Personalized medicine in cancer

Genome Variation | Data Formats | Resources | Sharing | Privacy

Michael Baudis **UZH SIB**
Computational Oncogenomics

University of Zurich[UZH]

1992

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

dkfz.

Heidelberg

Student of medicine | doctoral thesis in molecular cytogenetics @ DKFZ (Peter Lichter) | resident in clinical hematology/oncology | data, clinical studies & cancer systematics

2001

Stanford University

Stanford

Post-doc in hemato-pathology (Michael Cleary) | molecular mechanisms of leukemogenesis | transgenic models | expression arrays | systematic cancer genome data collection | *Progenetix* website

2003

UF | UNIVERSITY of FLORIDA

Gainesville

Assistant professor in paediatric haematology | molecular mechanisms of leukemogenesis | focus on bioinformatics for cancer genome data analysis

2006

RWTH AACHEN UNIVERSITY

Aachen

Research group leader in genetics | genomic array analysis for germline alterations | descriptive analysis of copy number aberration patterns in cancer entities

2007

SIB Swiss Institute of Bioinformatics

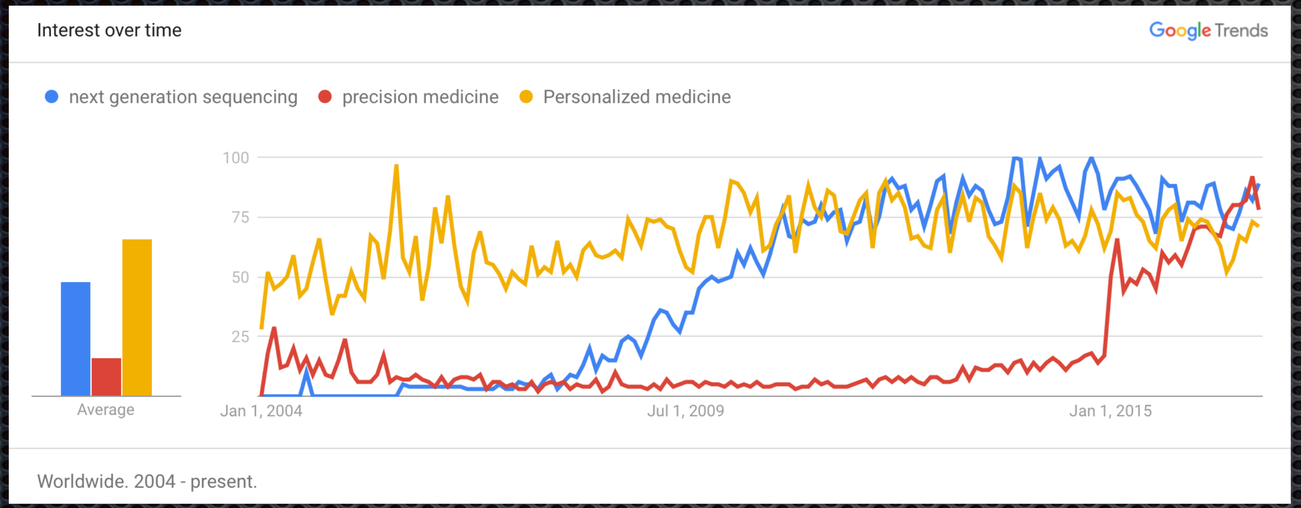
Universität Zürich

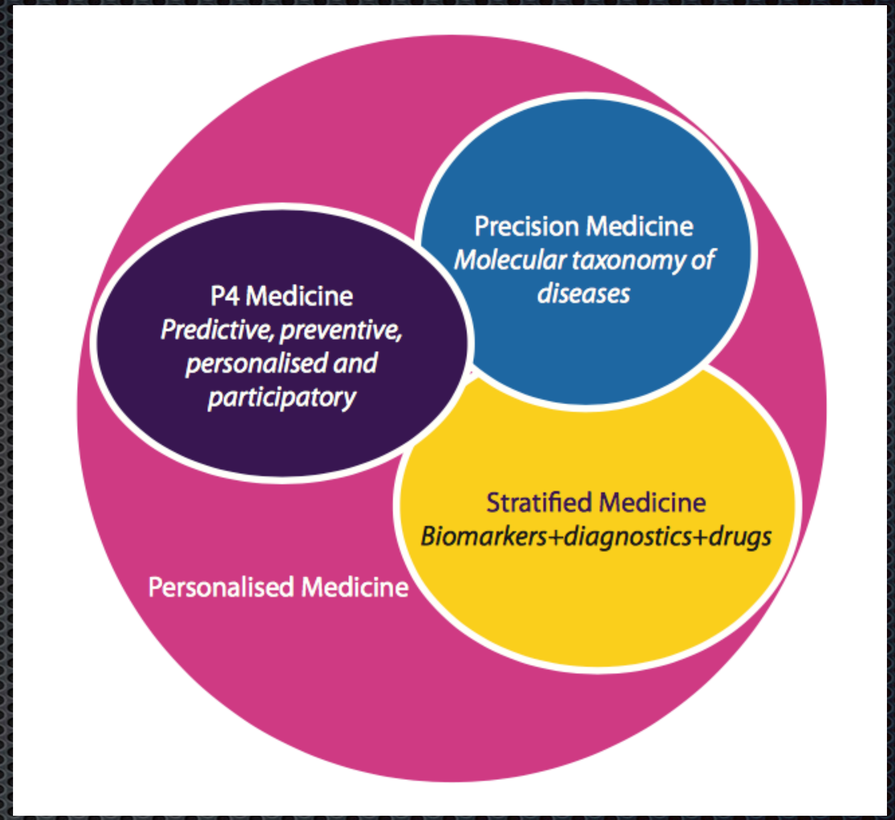
Zürich

Professor of bioinformatics @ IMLS (2015) | systematic assembly of oncogenomic data | databases and software tools | patterns in cancer genomes | *arraymap* online resource | GA4GH | SPHN

Many names for one concept or many concepts in one name?

Stratified, personalised, precision, individualised, P4 medicine or personalised healthcare – all are terms in use to describe notions often referred to as the future of medicine and healthcare. But what exactly is it all about, and are we all talking about the same thing?



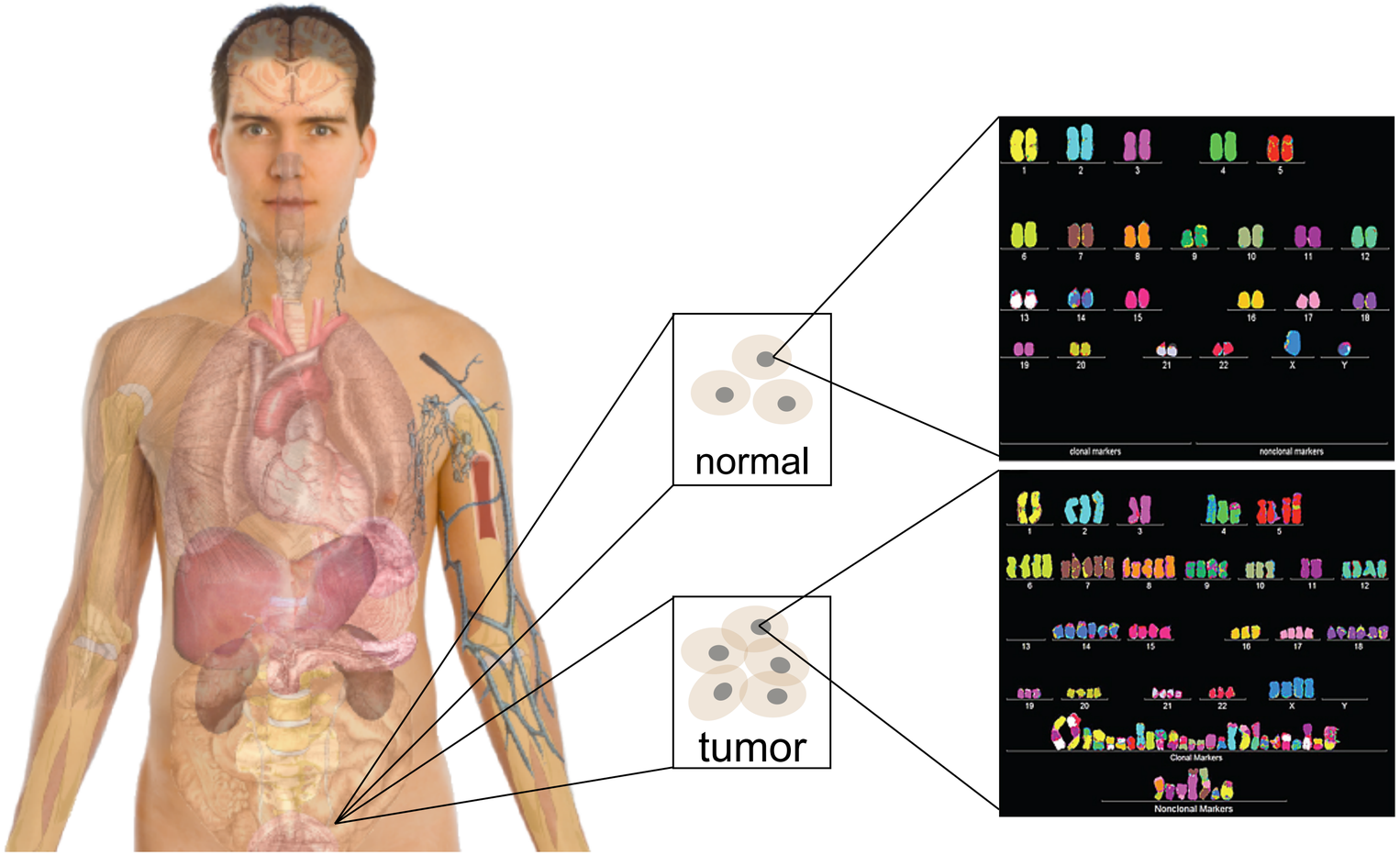


Source: PHG Foundation

While medicine has always been "personal" and "precise" in the given context of available knowledge and technologies, the concept of "**Personalised Medicine**" describes the use of individual genome information, concept based metadata and individually targeted therapies.

Personal Genomics as a Gateway into Biology

Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.



Genome analyses at the core of Personalized Health™

- Genome analyses (including transcriptome, metagenomics) are the **core technologies** for Personalized Health™ applications
- In the context of **academic medicine**, this requires
  - standard sample acquisition procedures & central **biobanking**
  - **core sequencing facility** (large throughput, cost efficiency, uniform sample and data handling procedures)
- secure **computing/analysis** platform
- Standardized **data formats** and **sample identification** procedures
- Metadata rich, reference **variant resource(s)** & expertise
- participation in reciprocal, international **data sharing** and **biocuration** efforts

Genome analyses at the core of Personalized Health™

Susceptibility, Pharmacogenomics, Classification, Infectious Diseases, Outcome Prediction, Lifestyle ...

doi:10.1038/nature19057

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2]*, Eric V. Minikel[1,2,5]*, Kaitlin E. Samocha[1,2,5,6]*, Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2], Daniel P. Birnbaum[2], Jack A. Kosmicki[1,2,6,13], Laramie E. Duncan[1,2,6], Karol Estrada[1,2], Fengmei Zhao[1,2], James Zou[2],

Barkur S. Shastry

SNP alleles in human disease and evolution

DISEASE MECHANISMS

Mechanisms underlying structural variant formation in genomic disorders

Claudia M. B. Carvalho[1,2] and James R. Lupski[1,3,4,5]

Abstract | With the recent burst of technological developments in genomics, and the clinical implementation of genome-wide assays, our understanding of the molecular basis of genomic disorders, specifically the contribution of structural variation to disease burden, is evolving

Rapid whole genome sequencing and precision neonatology

Joshua E. Petrikin, MD[a,], Laurel K. Willig, MD, FAAP[b], Laurie D. Smith, MD, PhD[c], and Stephen F. Kingsmore, MB, BAO, ChB, Dsc, FRCPath[d,e]*

insight progress

Cancer genetics

Bruce A. J. Ponder

Genomic Classification of Cutaneous Melanoma

The Cancer Genome Atlas Network[1,*,*,*]
[1]Cancer Genome Atlas Program Office, National Cancer Institute at NIH, 31 Center Drive, Bldg. 31, Suite 3A20, Bethesda, MD 20892, USA
*Correspondence: irwatson@mdanderson.org (I.R.W.), jgershen@mdanderson.org (J.E.G.), lchin@mdanderson.org (L.C.)
http://dx.doi.org/10.1016/j.cell.2015.05.044

Consequences of genomic diversity in *Mycobacterium tuberculosis*

Mireia Coscolla[a,b], Sebastien Gagneux[a,b,*]

[a] *Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Socinstrasse 57, 4002 Basel, Switzerland*
[b] *University of Basel, Petersplatz 1, Basel 4003, Switzerland*

doi:10.1111/pcn.12128

Psychiatry and Clinical Neurosciences

PCN Frontier Review

Copy-number variation in the pathogenesis of autism spectrum disorder

Emiko Shishido, PhD[1,2,3], Branko Aleksic, MD, PhD[3] and Norio Ozaki, MD, PhD[3]*

he Promotion of Science, Japan

Common gene variants, mortality and extreme longevity in humans

B.T. Heijmans[a,b], R.G.J. Westendorp[b], P.E. Slagboom[a,*]

RESEARCH ARTICLE Open Access

Integrative genome-wide expression profiling identifies three distinct molecular subgroups of renal cell carcinoma with different patient utcome

nfred Beleut[1,5*], Philip Zimmermann[2], Michael Baudis[3], Nicole Bruni[4], Peter Bühlmann[4], Oliver Laule[2], -Duc Luu[1], Wilhelm Gruissem[2], Peter Schraml[1*] and Holger Moch[1]

NEURODEVELOPMENT

Genes, circuits, and precision therapies for autism and related neurodevelopmental disorders

Mustafa Sahin* and Mriganka Sur*

RESEARCH ARTICLE Open Access

Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens

Haoyang Cai[1,2], Nitin Kumar[1,2], Homayoun C Bagheri[3], Christian von Mering[1,2], Mark D Robinson[1,2]* and Michael Baudis[1,2]*

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib

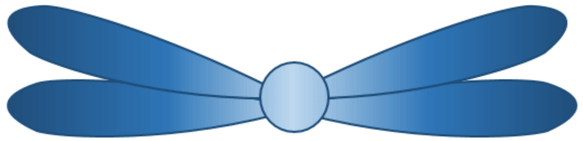
Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.
N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

The landscape of somatic copy-number alteration across human cancers

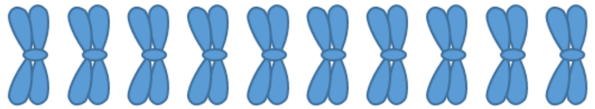
Rameen Beroukhim[1,3,4,5]*, Craig H. Mermel[1,3]*, Dale Porter[8], Guo Wei[1], Soumya Raychaudhuri[1,4], Jerry Donovan[8], Jordi Barretina[1,3], Jesse S. Boehm[1], Jennifer Dobson[1,3], Mitsuyoshi Urashima[9], Kevin T. Mc Henry[8], Reid M. Pinchback[1], Azra H. Ligon[4], Yoon-Jae Cho[6], Leila Haery[1,3], Heidi Greulich[1,3,4,5], Michael Reich[1], Wendy Winckler[1], Michael S. Lawrence[1], Barbara A. Weir[1,3], Kumiko E. Tanaka[1,3], Derek Y. Chiang[1,3,13], Adam J. Bass[1,3,4], Alice Loo[8], Carter Hoffman[1,3], John Prensner[1,3], Ted Liefeld[1], Qing Gao[1], Derek Yecies[3], Sabina Signoretti[3,4], Elizabeth Maher[10], Frederic J. Kaye[11], Hidefumi Sasaki[12], Joel E. Tepper[13], Jonathan A. Fletcher[4], Josep Tabernero[14], José Baselga[14], Ming-Sound Tsao[15], Francesca Demichelis[16], Mark A. Rubin[16], Pasi A. Janne[3,4], Mark J. Daly[1,17], Carmelo Nucera[7], Ross L. Levine[18], Benjamin L. Ebert[1,4,5], Stacey Gabriel[1], Anil K. Rustgi[19], Cristina R. Antonescu[18], Marc Ladanyi[18], Anthony Letai[3], Levi A. Garraway[1,3], Massimo Loda[3,4], David G. Beer[20], Lawrence D. True[21], Aikou Okamoto[22], Scott L. Pomeroy[6], Samuel Singer[18], Todd R. Golub[1,3,23], Eric S. Lander[1,2,5], Gad Getz[1], William R. Sellers[8] & Matthew Meyerson[1,3,5]

Genome Sequencing

whole genome sequencing (WGS)

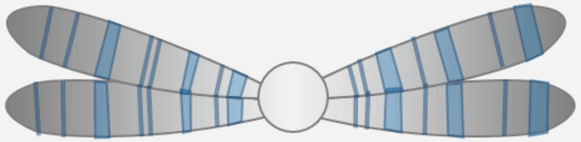


100%



all DNA
(3.1 billion base pairs)

exome sequencing



~1%

protein-coding DNA only
(~31 million base pairs)

What does it cost to sequence a genome?

