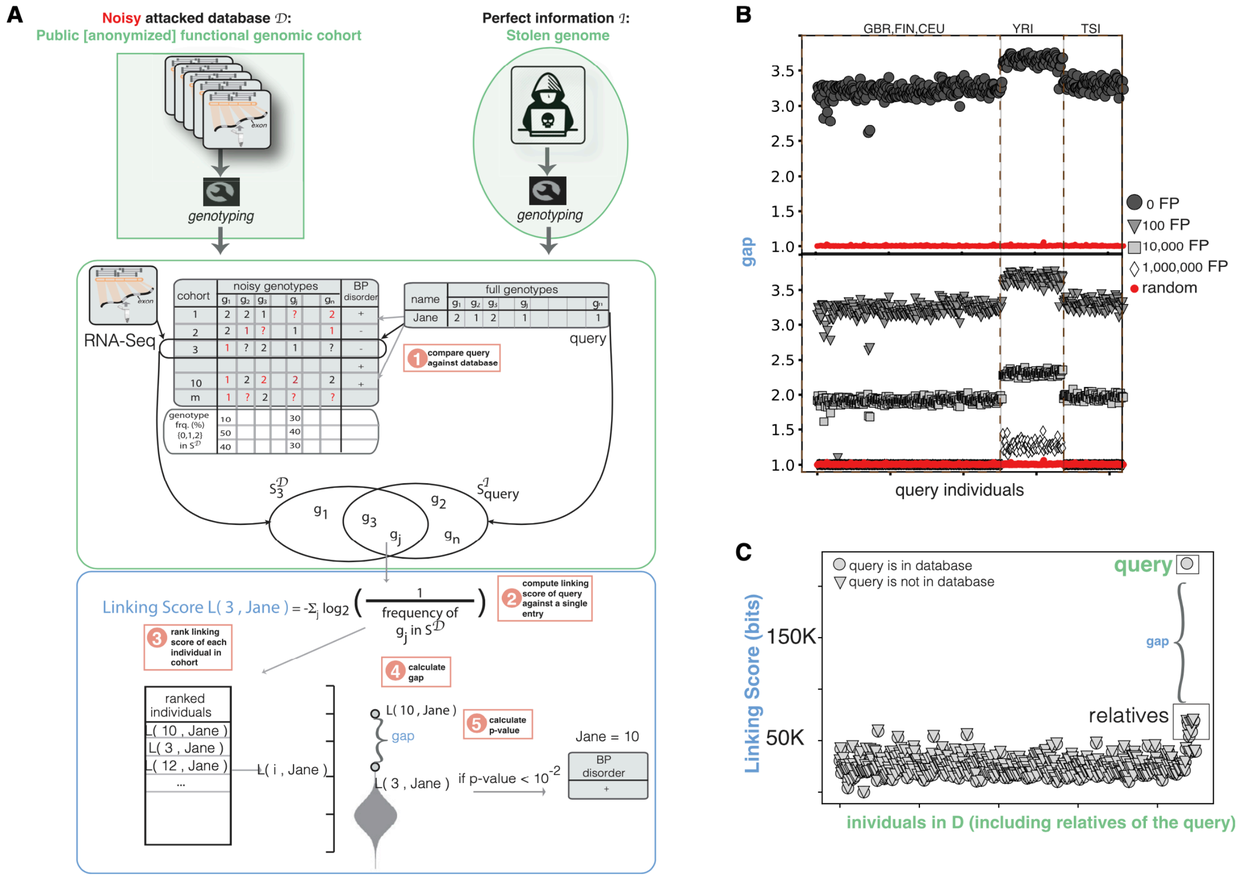


Figure 1. Functional Genomics Data De-anonymization Scheme with Perfect Genomes

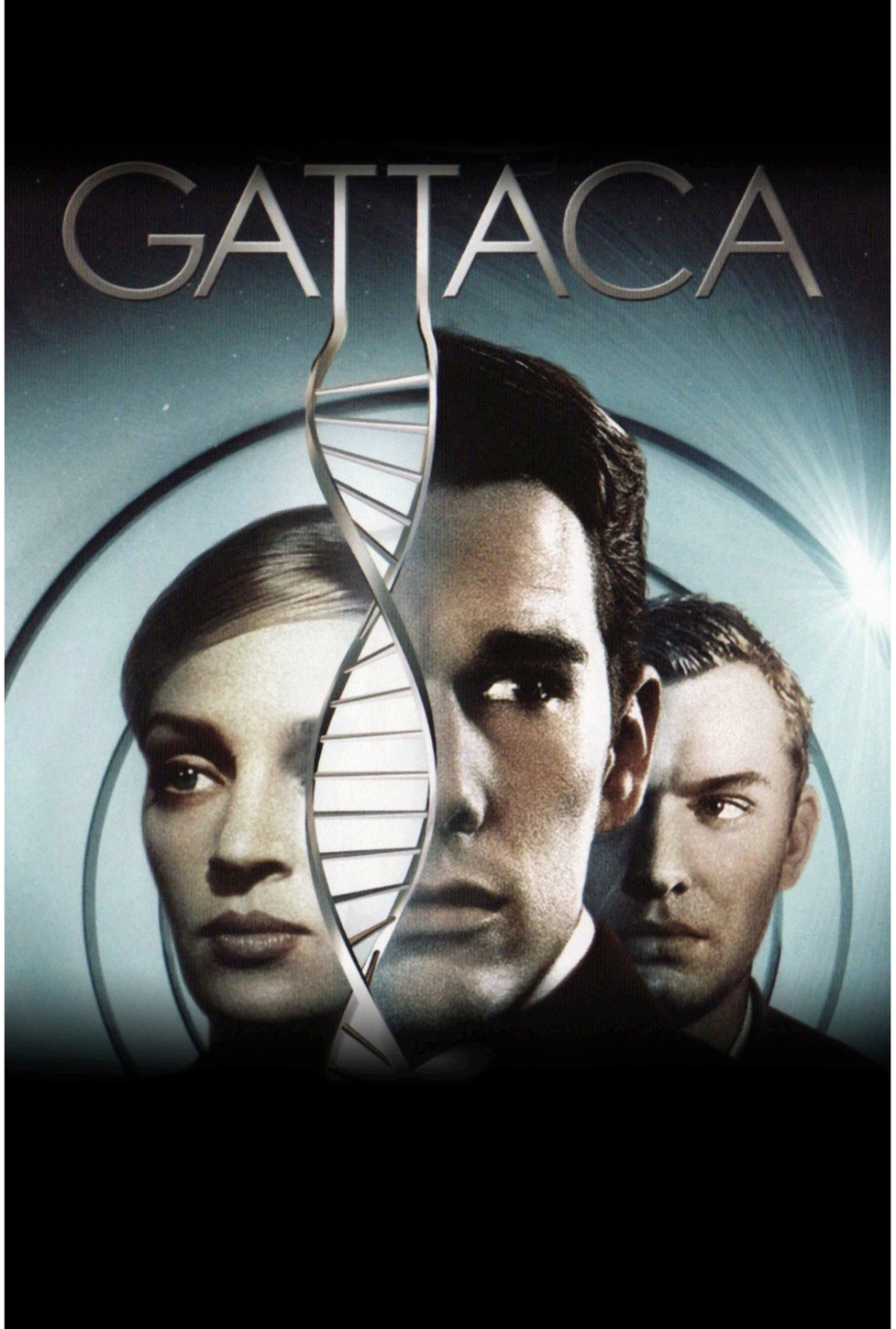
(A) Anonymized functional genomics data from a cohort of individuals can be seen as a database *D* to be attacked, which contains functional genomics reads and phenotypes for every individual in the cohort. The perfect information *I* about an individual can be the genome of an individual. After obtaining genotypes from the functional genomics reads, the attacker scores each individual in the cohort based on the overlapping genotypes between the known individual's genome and the noisy genotypes called from functional genomics. These scores are then ranked and the top-ranked individual in the cohort is selected as the known individual. See also Figure S1.

(B) *gap* values for the 1000 Genomes Project individuals in the gEUVADIS RNA-seq cohort. Red circles are the *gap* values obtained by linking a random set of genotypes to the RNA-seq panel. *gap* values are also shown after adding false-positive genotypes to the genotype set of each individual in the database.

(C) The linking scores for each individual in the functional genomics cohort after the addition of genetically related individuals to the query, with and without the query individual present in the database.

Gürsoy et al., 2020, Cell *183*, 905–917
November 12, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cell.2020.09.036

Genomes & Privacy

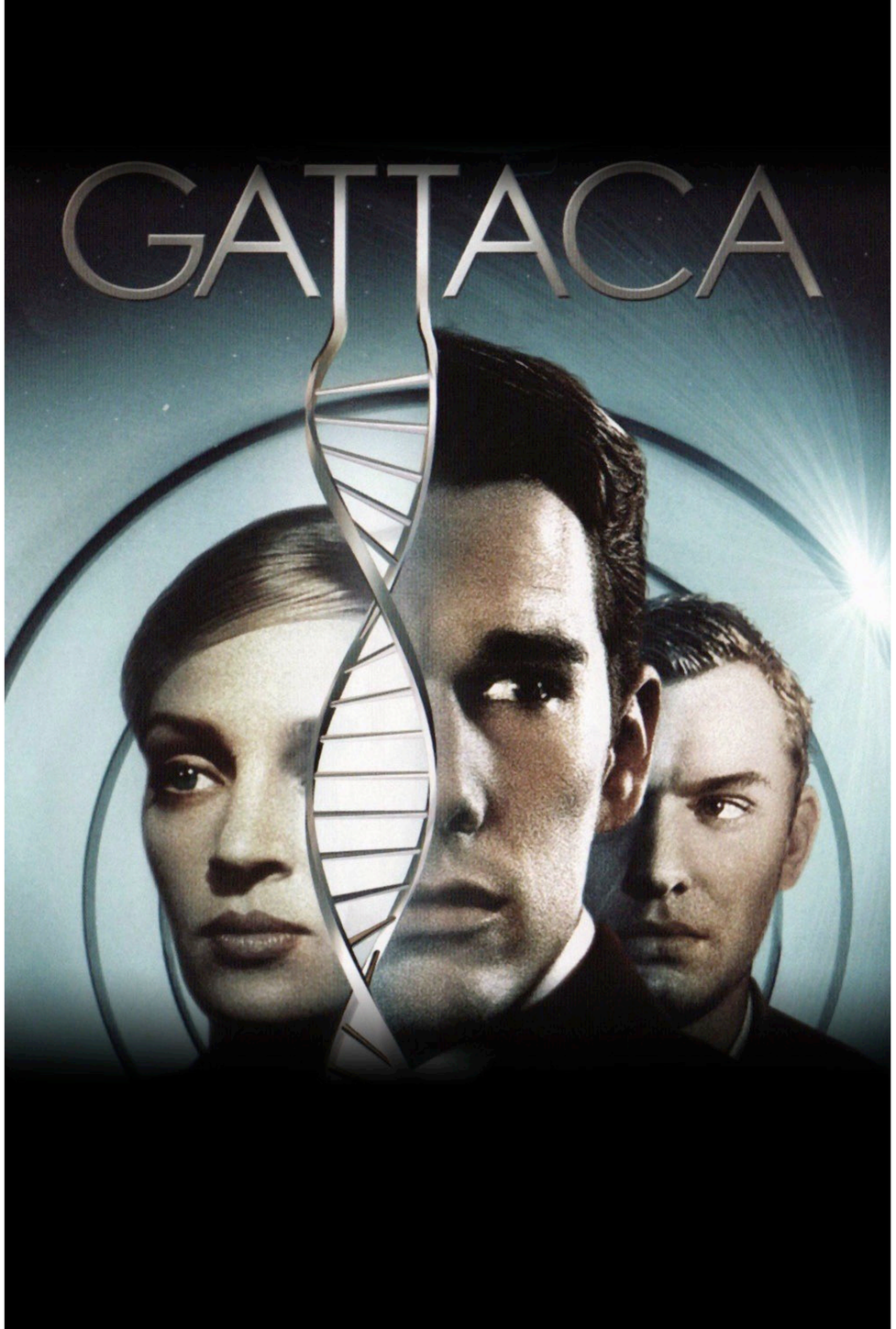


Gattaca (1997)

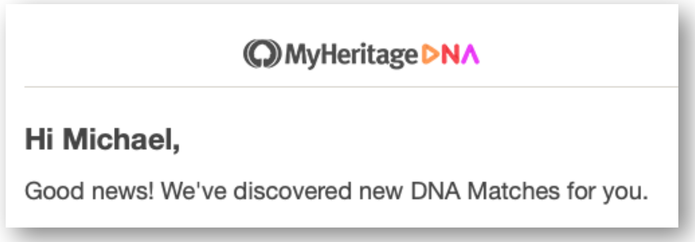
A genetically inferior man assumes the identity of a superior one in order to pursue his lifelong dream of space travel.

- genetic determinism
  - main character has been determined to be unsuitable for complex jobs based on genetic analysis
- genetic identification
  - the use of genetic sampling for personal identification is daily routine

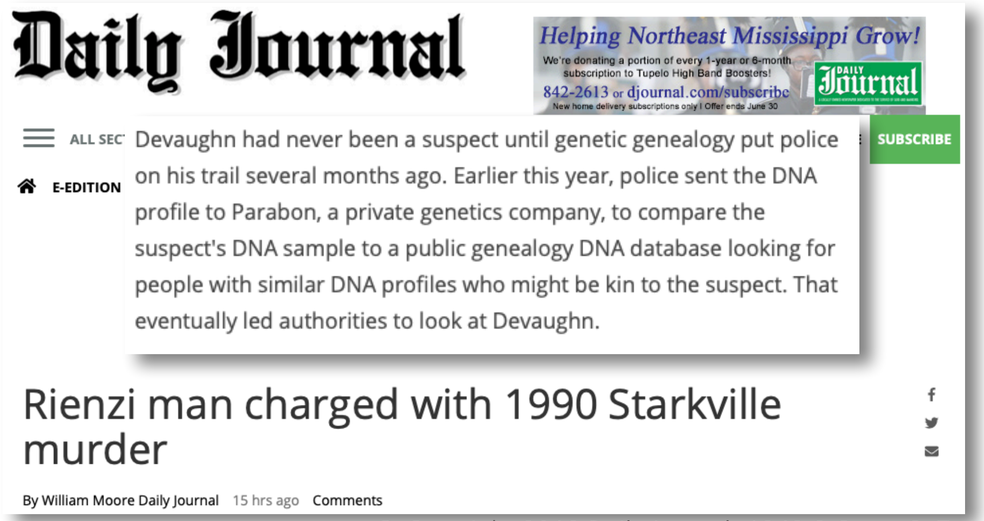
With information from https://www.imdb.com/title/tt0119177/