	Human Genome Project (HGP): 1991-2003	today: 2017
cost:	$2.7 billion	~$1,500
time:	12+ years	< 2 days

today:
2017

~$530
~3 days

Ryan L. Collins & Brad Wierbowski (2017)

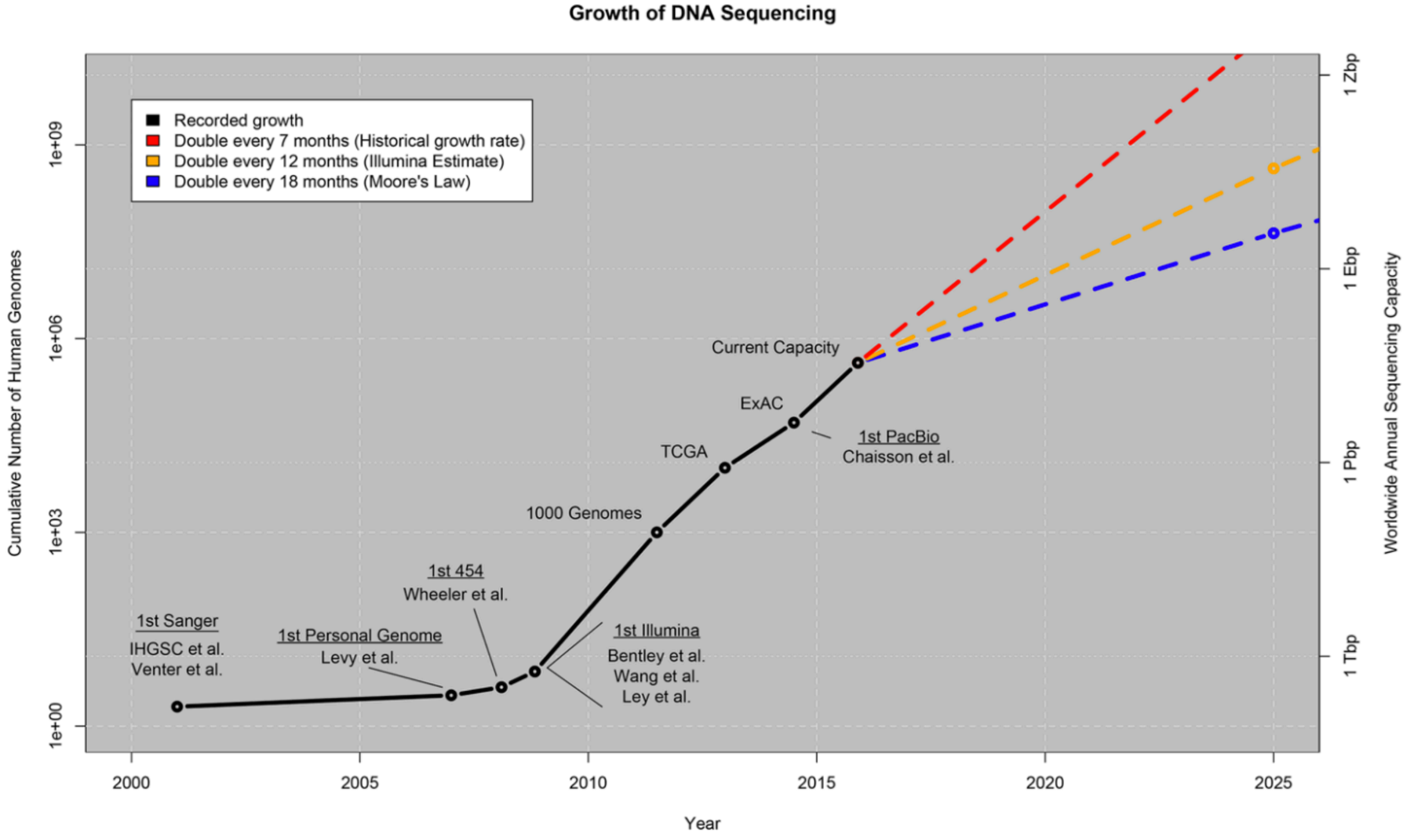


Fig 1. Growth of DNA sequencing. The plot shows the growth of DNA sequencing both in the total number of human genomes sequenced (left axis) as well as the worldwide annual sequencing capacity (right axis: Tera-basepairs (Tbp), Peta-basepairs (Pbp), Exa-basepairs (Ebp), Zetta-basepairs (Zbps)). The values through 2015 are based on the historical publication record, with selected milestones in sequencing (first Sanger through first PacBio human genome published) as well as three exemplar projects using large-scale sequencing: the 1000 Genomes Project, aggregating hundreds of human genomes by 2012 [3]; The Cancer Genome Atlas (TCGA), aggregating over several thousand tumor/normal genome pairs [4]; and the Exome Aggregation Consortium (ExAC), aggregating over 60,000 human exomes [5]. Many of the genomes sequenced to date have been whole exome rather than whole genome, but we expect the ratio to be increasingly favored towards whole genome in the future. The values beyond 2015 represent our projection under three possible growth curves as described in the main text.

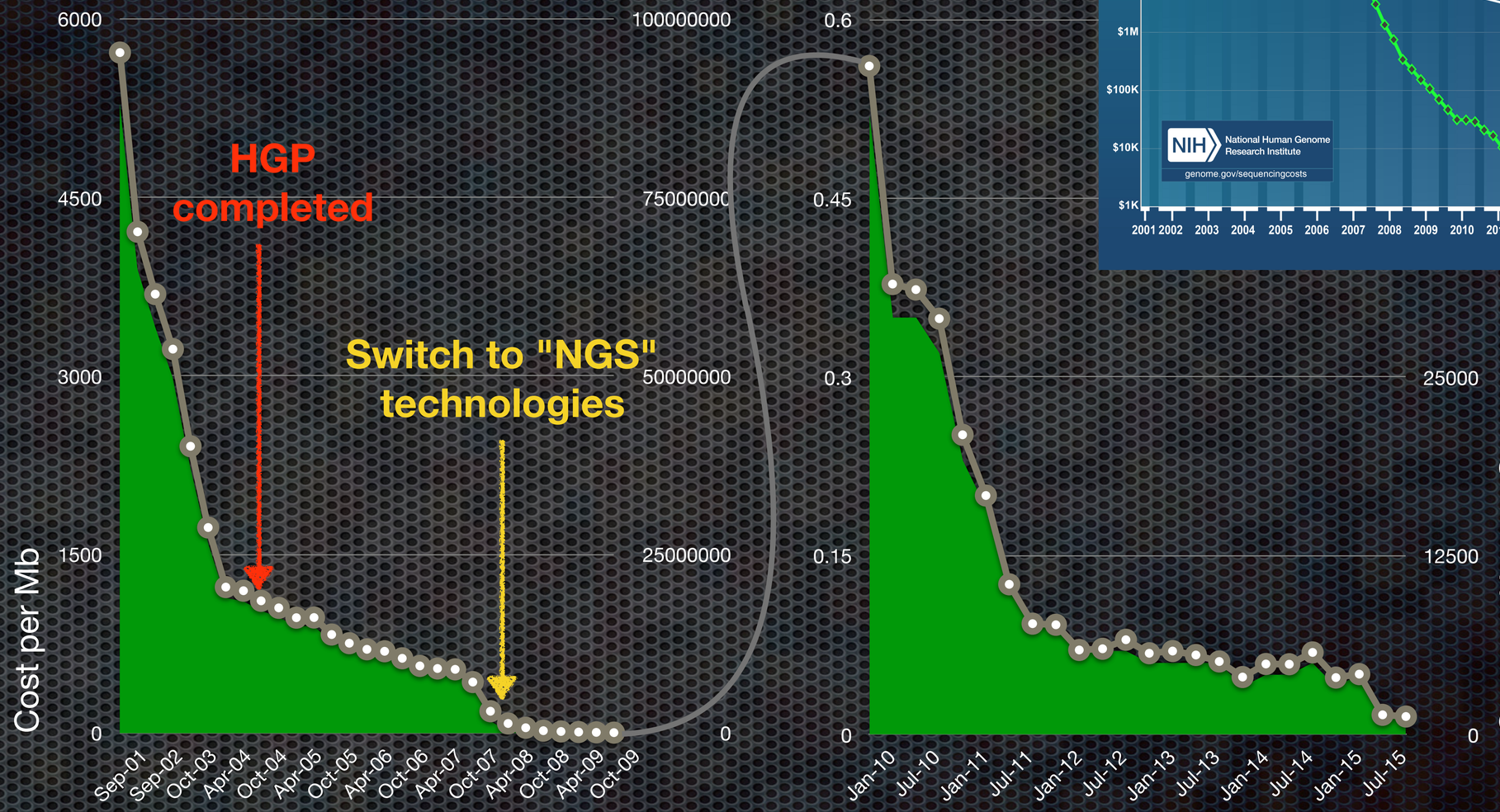
Citation: Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195

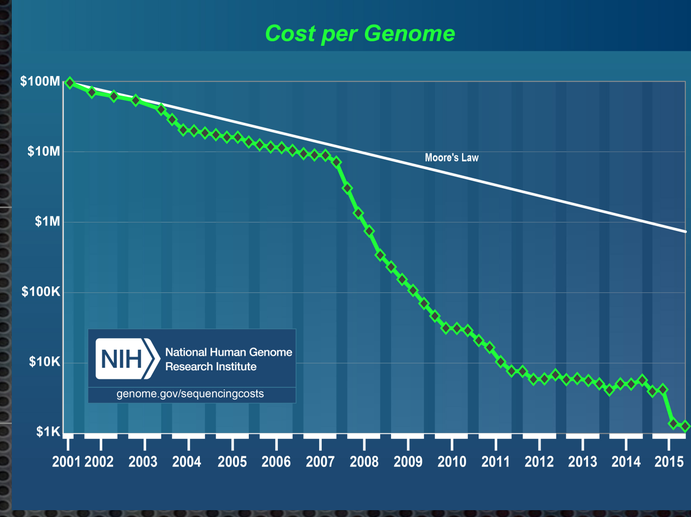
DNA Sequencing Costs: Data

Data from the NHGRI Genome Sequencing Program (GSP)

https://www.genome.gov/sequencingcostsdata/

Date	Cost per Mb	Cost per Genome
Sep-01	JS$ 5'292.39	US$ 95'263'072
Mar-02	JS$ 3'898.64	US$ 70'175'437
Sep-02	JS$ 3'413.80	US$ 61'448'422
Mar-03	JS$ 2'986.20	US$ 53'751'684
Oct-03	JS$ 2'230.98	US$ 40'157'554
Jan-04	JS$ 1'598.91	US$ 28'780'376
Apr-04	JS$ 1'135.70	US$ 20'442'576
Jul-04	JS$ 1'107.46	US$ 19'934'346
Oct-04	JS$ 1'028.85	US$ 18'519'312
Jan-05	US$ 974.16	US$ 17'534'970
Apr-05	US$ 897.76	US$ 16'159'699
Jul-05	US$ 898.90	US$ 16'180'224
Oct-05	US$ 766.73	US$ 13'801'124
Jan-06	US$ 699.20	US$ 12'585'659
Apr-06	US$ 651.81	US$ 11'732'535
Jul-06	US$ 636.41	US$ 11'455'315
Oct-06	US$ 581.92	US$ 10'474'556
Jan-07	US$ 522.71	US$ 9'408'739
Apr-07	US$ 502.61	US$ 9'047'003
Jul-07	US$ 495.96	US$ 8'927'342
Oct-07	US$ 397.09	US$ 7'147'571
Jan-08	US$ 102.13	US$ 3'063'820
Apr-08	US$ 15.03	US$ 1'352'982
Jul-08	US$ 8.36	US$ 752'080
Oct-08	US$ 3.81	US$ 342'502
Jan-09	US$ 2.59	US$ 232'735
Apr-09	US$ 1.72	US$ 154'714
Jul-09	US$ 1.20	US$ 108'065
Oct-09	US$ 0.78	US$ 70'333
Jan-10	US$ 0.52	US$ 46'774
Apr-10	US$ 0.35	US$ 31'512
Jul-10	US$ 0.35	US$ 31'125
Oct-10	US$ 0.32	US$ 29'092
Jan-11	US$ 0.23	US$ 20'963
Apr-11	US$ 0.19	US$ 16'712
Jul-11	US$ 0.12	US$ 10'497
Oct-11	US$ 0.09	US$ 7'743
Jan-12	US$ 0.09	US$ 7'666
Apr-12	US$ 0.07	US$ 5'901
Jul-12	US$ 0.07	US$ 5'985
Oct-12	US$ 0.07	US$ 6'618
Jan-13	US$ 0.06	US$ 5'671
Apr-13	US$ 0.06	US$ 5'826
Jul-13	US$ 0.06	US$ 5'550
Oct-13	US$ 0.06	US$ 5'096
Jan-14	US$ 0.04	US$ 4'008
Apr-14	US$ 0.05	US$ 4'920
Jul-14	US$ 0.05	US$ 4'905
Oct-14	US$ 0.06	US$ 5'731
Jan-15	US$ 0.04	US$ 3'970
Apr-15	US$ 0.05	US$ 4'211
Jul-15	US$ 0.015	US$ 1'363
Oct-15	US$ 0.014	US$ 1'245

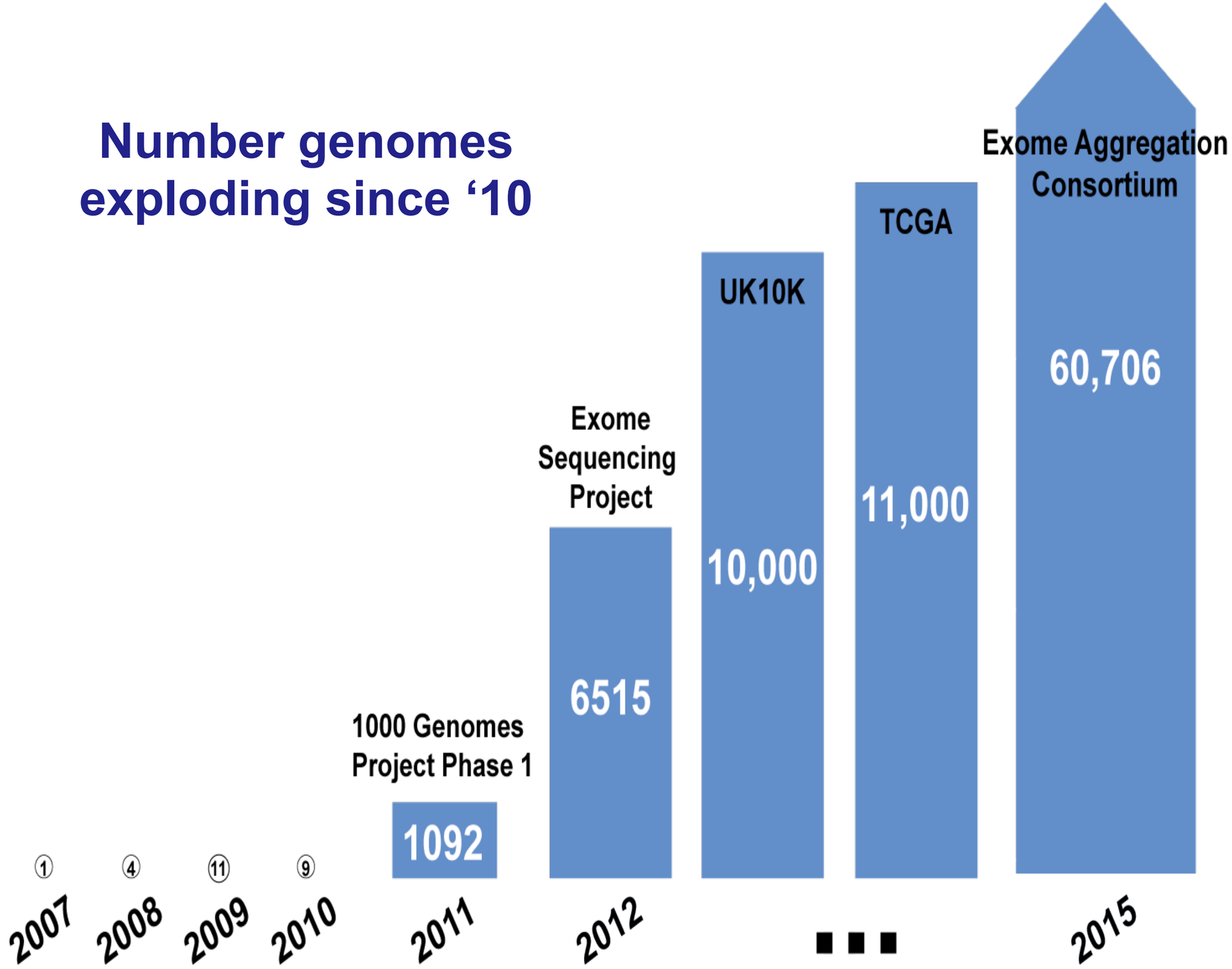




- Labor, administration, management, utilities, reagents, and consumables
- Sequencing instruments and other large equipment (amortized over three years)
- Informatics activities directly related to sequence production (e.g., laboratory information management systems and initial data processing)
- Submission of data to a public database
- Indirect Costs (http://oamp.od.nih.gov/dfas/faq/indirect-costs#difference) as they relate to the above items

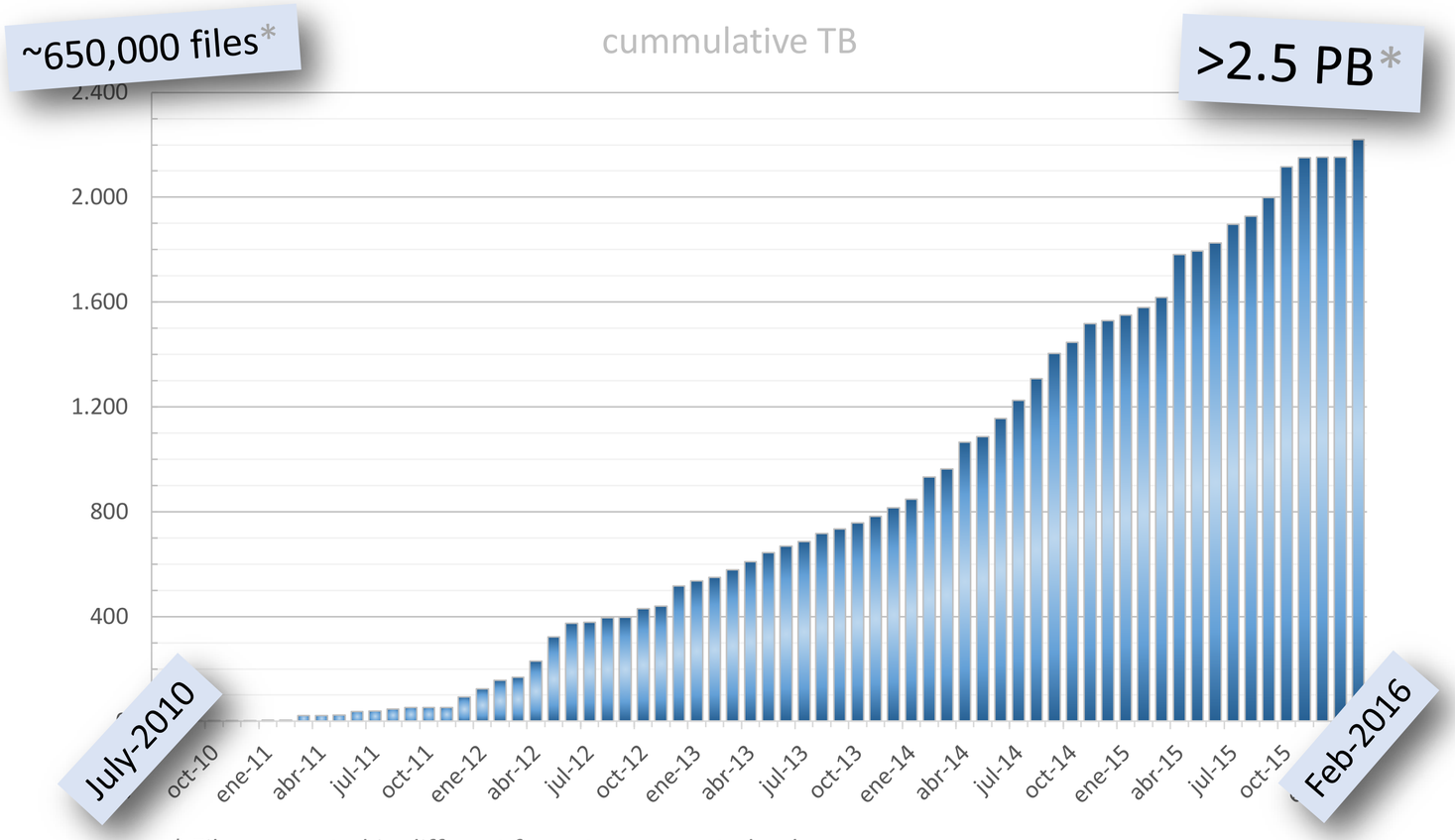
Slides freely downloadable from lectures.GersteinLab.org & “tweetable” (via @markgerstein)

Number genomes exploding since ‘10



Growth of Genome Data Repositories: Example EGA

The EGA contains a growing amount of data



International Cancer Genome Consortium

IRDiRC
INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

UK 10K

wellcome trust sanger institute

** Files encrypted in different formats are counted only once*

european genome-phenome archive

Source: Jordi Rambla (CRG Barcelona) 2016

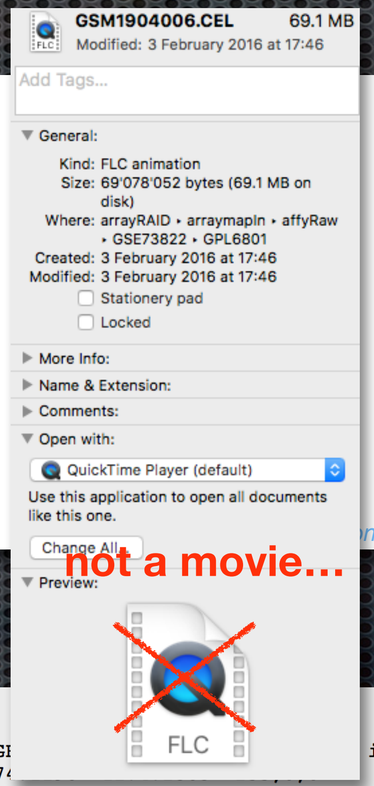
What is a PB, for human genomes? It depends.

- 2 bits per base are sufficient to encode TCGA
  - using 00, 01, 10, 11
  - [TCGA]{3’000’000’000}
  - 2 * 3 * 10^9b = 6, 000, 000, 000 b
  - perfect genome (no overhead): ~715 MB
  - 1PB => ~1400000 genomes
- according to Swiss online store (Sep 2017) ~45'000CHF (100x10TB disks)
- this is less than a PhD position per year in Switzerland …
- (real costs are 2x that, + duplication, facilities, service … => ~500’000CHF)
- **However: A single 30x BAM file => 100GB**
- Still: 500’000CHF => 1PB => 10’000 genomes => 50CHF/genome (BAM format)

Bioinformatics: **File Formats**

- **text** or **binary** file formats, optimised for specific types of biological data
- examples from genomics:
  - **BAM** - compressed binary version of Sequence Alignment/ Map (SAM)
  - **BED** (Browser Extensible Data) - flexible way to define the data lines in an genome browser annotation tracks
  - **VCF** (Variant Call Format)

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- barChart and bigBarChart format
- bigBed format
- bigGenePred table format
- bigPsl table format
- bigMaf table format
- bigChain table format
- bigWig format
- Chain format



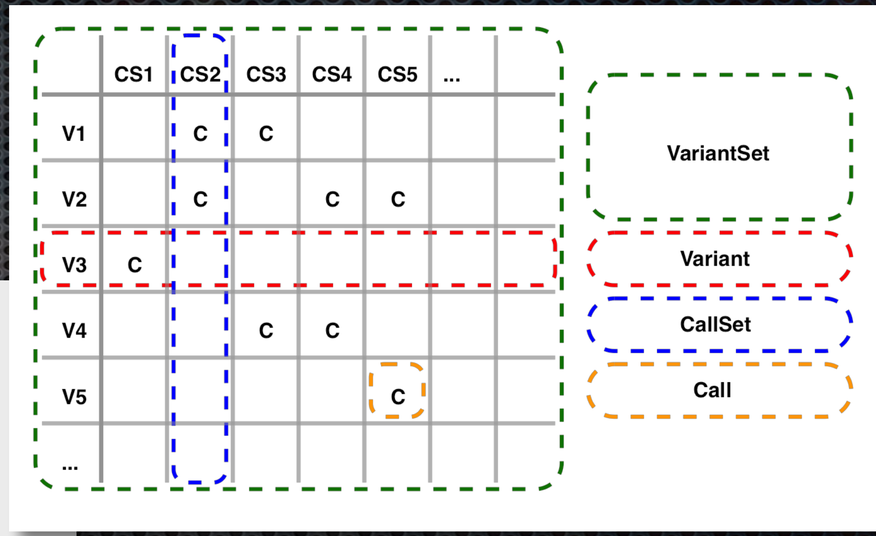
- CRAM format
- GenePred table format
- GFF format
- GTF format
- HAL format
- MAF format
- Microarray format
- Net format
- Personal Genome SNP format
- PSL format
- VCF format
- WIG format

me.ucsc.edu/FAQ/FAQformat.html

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB                                    itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127                                    ,
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

BED file example

The VCF file format
Standard for variant representation



Example

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0|1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1|0:77	1/1:95
1	100		T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Deletion

SNP

Large SV

Insertion

Other event

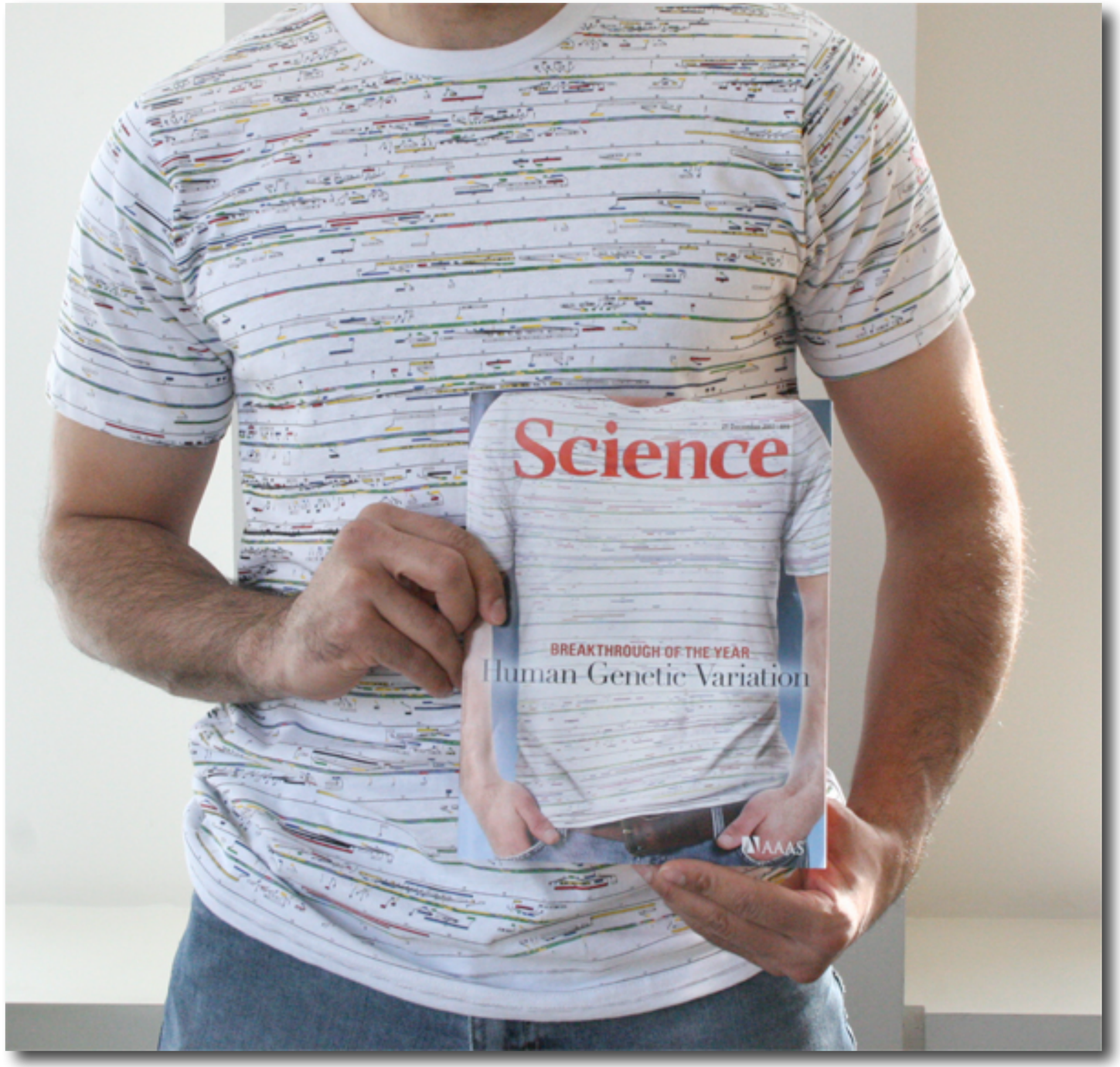
Phased data (G and C above are on the same chromosome)