Long-Range Familial Searches



- 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**



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

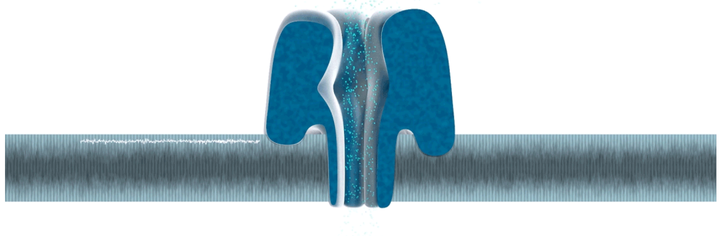
Identification of abandoned meterial using DNA fingerprinting is a common practice. The main challange currently being: time. Our method allows rapid sample preparation at the crime scene (see movie). We envision that the method can be adopted in the field for rapid checks, after a mass disaster, and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



The MinION (Oxford Nanopore)
Source: Sophie Zaaijer
https://medium.com/neodotlife/nanopore-6443c81d76d3

DEMOCRATIZING DNA FINGERPRINTING

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

ddf.teamerlich.org

MinION by Oxford Nanopore Technologies

The MinION is the smallest DNA sequencer currently around. Its the size of a Mars bar, and can be simply plugged into a laptop with a USB3.0 port.

For more information about the MinION please click:

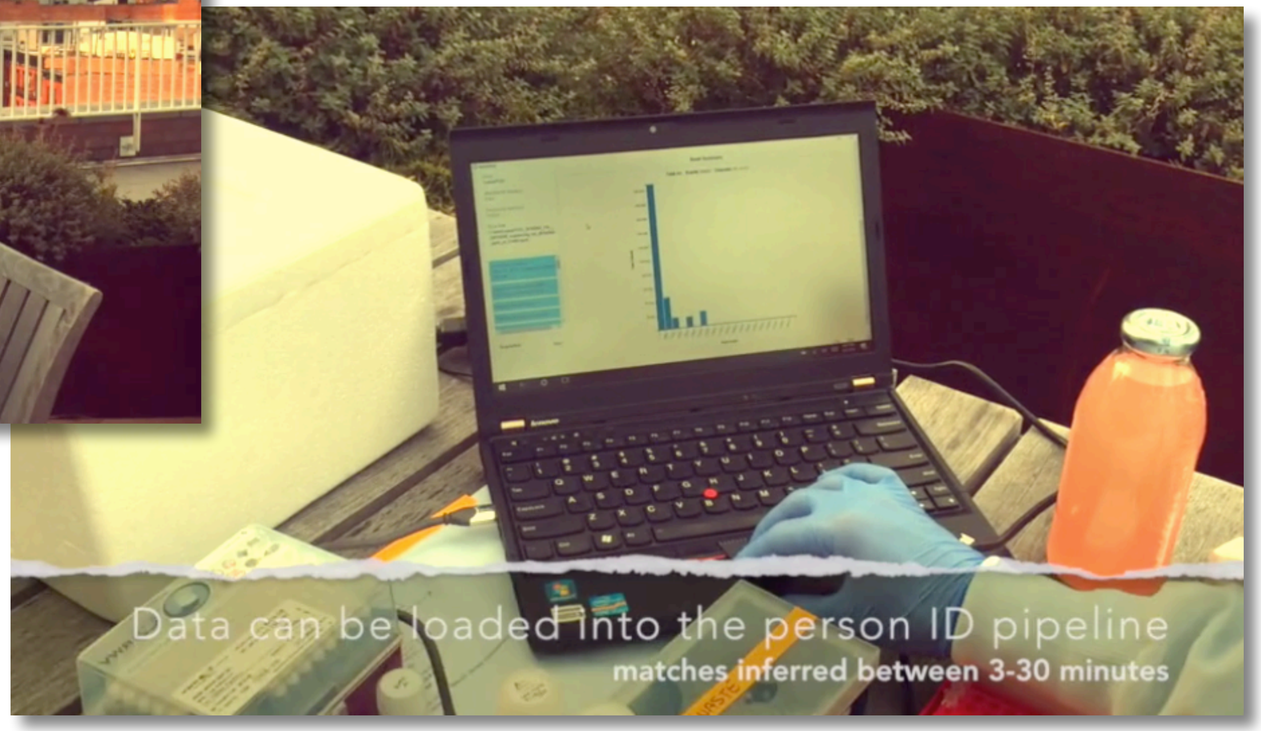
Oxford Nanopore Technologies

Bento Lab

The Bento lab is a miniature lab with a centrifuge, thermocycler and a electrophoresis compartment.

For more information about the Bento-lab please click:

Bento Lab



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

Rapid DNA

Legalizing DNA Tests for DNA Indexing

Congress / Bills / H.R. 510 (115th) / Summary

H.R. 510 (115th): Rapid DNA Act of 2017

Overview | **Summary** | Details | Text | Study Guide

GovTrack's Summary | Library of Congress

Rapid DNA is a new technique that can analyze DNA samples in about 90 minutes, instead of days or even weeks as it took previously. A bill that passed the Senate and House last week would expand the use of this technology.

What the bill does

The Rapid DNA Act establishes a system for Rapid DNA's nationwide coordination among law enforcement departments, by connecting it to the FBI's Combined DNA Index System.

Labelled S. 139 in the Senate and H.R. 510 in the House, the legislation was introduced by Sen. Orrin Hatch (R-UT) and Rep. James Sensenbrenner (R-WI5).

Former FBI Director James Comey cited a real-life example of how the technology could be used effectively. "[It will] allow us, in booking stations around the country, if someone's arrested, to know instantly—or near instantly—whether that person is the rapist who's been on the loose in a particular community before they're released on bail and get away or to clear somebody, to show that they're not the person," Comey said in testimony.

Rapid DNA was used for the first time ever in a criminal investigation in 2013, to nail burglars who stole more than $30,000 worth of items from an Air Force Member's Florida home while they were serving in Afghanistan. Presumably more such cases would be solved and quickly with expanded use of rapid DNA.

What supporters say

Supporters say it will save both time and taxpayer dollars by speeding up the DNA analysis process in a manner that's no less effective, reducing the backlog of samples waiting to be tested.

"It will enable officers to take advantage of exciting new developments in DNA technology to more quickly solve crimes and exonerate innocent suspects," Senate lead sponsor Hatch said in a press release. "Under this legislation, rather than having to all send DNA samples to crime labs and wait weeks for results, trained officers will be able to process many samples in less than two hours."