Variant
Call
Format

- stores the results of a single or multiple interpretations of genome sequencing datasets, in comparison to a reference genome
- standard format for file-based storage of human genome variants

Global Alliance for Genomics & Health

The trouble with human genome variation



Conclusions from the analysis of variation in the human genome

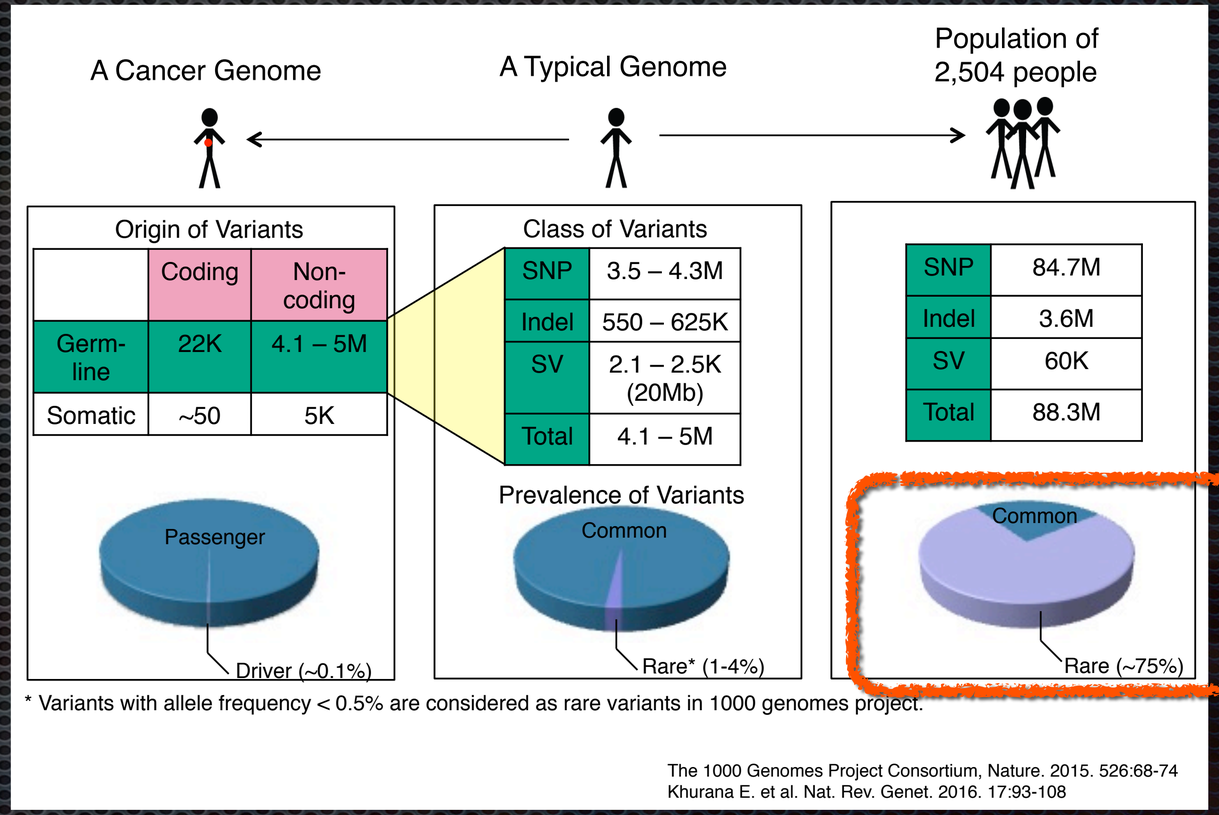
- 1. Humans are all very similar to each other
  - Two humans will show about 99.9% sequence identity with each other. In other words, only about 1 in 1’000 bp is different between two individuals.
  - Humans show about 98% sequence identity to chimps. So two humans are still much more similar to each other than either is to the monkey.

- 2. Humans are very different from each other
  - Two typical humans will likely have over 1’000’000 independent sequence differences in their genomes.

Slide from Michael Hengartner’s Human Genome Lecture

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

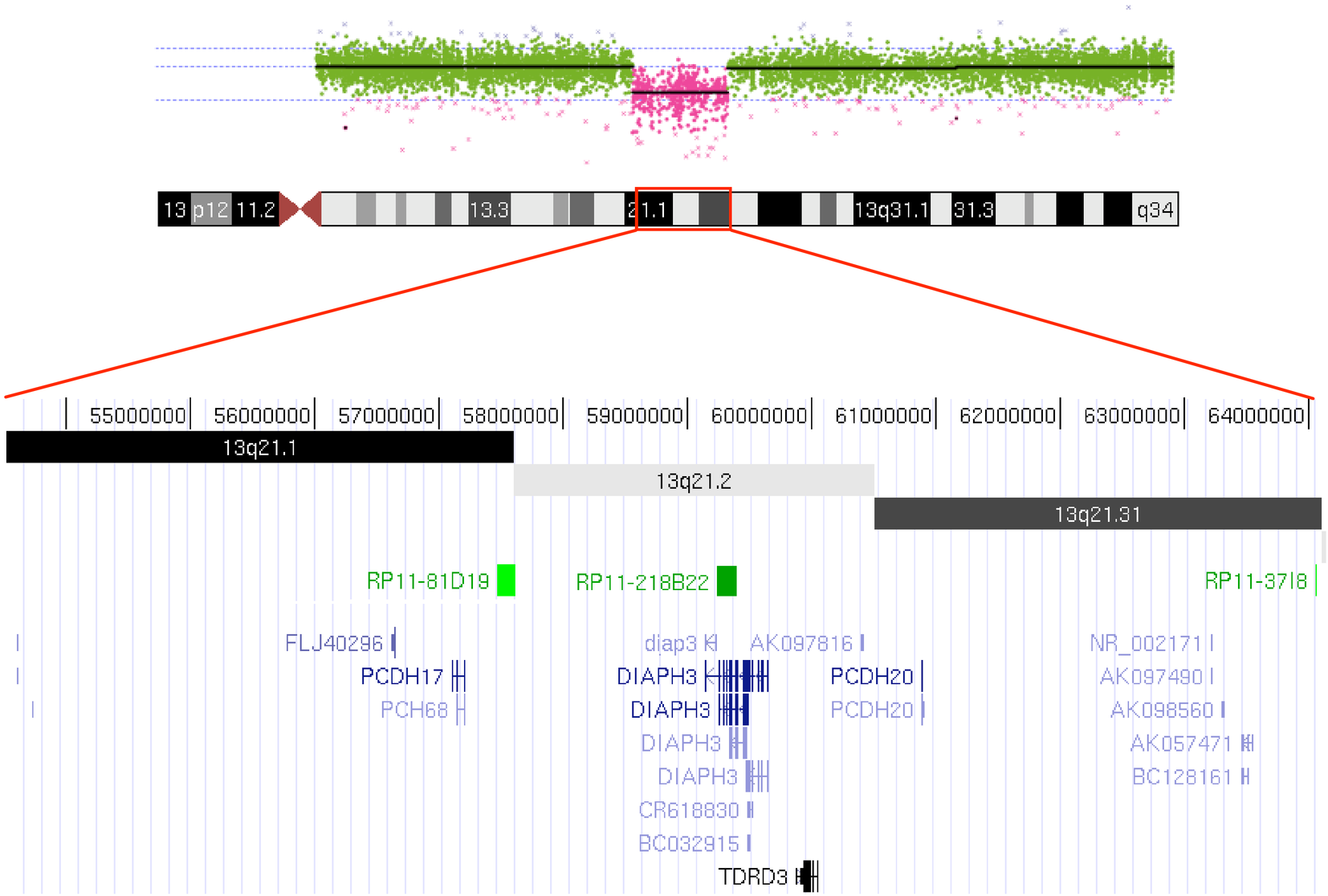


The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

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

Nobody is perfect (?)

A 10.7 Mb Interstitial Deletion of 13q21 Without Phenotypic Effect Defines a Further Non-Pathogenic Euchromatic Variant
Andreas Roos, Miriam Elbracht, Michael Baudis, Jan Senderek, Nadine Schönherr, Thomas Eggermann, and Herdit M. Schüler
American Journal of Medical Genetics Part A 146A:2417 – 2420 (2008)



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

Reference Resources for Human Genome Variants

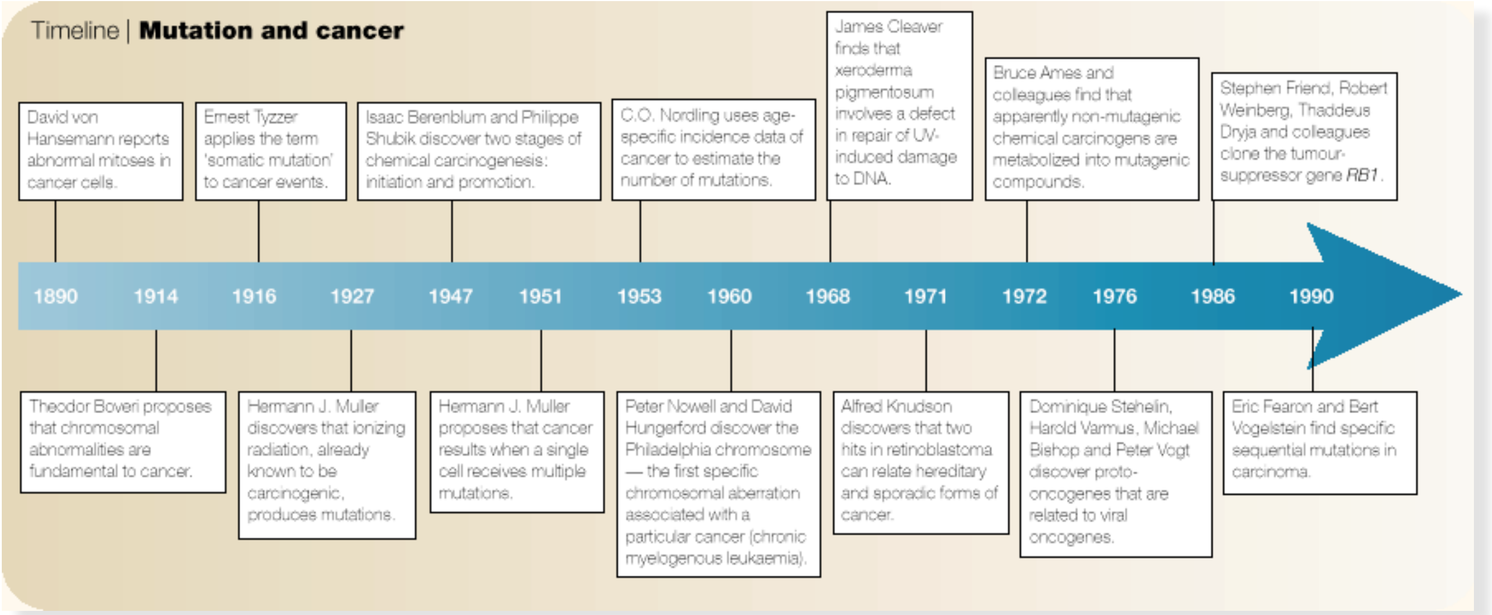
- NCBI:dbSNP
  - single nucleotide polymorphisms (SNPs) and multiple small-scale variations
  - including insertions/deletions, microsatellites, non-polymorphic variants
- NCBI:dbVAR
  - genomic structural variation
  - insertions, deletions, duplications, inversions, multinucleotide substitutions, mobile element insertions, translocations, complex chromosomal rearrangements
- NCBI:ClinVar
  - aggregates information about genomic variation and its relationship to human health
- EMBL-EBI:EVA
  - open-access database of all types of genetic variation data from all species
- Ensembl
  - portal for many things genomic...

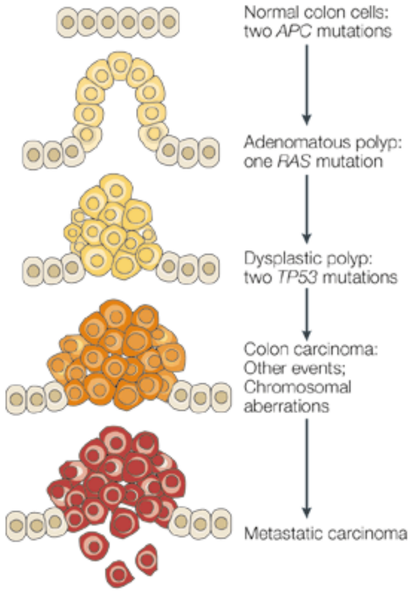
Personalized medicine in cancer

Data Formats | Genome Variation | Techniques | Resources | Sharing

Michael Baudis **UZH SIB**
Computational Oncogenomics

University of Zurich[UZH]



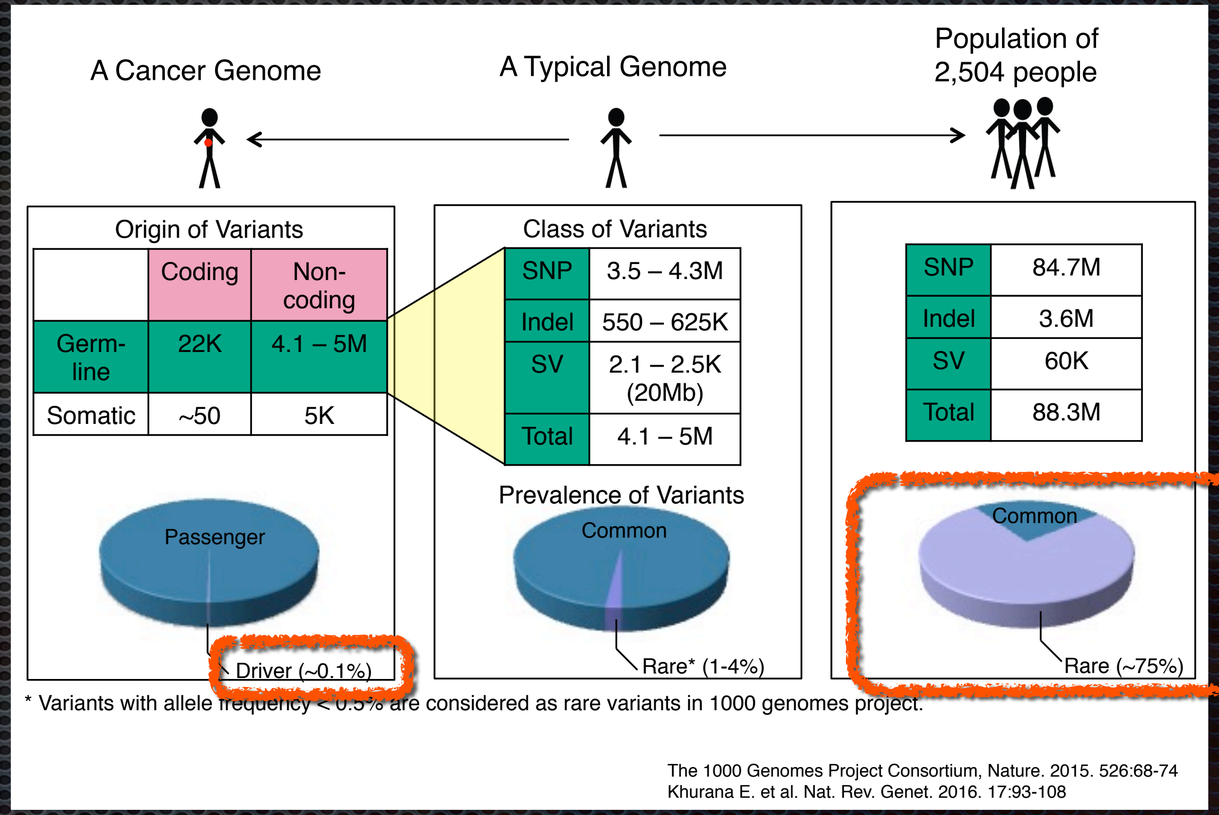


Cancers are based on acquired and inherited genomic mutations

Knudson, A. G. (2001). Two genetic hits (more or less) to cancer. Nature Reviews Cancer, 1(2), 157–162.

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

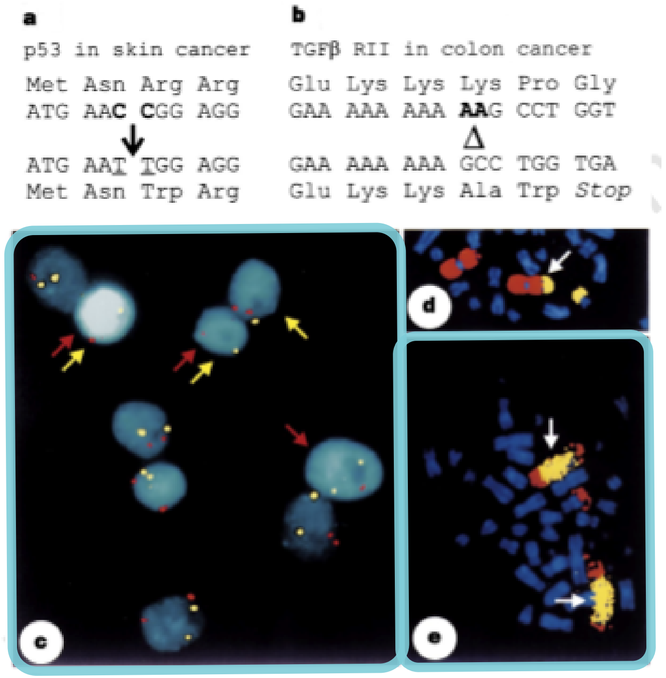
- a typical human genome (~3 billion base pairs) has ~5 million variants
- most of them are "**rare**"; i.e. can only be identified as recurring when sequencing thousands of people
- cancer cells accumulate additional variants, only **few** of which ("**drivers**") are relevant for the disease



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

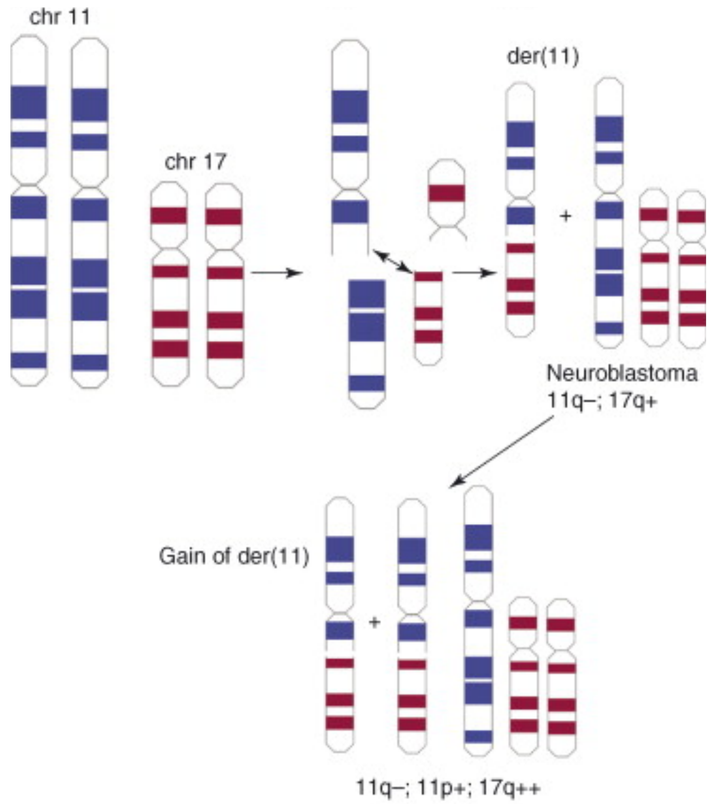
Graphic adapted from Mark Gerstein (GersteinLab.org; @markgerstein)

Mutations & genomic rearrangements in cancer



Lengauer et al. Genetic instabilities in human cancers. Nature (1998) vol. 396 (6712) pp. 643-9

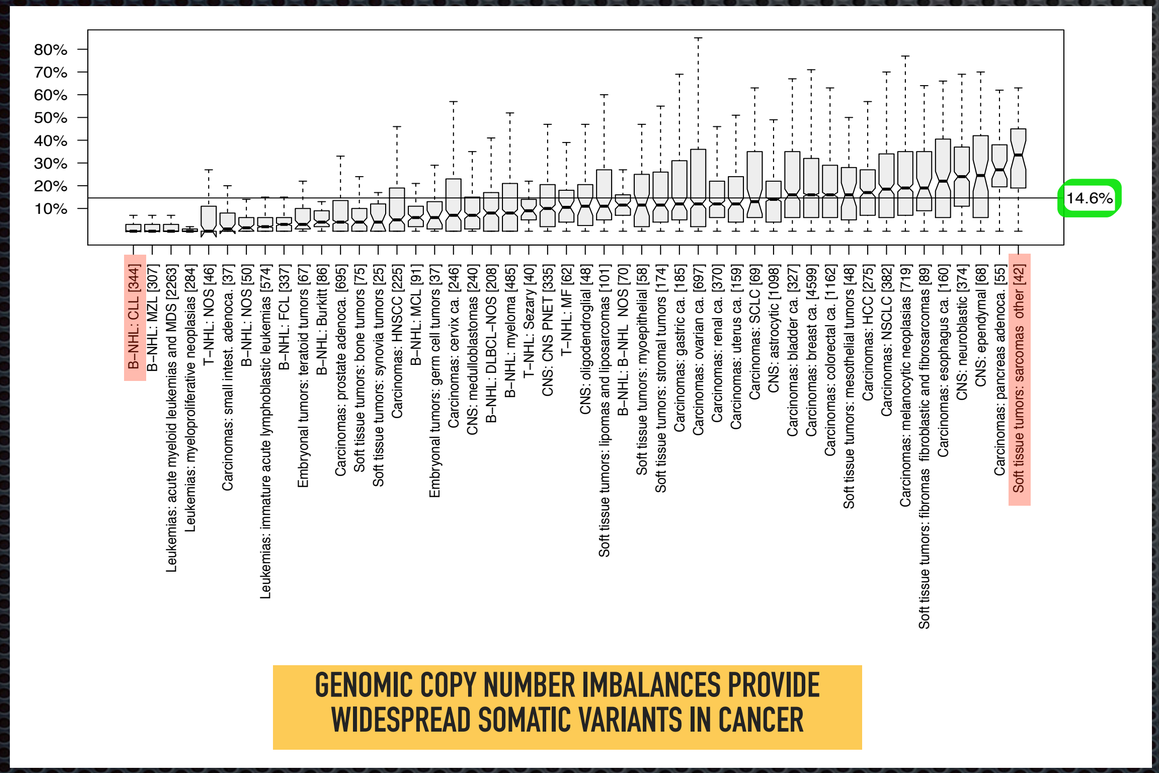
a. small mutation (di-pyrimidine exchange at p53 in Xeroderma pigmentosum patient)
b. two-base deletion in *TGFB* in a colorectal cancer patient with mismatch repair deficiency
c. chromosomal losses (FISH; red=3, yellow=12) in CRC
d. t(1;17) in neuroblastoma, whole-chromosomal painting
e. *MYCN* gene amplification (multiple copies inserted into chromosome 1 derived marker)



Generation of copy number imbalances in cancer through imbalanced cytogenetic rearrangements - partial deletion of 11q, gain of 11pterq21 and 2 addl. copies of 17q

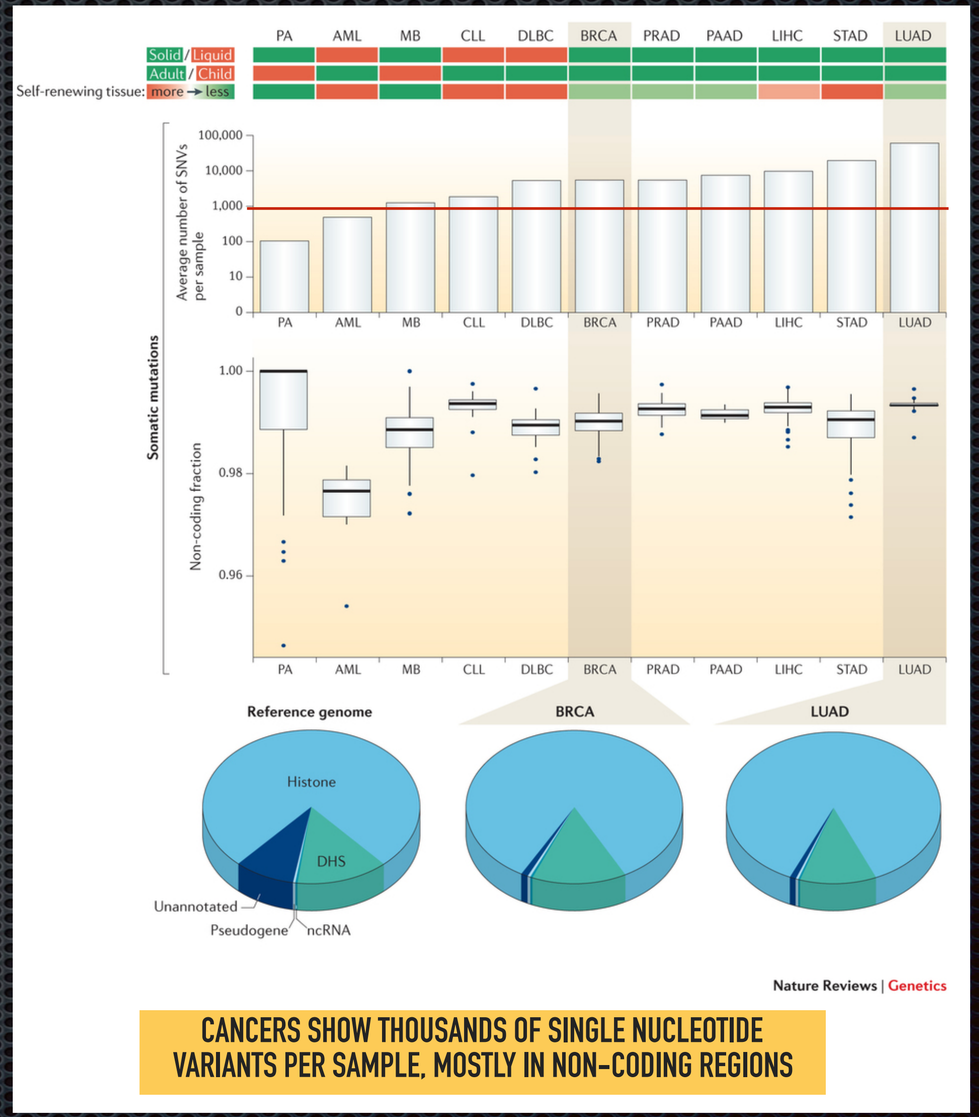
RL Stallings: Are chromosomal imbalances important in cancer? Volume 23, Issue 6, p278–283, 2007

Quantifying Somatic Mutations In Cancer



GENOMIC COPY NUMBER IMBALANCES PROVIDE WIDESPREAD SOMATIC VARIANTS IN CANCER

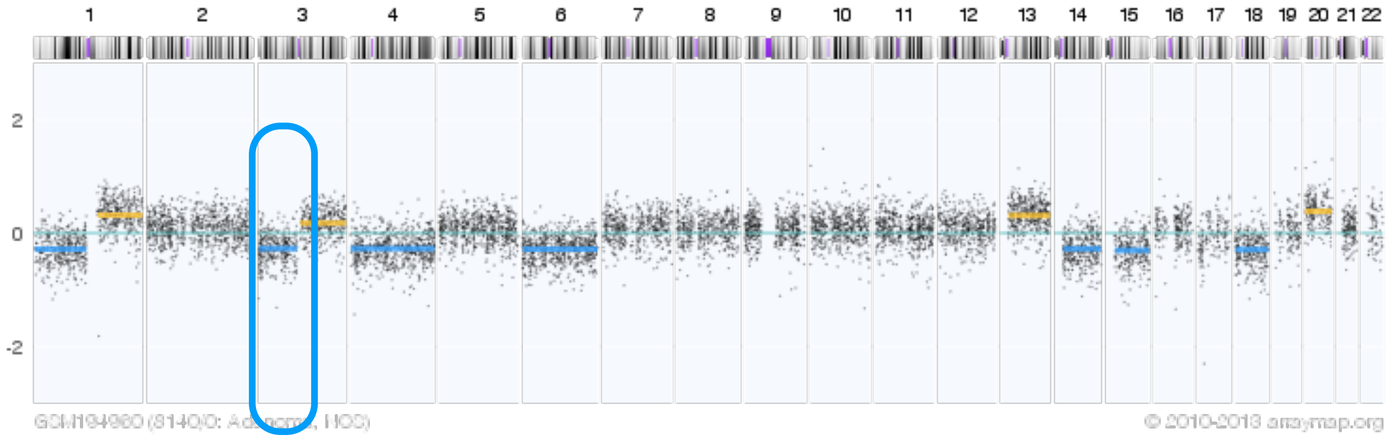
On average ~15% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on >30'000 cancer genomes from arraymap.org