What opponents say

GovTrack Insider could not locate any members of Congress who expressed public opposition to the legislation, but some members of the public are concerned. The *New Republic* called the rise of rapid DNA "troubling," citing the potential for privacy violations and misuses by immigration authorities. They also noted that the FBI already has DNA samples from more than 3.5 percent of Americans, a number likely to grow thanks to a 2015 Supreme Court decision allowing DNA samples to be taken without a warrant.

The Electronic Frontier Foundation expressed doubts about the accuracy of Rapid DNA. "Rapid DNA has only been tested on single-source samples—like a swab taken directly from a person's inner cheek," the EFF writes. "And yet, Rapid DNA manufacturers are trying to convince law enforcement agencies to buy these machines to get through their backlog of rape kits and for low-level property crimes—situations where there's a very good chance the DNA came from multiple people—some of whom may have had no connection to the crime at all.

Votes and odds of passage

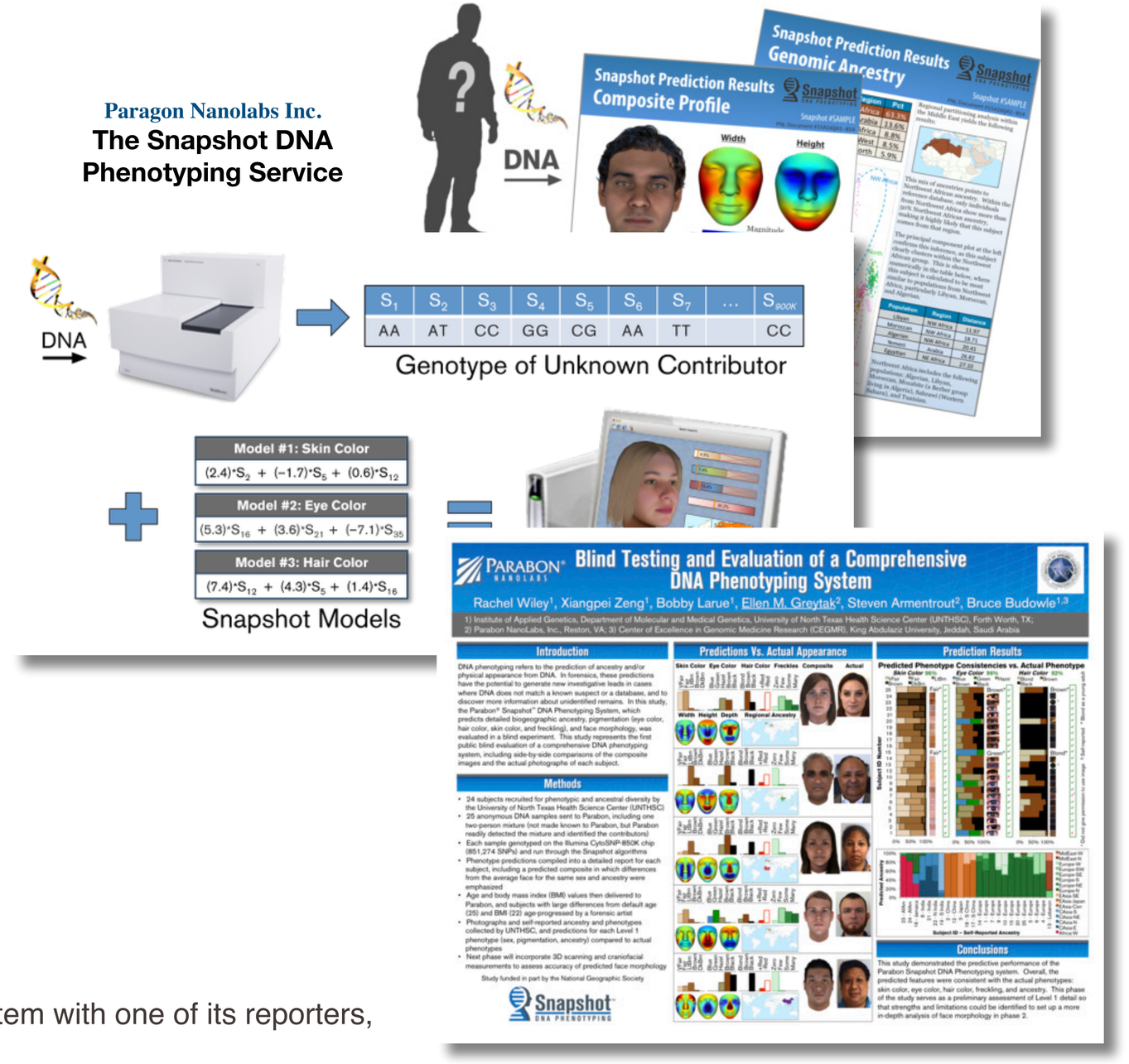
The legislation attracted a bipartisan mix of 12 Senate cosponsors, seven Republicans and five Democrats, and 24 House cosponsors, 17 Republicans and seven Democrats. It passed both the House and Senate on May 16, by a unanimous consent voice vote in both chambers, meaning no record of individual votes was recorded. It now goes to President Trump's desk, where he appears likely to sign it.

https://www.govtrack.us/congress/bills/115/hr510/summary

Phenotyping from DNA

From DNA to "Wanted" Posters?

- association of genomic variants with phenotypic data collection
- while hair, eye color are easy targets not useful for relevant phenotypic features especially if large environmental component
- huge biases based on input/collection data
- Belgium and Germany do not allow forensic DNA phenotyping
- Switzerland: Bundesrat decision on 2020-12-04 to allow phenotyping for law enforcement purposes



"When the New York Times ran an informal test of the Parabon system with one of its reporters, it failed badly." (ACLU.org)

https://snapshot.parabon-nanolabs.com/phenotyping

Federal Act on the Use of DNA Profiles in Criminal Proceedings and for Identifying Unidentified or Missing Person, DNA Profiles Act

An Area in Transition...

- Currently: «Genetic Fingerprint»
- Future: Will it be allowed to take a deeper look and how far can genetic data be used to determine the characteristics of an unknown perpetrator (colour of hair and eyes, height, ethnicity, etc.)
- Switzerland: Bundesrat decision on 2020-12-04 to allow phenotyping for law enforcement purposes

Content provided by Julian Mausbach | Oberassistent für Strafrecht und Strafprozessrecht

HGTA

Federal Act on Human Genetic Testing

- Currently limited to
  - a. in the medical context;
  - b. in the context of employment;
  - c. in the context of insurance;
  - d. in the context of liability.
- Therefore: Direct to Consumer Tests (DCT) ***not*** allowed in Switzerland
- But: Changes in the newer future are to be expected… e.g. DTC will be possible in limited ways.

HGTA : Federal Act on Human Genetic Testing

HGTA new (probably 2021)	medical field	outside the medical field	
Investigated characteristics	medical relevant	especially protective values characteristics	other characteristics
General Requirements	Non-discrimination, information and consent, right to information, right not to know, avoidance of surplus information, protection of samples and genetic data, Circulation concerning public advertising, state of science and technology, penal provisions		
Initiation	Physician	Health professional (controlled taking of samples)	Consumer **(DTC)**
Persons concerned	Persons with **and** without capacity of judgement, pregnant woman (PND)	ONLY persons with Capacity of judgement	ONLY persons with Capacity of judgement
Communication of surplus information	as a rule according to decision of the person concerned	Not allowed	Not allowed
Laboratory	subject to authorization (cyto and molecular genetic studies)	subject to authorization (cyto and molecular genetic studies)	not subject to authorisation
Employers and Insurance institutions	Studies and Recovery of Results / Data only in regulated exceptional cases	Prohibition to carry out investigations and the Recovery of Results / Data	Prohibition to carry out investigations and the Recovery of Results / Data

Data Ownership

- Within Switzerland, there is no coherent approach on ownership of data as such (but academic discussion is ongoing, if that is needed).
- Restrictions of usage and disclosure of data other than personal data mainly stem from contractual relationships.
- In the field of research this leads mostly to a data ownership by the research institution.