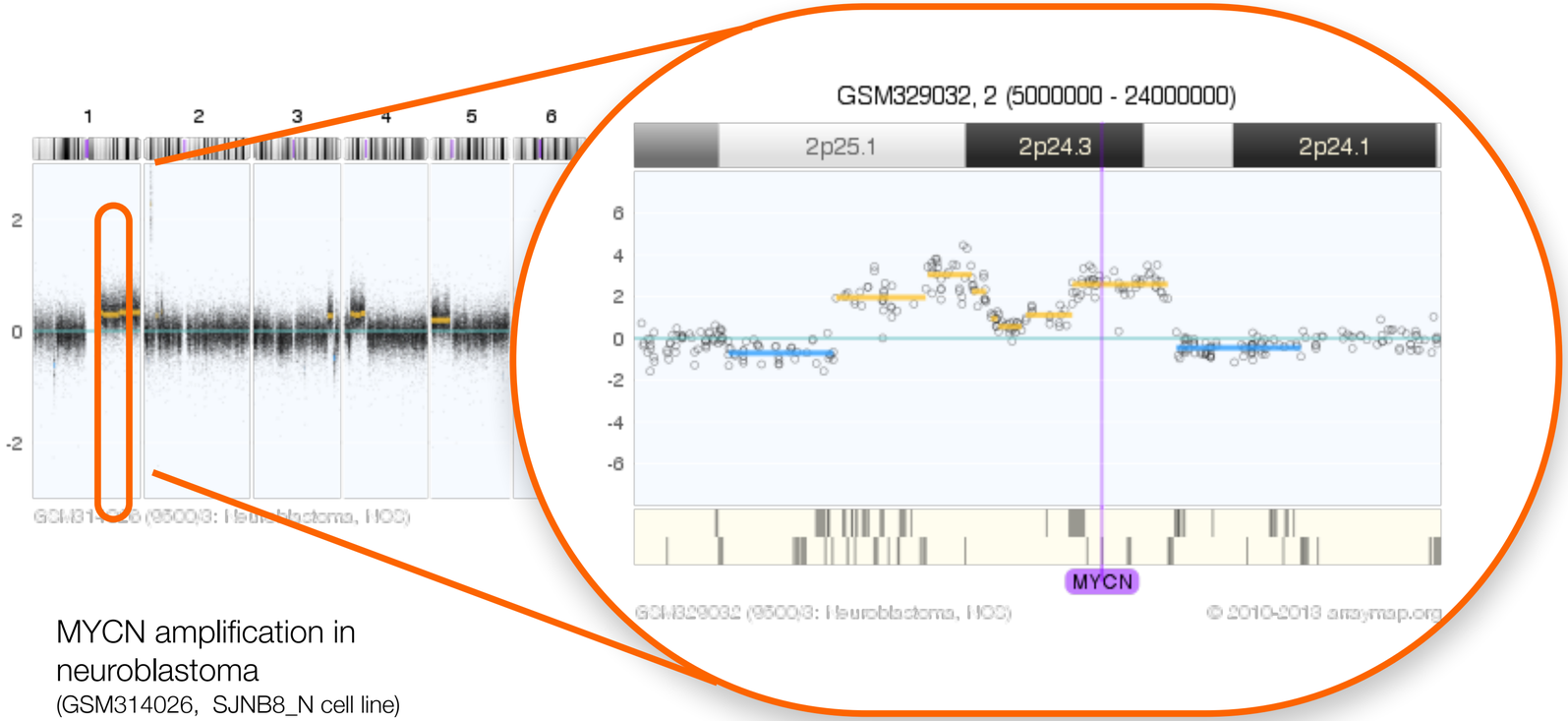
CANCERS SHOW THOUSANDS OF SINGLE NUCLEOTIDE VARIANTS PER SAMPLE, MOSTLY IN NON-CODING REGIONS

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)

Genomic arrays: Many probes + bioinformatics determine copy number aberrations



Gain of chromosome arm 3q in colorectal carcinoma

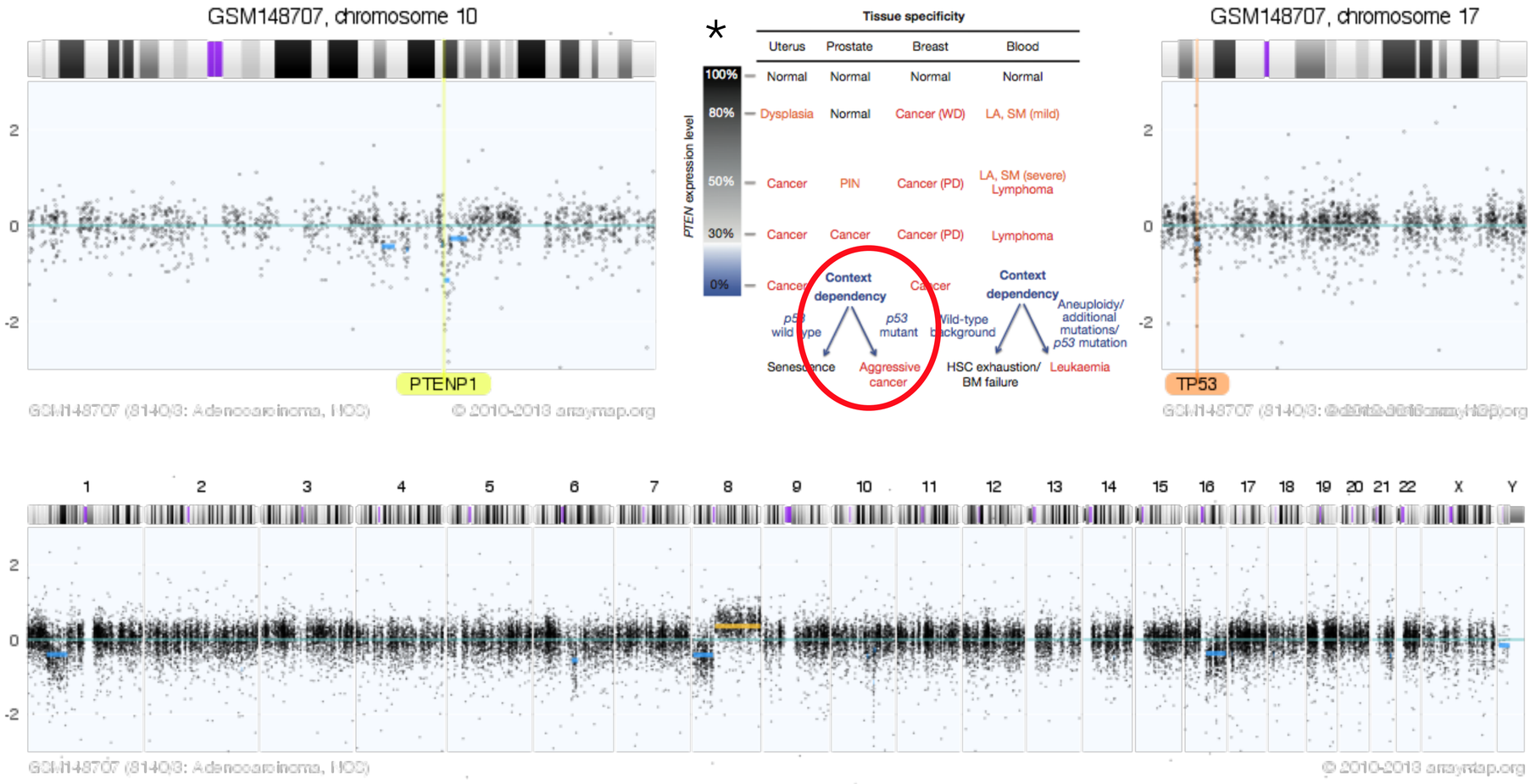


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

low level/**high level** copy number alterations (CNAs)

arrayMap

Gene dosage phenomena beyond simple on/off effects



Combined heterozygous deletions involving *PTEN* and TP53 loci in a case of prostate adenocarcinoma
(GSM148707, PMID 17875689, Lapointe *et al*., CancRes 2007)

* A. H. Berger, A. G. Knudson, and P. P. Pandolfi, “A continuum model for tumour suppression,” *Nature*, vol. 476, no. 7359, pp. 163–169, Aug. 2011.

arrayMap

Challenges in aCGH data collection: Constitutional CNVs vs. imbalances

Segmental copy number variations of unique sequences have been described as normal feature of human DNA.
These CNVs may be up to 2-3 megabases in size, and can involve coding regions.

Everybody has them ...

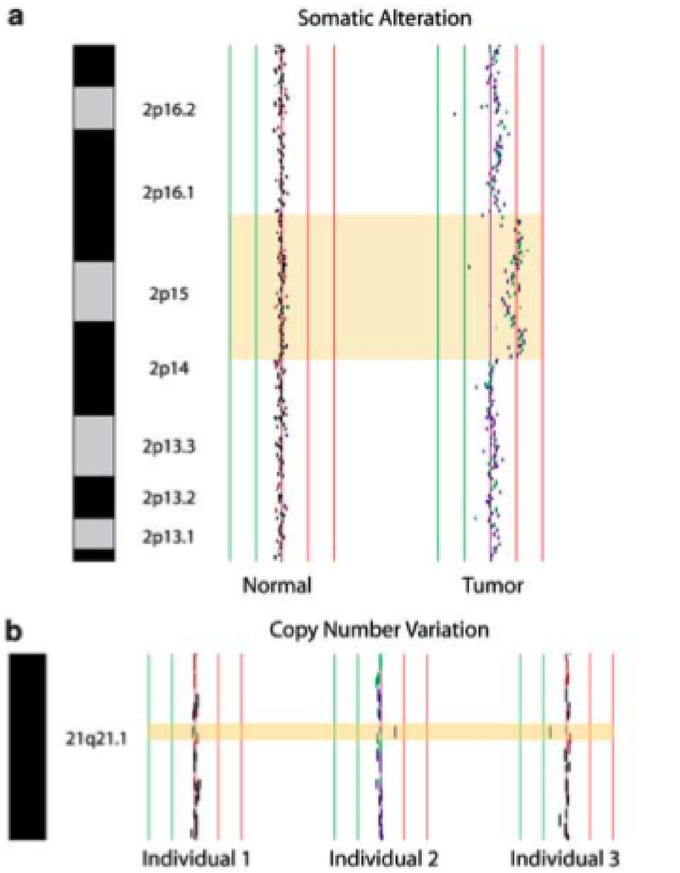


Figure 3 Somatic alterations and copy number variations. (**a**) Example of a segmental duplication observed at chromosome arm 2p present in the cancer cells but absent in the normal cells from the same individual. Each black dot represents a single BAC clone spotted on the array. The purple line represents equal fluorescent intensity ratio between sample and reference. Copy number gain (and loss) shifts the ratio to the right (and left). (**b**) Illustrates a copy number variation observed at chromosomal region 21q21.1. Three normal individuals exhibit equal, more and fewer copies relative to the reference DNA, indicating variation in the population.

Lockwood et al. Recent advances in array comparative genomic hybridization technologies and their applications in human genetics.
Eur J Hum Genet (2006) vol. 14 (2) pp. 139-48

CNV vs. CNA: Size matters

- no unambiguous criterium for CNV vs. CNA
- statistic argument: CNVs are recurring copy number variations found in the germline DNA of “healthy” individuals
- size argument: CNVs are rarely larger than 1Mb

Perry et al. The Fine-Scale and Complex Architecture of Human Copy-Number Variation. Am J Hum Genet (2008)

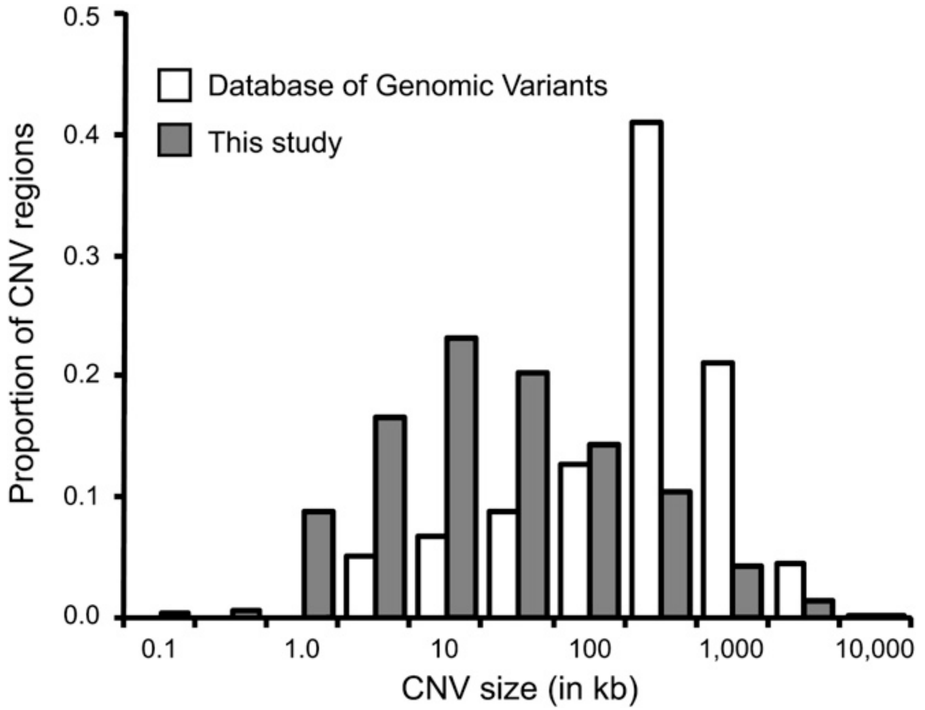


Figure 1. Size Distribution of CNVs from the Database of Genomic Variants, with Corresponding CNVs from This Study
We identified CNVs in at least one individual for 1153 of 2191 putative CNV regions annotated in the Database of Genomic Variants (DGV) as of 30 November 2006. Size distributions for these regions are shown in log scale, with 10-fold multiples of 1 and $\sqrt{10}$, based on the size of each region from DGV and the estimates from our study of the total amount of copy-number-variable sequence within and overlapping the DGV-defined region. Our estimates were smaller than the corresponding DGV region for 1020 of the 1153 loci (88%) and smaller by more than 50% for 876 regions (76%).

Where to find cancer genome data ...

RESOURCES FOR CANCER GENOMICS

COSMIC
Catalogue of somatic mutations in cancer

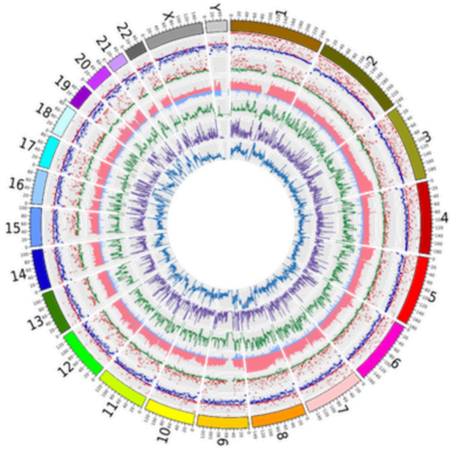
Home ▼ | Resources ▼ | Curation ▼ | Tools ▼ | Data ▼ | News ▼ | Help ▼ | About ▼ | Search COSMIC... | Login ▼

COSMIC v79, released 14-NOV-16

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below, or by browsing a region of the human genome using the map to the right.

eg: Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell SEARCH



Browse the **genomic landscape** of cancer

R Resources

Key COSMIC resources

Cell Lines Project
COSMIC
Whole Genomes
Cancer Gene Census
Drug Sensitivity
Mutational Signatures
GRCh37 Cancer Archive

T Tools

Additional tools to explore COSMIC

Cancer Browser
Genome Browser
GA4GH Beacon
CONAN

C Expert Curation

High quality curation by expert postdoctoral scientists

Drug Resistance New
Cancer Gene Census
Curated Genes
Gene Fusions
Genome-Wide Screens