Of course the restrictions of the different acts that are in the field need to be respected (procuring data lawfully, consent for further use, etc.)

Content provided by Julian Mausbach | Oberassistent für Strafrecht und Strafprozessrecht

Is Genomic Data Special?

Health Related Data & Privacy

Considerations when evaluating risks of data sharing

- 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

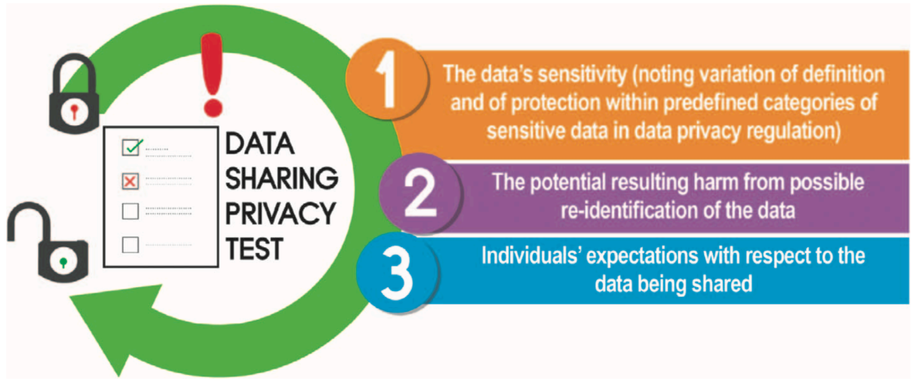


Figure 1. The three steps of a Data Sharing Privacy Test to distinguish degrees of data sensitivity within categories of data recognised as 'sensitive'.

Typical Data Scopes in Genomics (Research) Collections

Biomedical and procedural "Meta"data types

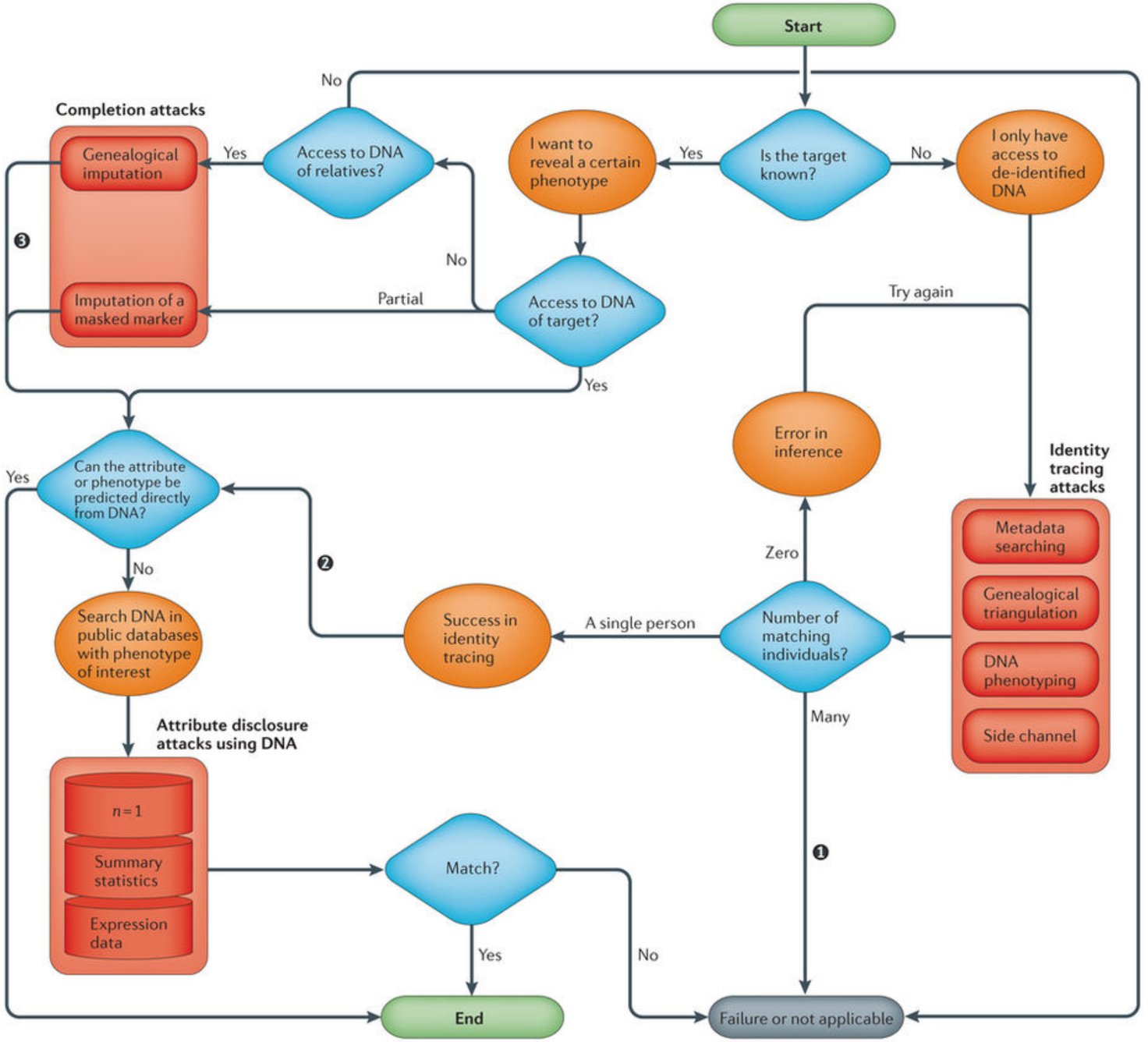
- Diagnostic classification
  - mapping text-based cancer diagnoses to standard classification systems
- Provenance data
  - store identifier-based pointers
  - geographic attribution (individual, biosample, experiment)
- Clinical information
  - **core set** of typical cancer study values:
    - ➡ stage, grade, followup time, survival status, genomic sex, age at diagnosis
  - balance between annotation effort and expected usability

progenetix

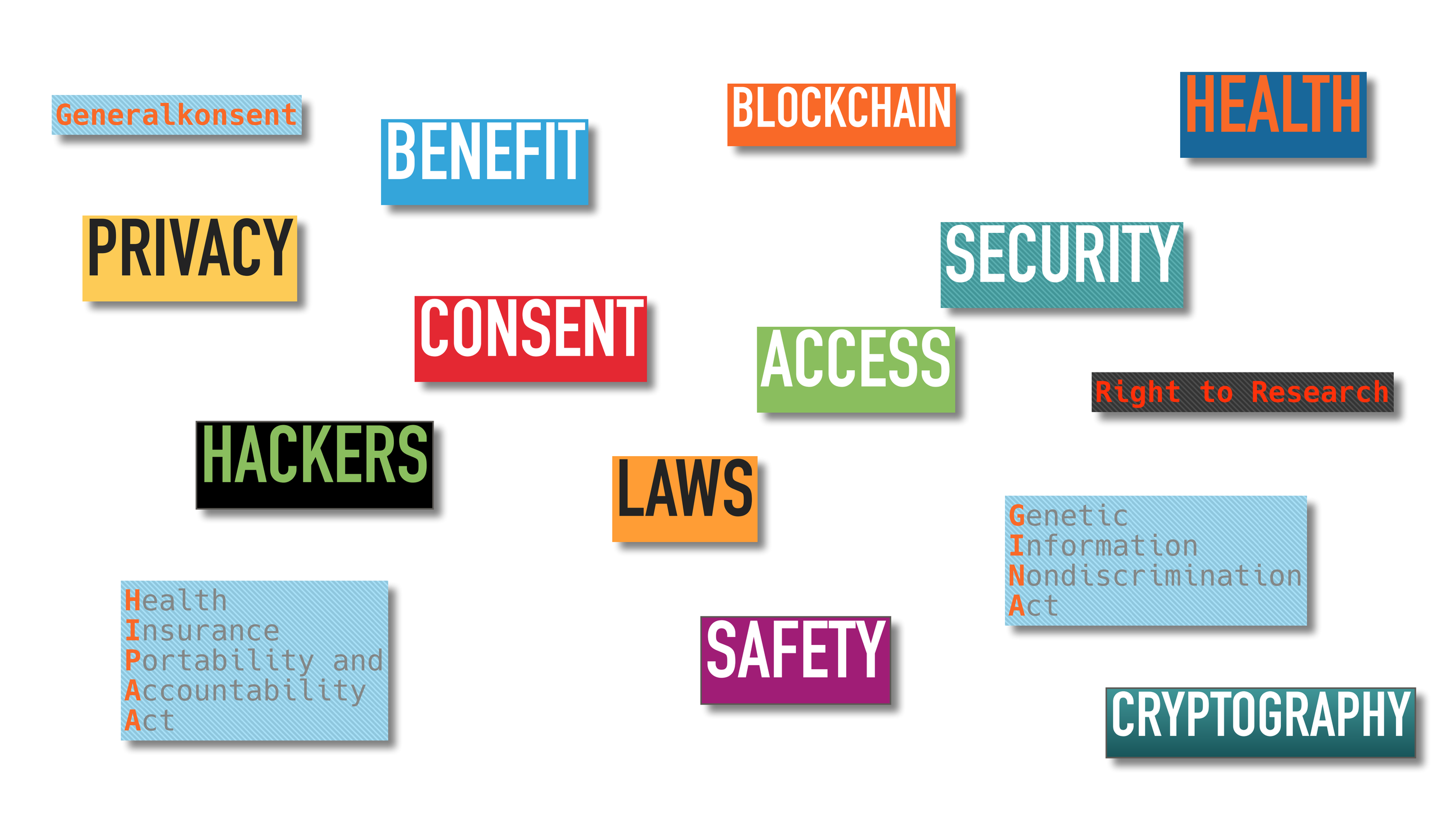
Routes for breaching and protecting genetic privacy

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)


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

Way forward...

The vision: Federation of data

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 ...

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

Art des Forschungs-materials / Personenbezug	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

SAMWASSM

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

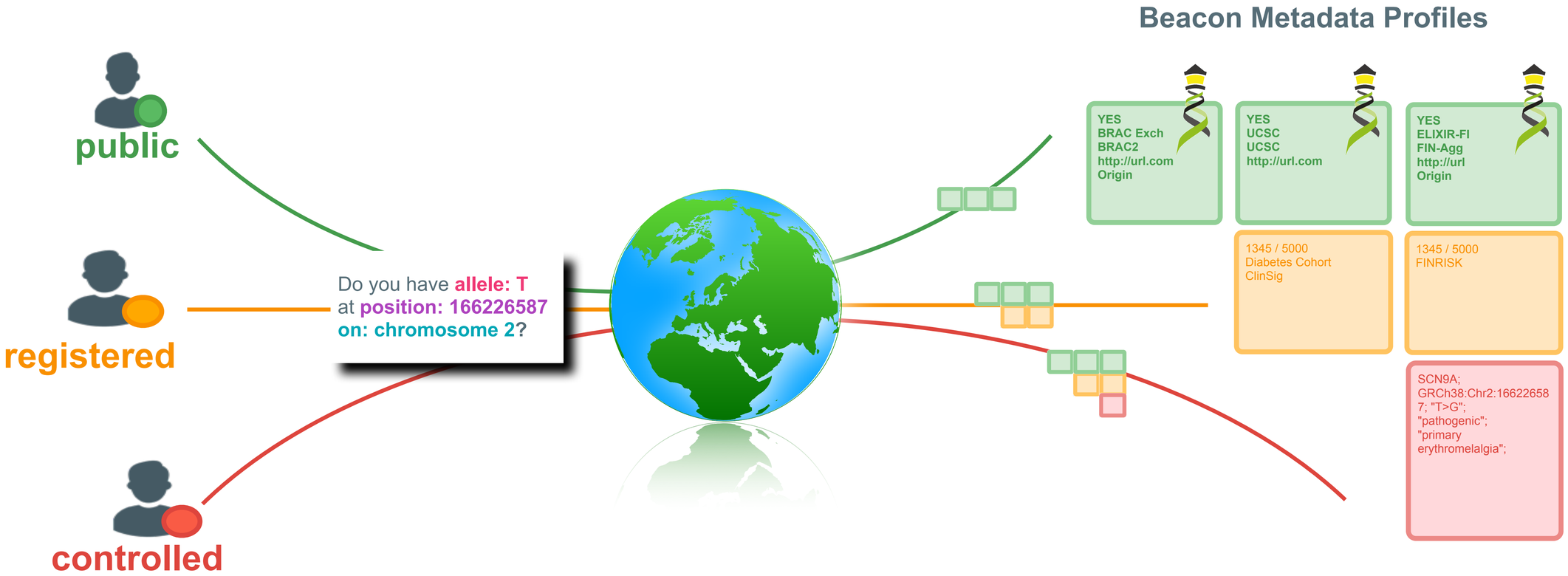
Name	Abbreviation	Description
Primary Categories (I^{ry})		
no restrictions	NRES	No restrictions on data use.
general research use and clinical care	GRU(CC)	For health/medical/biomedical purposes and other biological research, including the study of population origins or ancestry.
health/medical/biomedical 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^{ry}) (can be one or more extra conditions, in addition to I^{ry} 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)