D Data

Further details on using COSMIC's content

Downloads
License
Submission
Genome Annotation
Datasheets
Help
FAQ

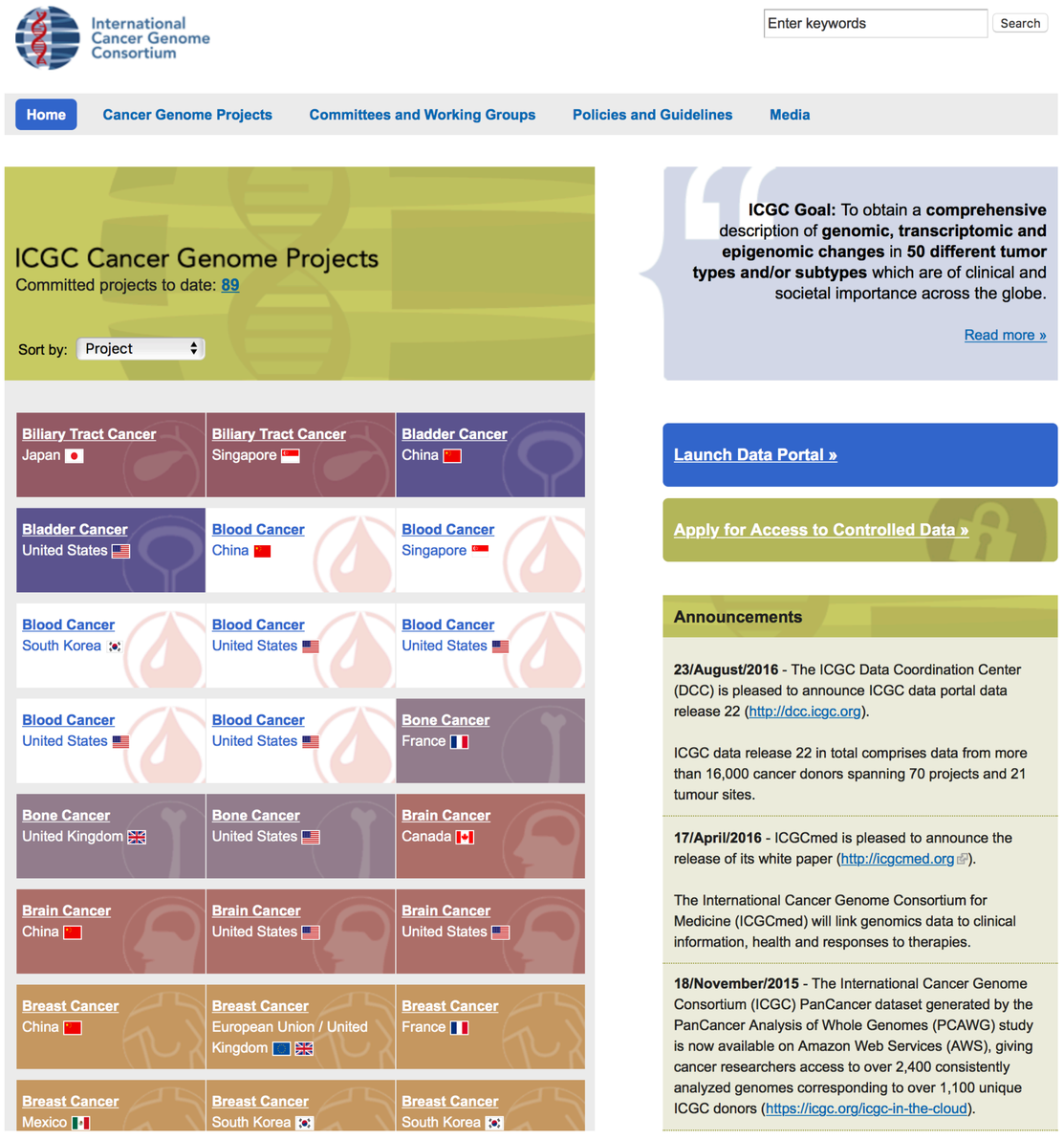
Cancer Gene Census Update

7 genes have been added to the Cancer Gene Census -

EPAS1 - Endothelial PAS domain protein 1.
PTPRT - Protein tyrosine phosphatase, receptor type T.
PPM1D - Protein phosphatase, Mg2+/Mn2+ dependent 1D.
BTK - Bruton tyrosine kinase.
PREX2 - Phosphatidylinositol-3,4,5-trisphosphate dependent Rac exchange factor 2.
TP63 - Tumour protein p63.
QKI - QKI, KH domain containing RNA binding.

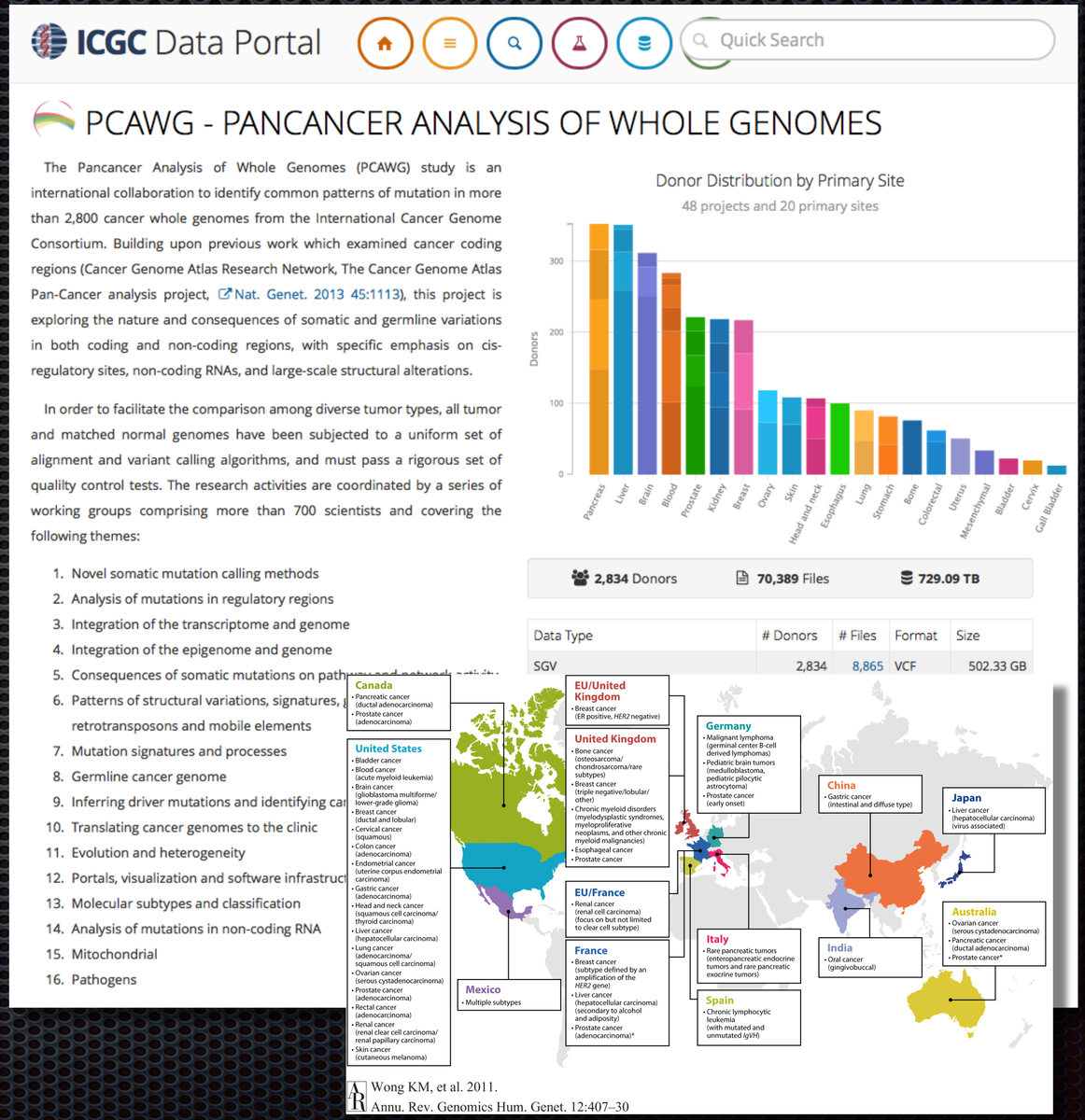
For full details, see the Datasheet.

RESOURCES FOR CANCER GENOMICS



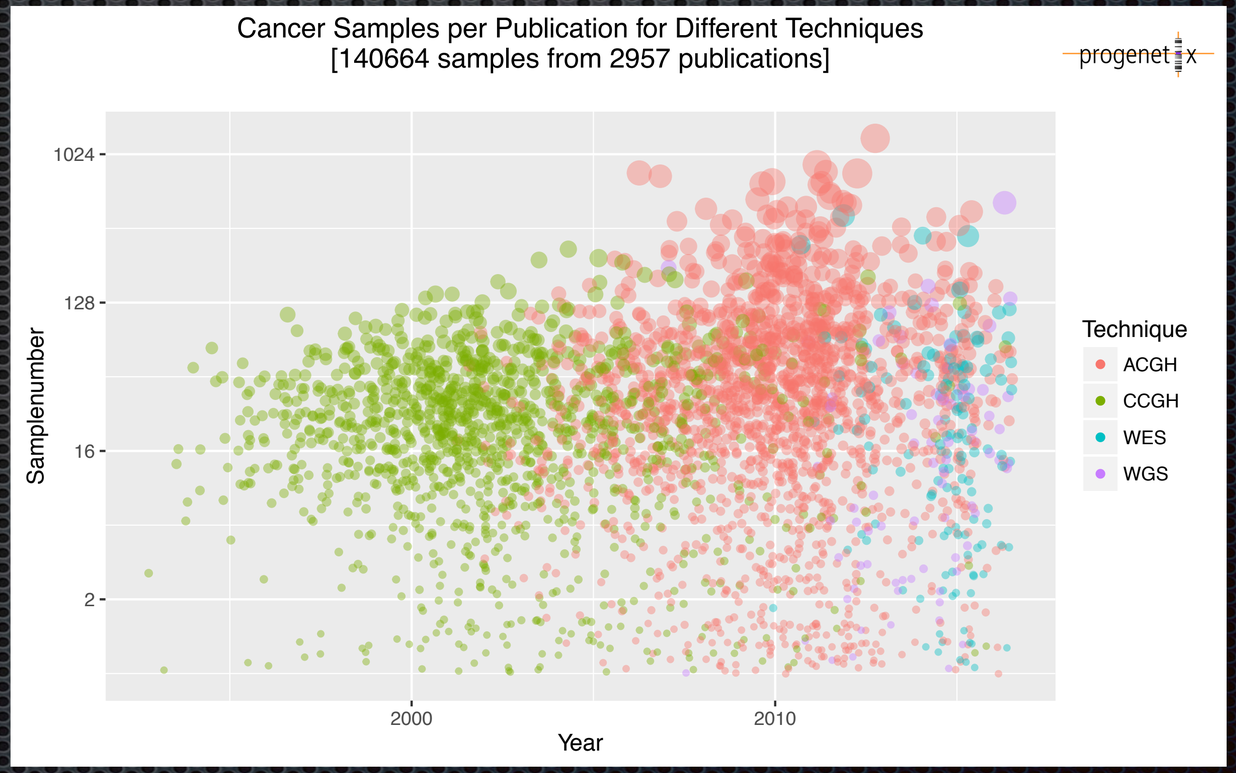
Genome-wide multi"*omics*" data generation for understanding tumor biology

- the International Cancer Genome Consortium (ICGC) as leading example of deep analysis of multiple cancer entities
- international collbaoration of leadinc research centers for each of ~20 tumor types
- limitations:
  - focus on prominant cancer types w/ limited representation of rare entities
  - data access policies influenced by national regulations and legal frameworks
  - technical heterogeneity



Molecular Cytogenetic & Sequencing Studies for **Whole Genome Profiling** in Cancer

- genome screening to identify mutations in cancer samples
- for diagnostic purposes and therapeutic target identification
  - karyotyping (~1968)
  - **C**omparative **G**enomic **H**ybridization (1992)
  - genome **microarrays** (aCGH, SNP arrays ...; 1997)
  - **W**hole **E**xome **S**equencing (2010)
  - **W**hole **G**enome **S**equencing (2011)



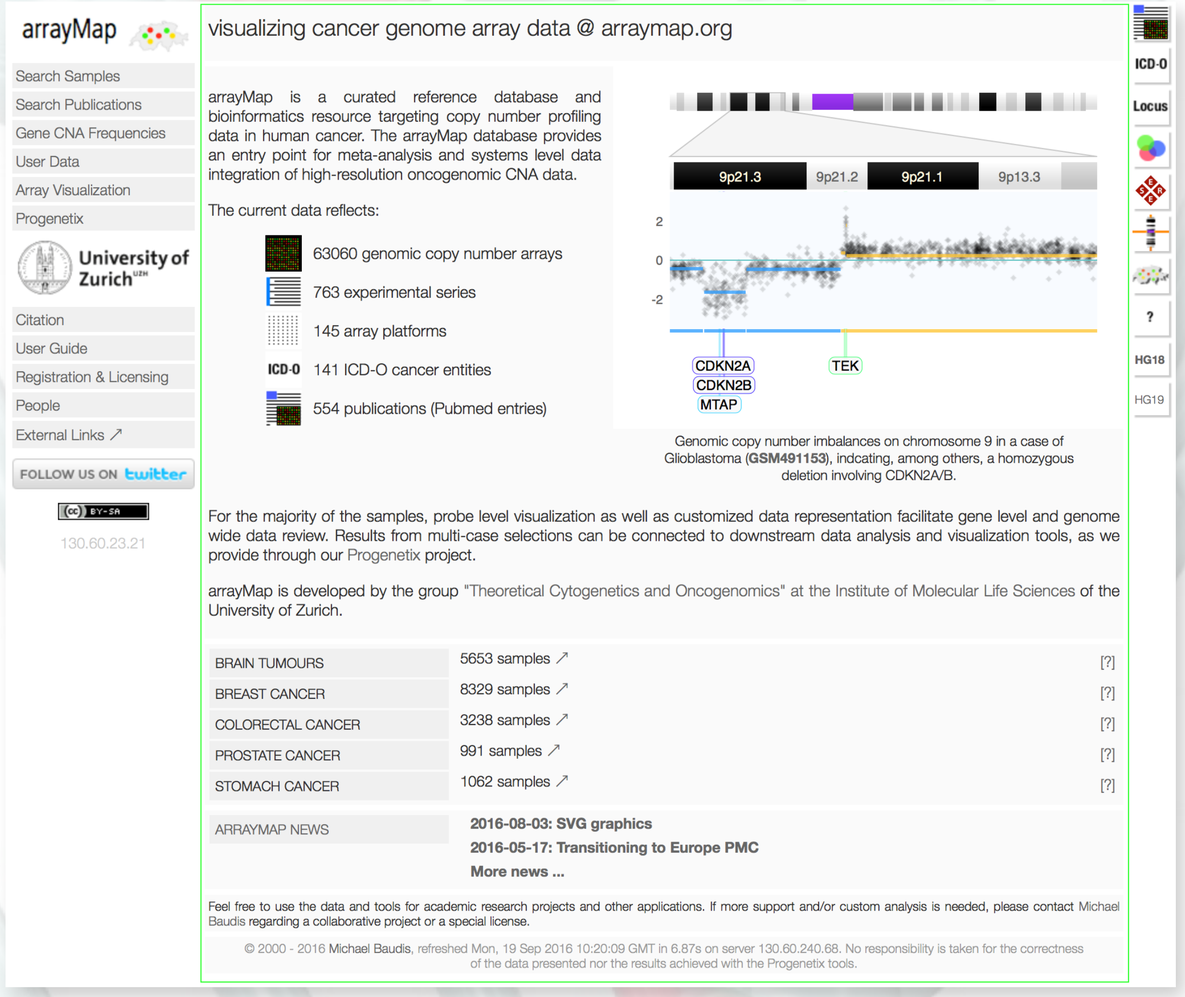
Overview of publications reporting whole-genome screening analysis of cancer samples, by molecular-cytogenic or genome sequencing methods. The data represents articles assessed for the progenetix.org cancer genome data resource (M. Baudis, 2001-2016)

Reference Resources for Cancer Genome Profiling

- continuously updated reference resources for cancer genome profiling data and related information
- basis for own research activities, collaborative projects and external use
- structured information serves for implementing GA4GH concepts

	progenetix	arrayMap
techniques	cCGH, aCGH, WES, WGS	aCGH (+?)
scope	sample (e.g. combination of several experiments)	experiment
content	**>31000** samples	**>60000** arrays
raw data presentation	no (link to sources if available)	yes (raw, log2, segmentation if available)
per sample re-analysis	no; supervised result (mostly as provided through publication)	yes (re-segmentation, thresholding, size filters ...)
final data	annotated/interpreted CN status for GP and cytogenetic regions	unsupervised CN status for GP and cytogenetic regions
main purposes	• Distribution of CNA target regions in most tumor types (>350 ICD-O) • Cancer classification	• Gene specific hits • Genome feature correlation (fragile sites ...)