Empowering Beacon use through Access Levels

Integrating permissions and discovery

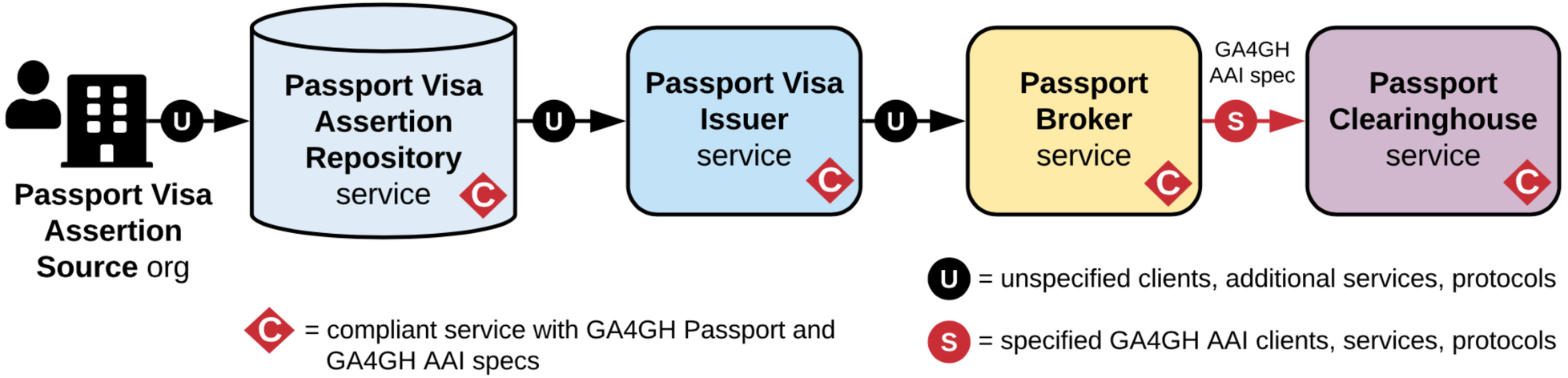


https://www.youtube.com/watch?v=LyfmvAs7LtQ&feature=youtu.be

GA4GH Passports

Global Alliance for Genomics & Health

Communicating a user's data access authorizations



www.ga4gh.org/ga4gh-passports/

- format to communicate a user's data access authorizations based on either their role (e.g. researcher), affiliation, or access status
- works together with the GA4GH Authentication and Authorization Infrastructure (AAI) OpenID Connect Profile to streamline researchers' data access over federated data access protocols
- both standards approved in Dec 2019 with early implementation by Google Cloud services and ELIXIR

Google Cloud

elixir

Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

Beacon *v2* API

The Beacon API v2 proposal opens the way for the design of a simple but powerful "**genomics API**".

Improving Data Privacy but Empowering Beneficial Use

Intersecting Areas of Development

- Make genomic (and functional) data "obfuscated" for malicious use
  - e.g. spiking / randomization of variants in "not-disease" loci
- access protection with defined user access using standardized protocols for users' roles and permissions, in contrast to individual per user, per dataset access requests over data access committees (DACs)
  - digital "differential" consent using e.g. data use ontologies
- intentional and unintentional (!) data providers have to be protected from abuse by legal regulations - though thin line regarding "overzealous" use by law enforcement
- alternative solution for active consent
  - encrypted wide-area networking solutions with managed access control (e.g. SPHN's BiomedIT) and limited access to anonymized data (e.g. using the Beacon protocol with "handover" scenarios)
  - (genomic) data ownership by the individual "data donors, together with strong privacy protection by law

Genomic Data & Privacy - Key Areas

- **Re-identification**
  - identification of an individual based on sets of genomic variants they (or close relatives) carry - so one needs some genome data first
  - information to be gained is circumstantial (e.g. their genome is in a particular disease related dataset)
  - currently only risk with some practical use (e.g. **long-range familial attacks**)
- **Genotype-to-Phenotype (G2P) attacks**
  - determination of some disease risk or phenotypic features from a genome itself
  - needs access to genome data which is illegal in many jurisdictions (but technically more & more feasible)
  - real-world use cases are limited but abuse through wrong perception of utility
- **Genomic Determinism**
  - assignment of individual abilities and personal development trajectories from genomic profiling
  - topic of (some good, most bad) SciFi
  - but: **Wehret den Anfängen**!

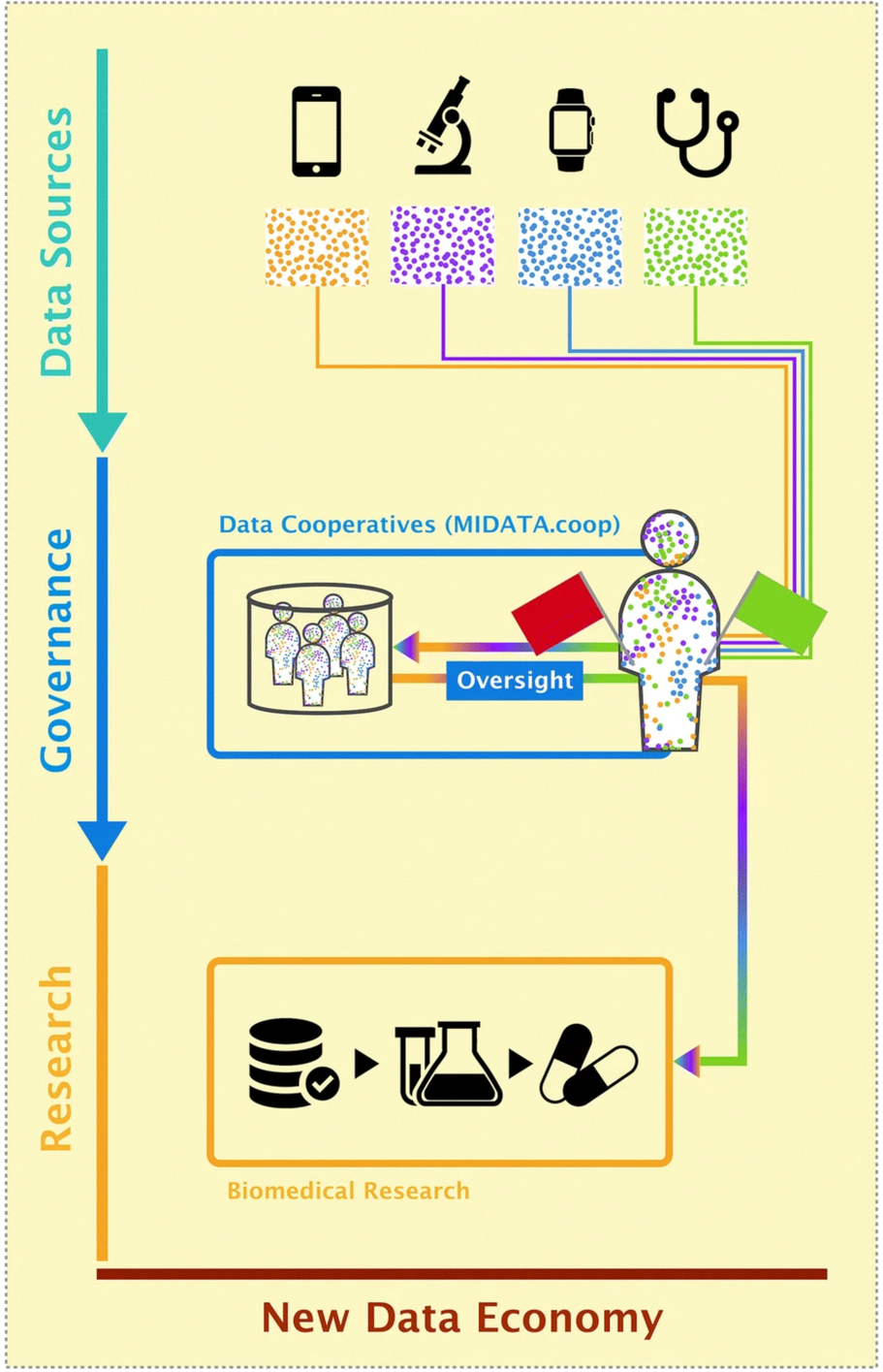
Genomic Data & Privacy - Some Take-Home Messages

- Many clinical and research applications in genomics **need vast numbers of genomes** to evaluate e.g. genotype-phenotype relationships
- Such data cannot simply be provided by a few reference data curation resources - and those again rely on multitudes of original data resources > **federated data access** + **data curation**
- Genomic data is considered to potentially expose unwilling individuals through **re-identification**/de-anonymization but also through direct information (genotype -> phenotype/disease)
- Legislative bodies and law enforcement have varying and *curious* approaches to "genomic privacy", with a mix of de-legalizing genomic data generation (e.g. in Switzerland) or strictly limiting its use while also using "eminent domain" to co-opt such data for criminal persecution in a possibly extending set of use cases

Power to the People?!

Individuals as Owners & Managers of their Data

- (genomic) data ownership by the individual "data donors"
- supported by technological frameworks for data management and arbitration
- one vision here are "data cooperatives"
- need strong support from policy makers and financial sustainability support



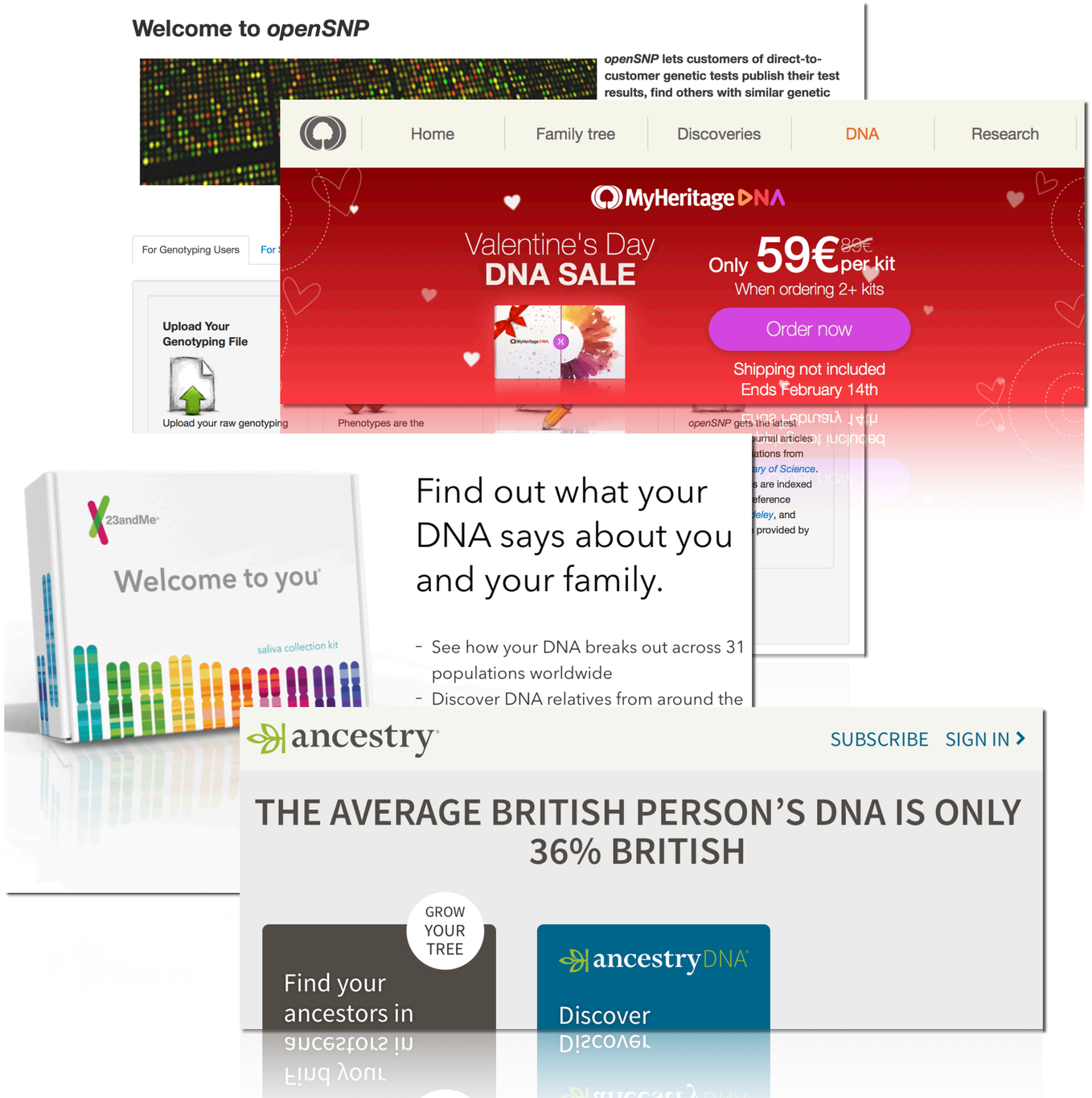
Citizens aggregate data from different sources and make them available for research through data cooperatives. Cooperatives offer oversight mechanisms to filter data access requests and tools for the democratic governance of the data.
Blasimme, A., Vayena, E. & Hafen, E. **Democratizing Health Research Through Data Cooperatives**. *Philos. Technol.* 31, 473–479 (2018). https://doi.org/10.1007/s13347-018-0320-8

graphic credit Manuel Schneider

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!



John Yuyi, NYT 2018-02-09

Welcome to *openSNP*

openSNP **lets customers of direct-to-customer genetic tests publish their test results, find others with similar genetic**

For Genotyping Users

Upload Your Genotyping File

Upload your raw genotyping

Phenotypes are the

openSNP gets the latest

Home | Family tree | Discoveries | DNA | Research

MyHeritage DNA

Valentine's Day
DNA SALE

Only **59€** ~~89€~~ per kit
When ordering 2+ kits

Order now

Shipping not included
Ends February 14th

23andMe

Welcome to you

saliva collection kit

Find out what your DNA says about you and your family.

- See how your DNA breaks out across 31 populations worldwide
- Discover DNA relatives from around the

ancestry

SUBSCRIBE SIGN IN

THE AVERAGE BRITISH PERSON'S DNA IS ONLY 36% BRITISH

GROW YOUR TREE

Find your ancestors in

ancestryDNA

Discover

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 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).

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

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,

Exercises

Read, Think, Opinionate ...

- The course material contains a number of articles (scientific, news) about topics touched upon in the course.
- Please use the time after the course to pick some you find interesting! Files are available through the course page:
  - "Longe-range familial attacks" has both news and scientific write-ups (Ehrlich et al. 2018)
  - the Beacon protocol has article but also online resources at http://beacon-project.io
  - ga4gh.org has links to many related topics