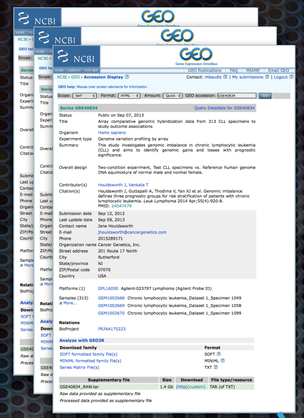
The arrayMap Cancer Genome Resource

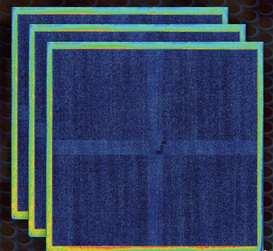


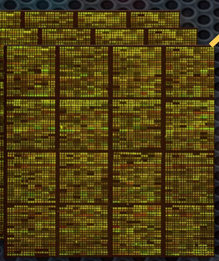
UID	SERIESID	PMID	ICDMORPHOLOGYCODE	ICDTOPOGRAPHYCODE
GSM1000061	GSE36942	23457519	8070/3	C10
GSM1000062	GSE36942	23457519	8070/3	C10
GSM1001316	GSE40777	23571474	8070/3	C53
GSM1001317	GSE40777	23571474	8010/3	C34
GSM1001318	GSE40777	23571474	8070/3	C09
GSM1001319	GSE40777	23571474	8010/3	C34
GSM1002668	GSE40834	24047479	9823/3	C42
GSM1002669	GSE40834	24047479	9823/3	C42
GSM1002670	GSE40834	24047479	9823/3	C42
GSM1002671	GSE40834	24047479	9823/3	C42
GSM1002672	GSE40834	24047479	9823/3	C42
GSM1002673	GSE40834	24047479	9823/3	C42
GSM1002674	GSE40834	24047479	9823/3	C42
GSM1002675	GSE40834	24047479	9823/3	C42
GSM1002676	GSE40834	24047479	9823/3	C42
GSM1002677	GSE40834	24047479	9823/3	C42
GSM1002678	GSE40834	24047479	9823/3	C42
GSM1002679	GSE40834	24047479	9823/3	C42
GSM1002680	GSE40834	24047479	9823/3	C42

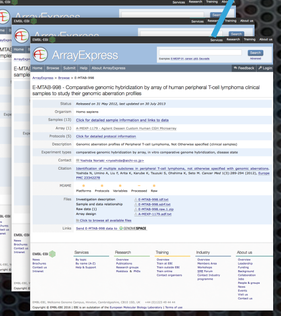
arrayMap

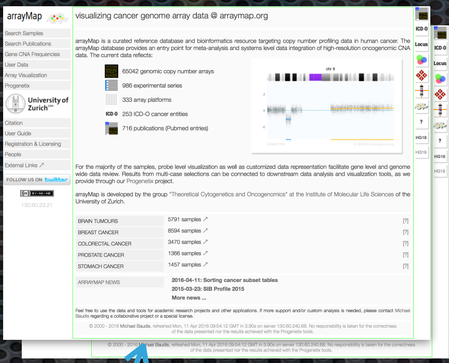
ARRAYMAP DATA PIPELINE

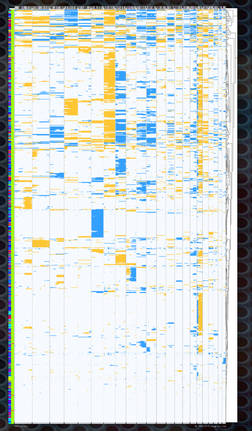


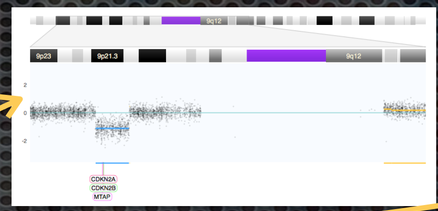


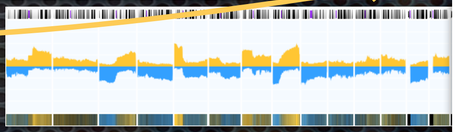


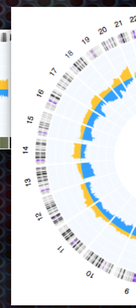


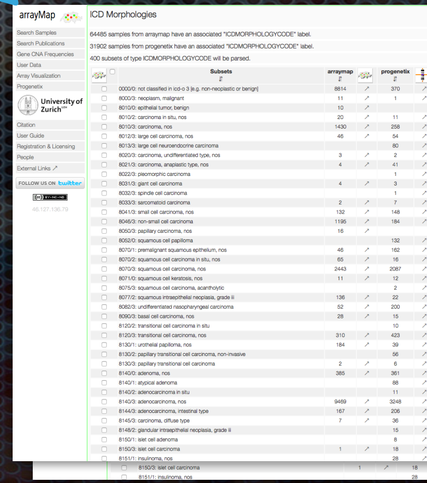












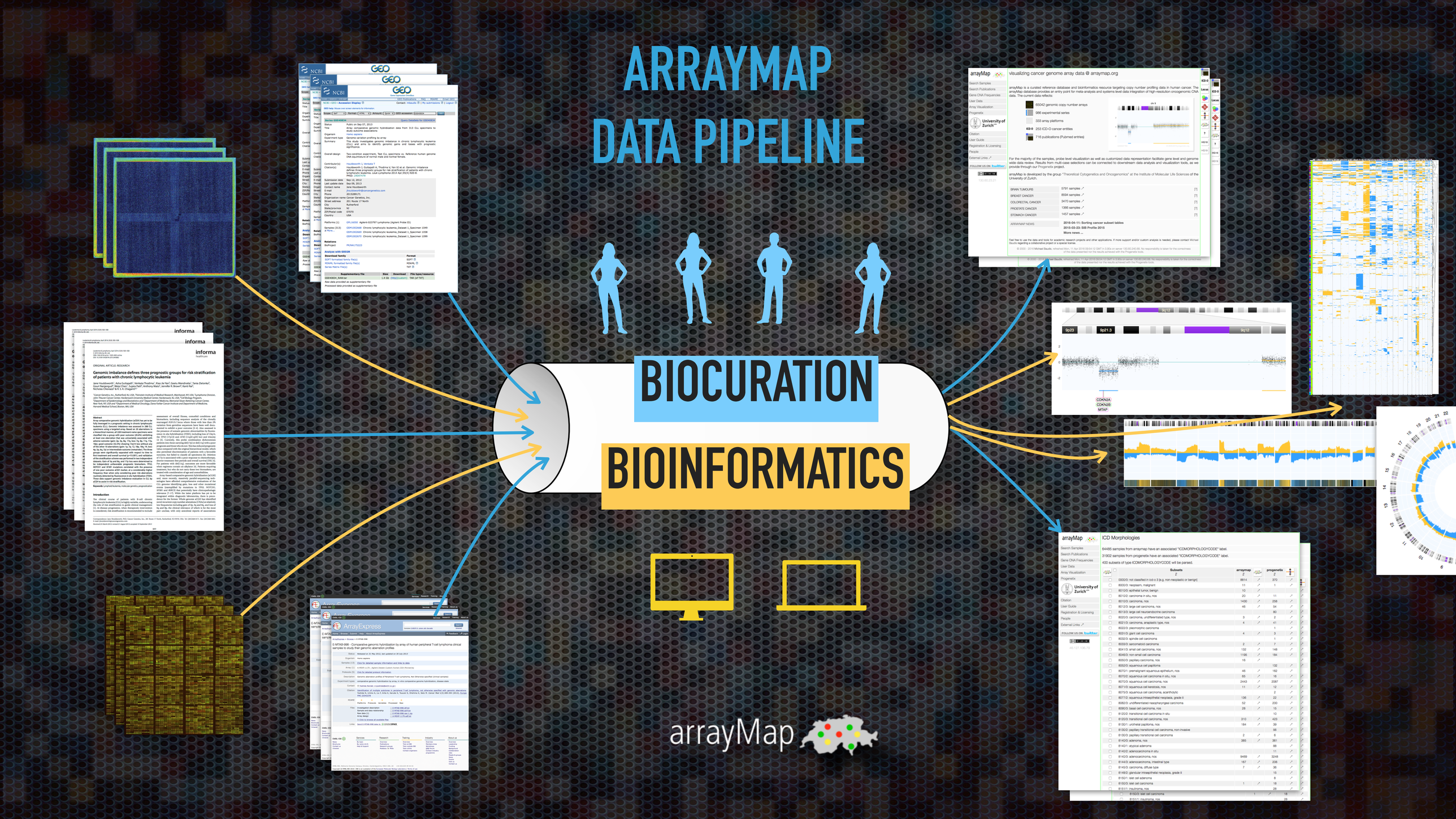
Automagical Processing Engine™

arrayMap

ARRAYMAP DATA PIPELINE

BIOCURATION

BIOINFORMATICS



BIOINFORMATICS & CURATION
ARRAYMAP DATA
"PIPELINE"

GEO
ArrayExpress

amDownload

Metadata
(GEO soft files…)

amParseMeta

metadata.ods
(multi-sample)

amUpdateMeta

db.metadata

Array Probe Data
(log2
tables, .CEL)

Publications
Direct
Submissions

Array
Processing
Pipeline

Metadata Review
Selection
Entity Mapping
(ICD-O, Biosample,
Individual, clinical...)

Manual interpretation
of associated
information (article,
series descriprtion...)

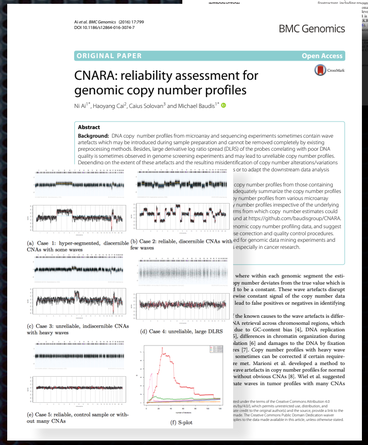
db.experiments

Array
Processing
Pipeline


arrayMap: A Reference Resource for Genomic Copy Number Imbalances in Human Malignancies

arrayMap 2014: an updated cancer genome resource

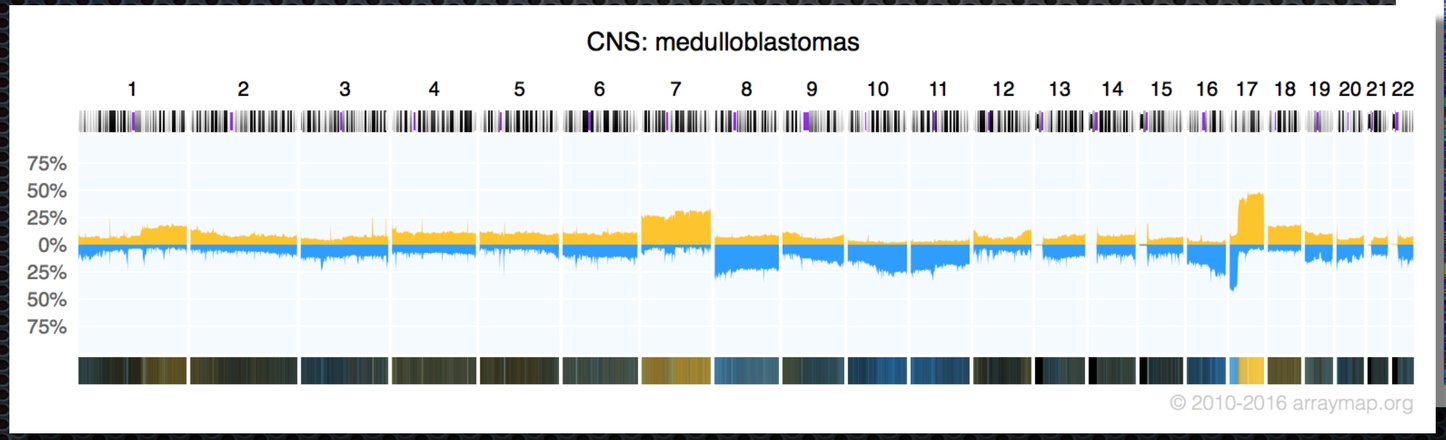

The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases

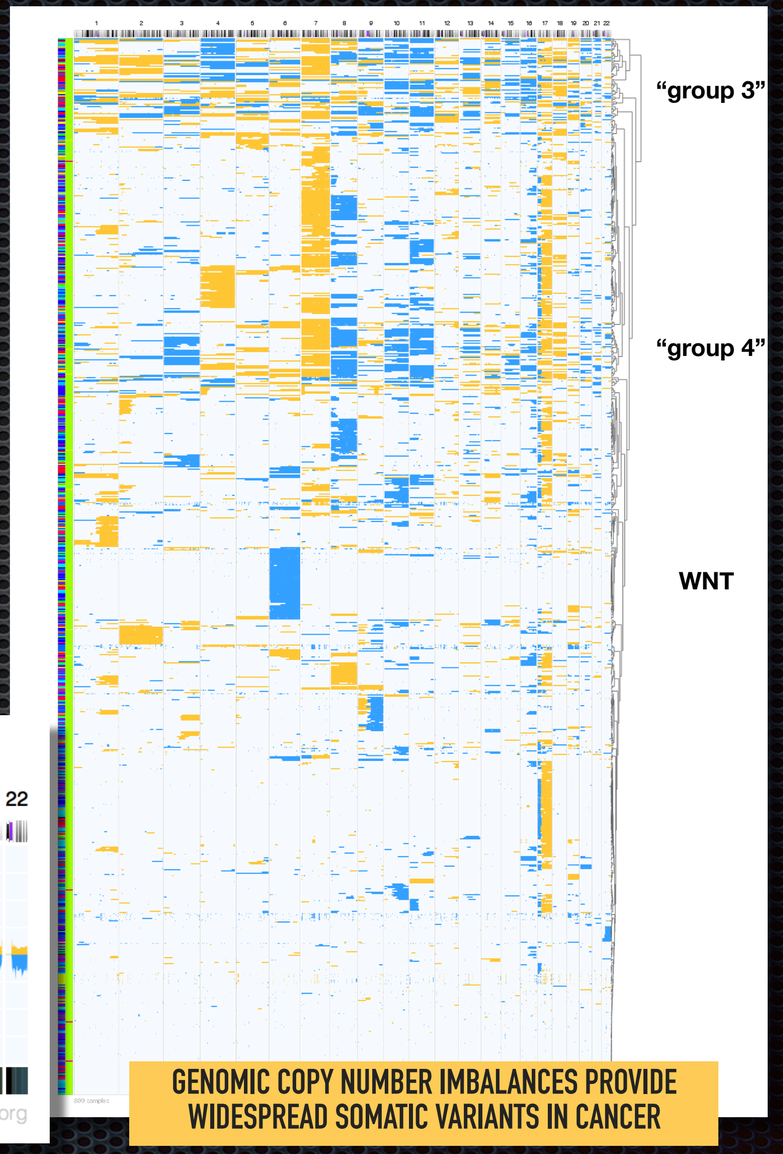

CNARA: reliability assessment for genomic copy number profiles

arrayMap

Somatic Mutations In Cancer: Patterns

- many tumor types express **recurrent mutation patterns**
- **How can** those patterns be used for classification and determination of biological mechanisms?





A genomic copy number histogram for malignant medulloblastomas, the most frquent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional relevance and clinical correlation. From arraymap.org

Somatic Mutations In Cancer: Patterns III
Making the case for genomic classifications
Some related cancer entities show similar copy number profiles

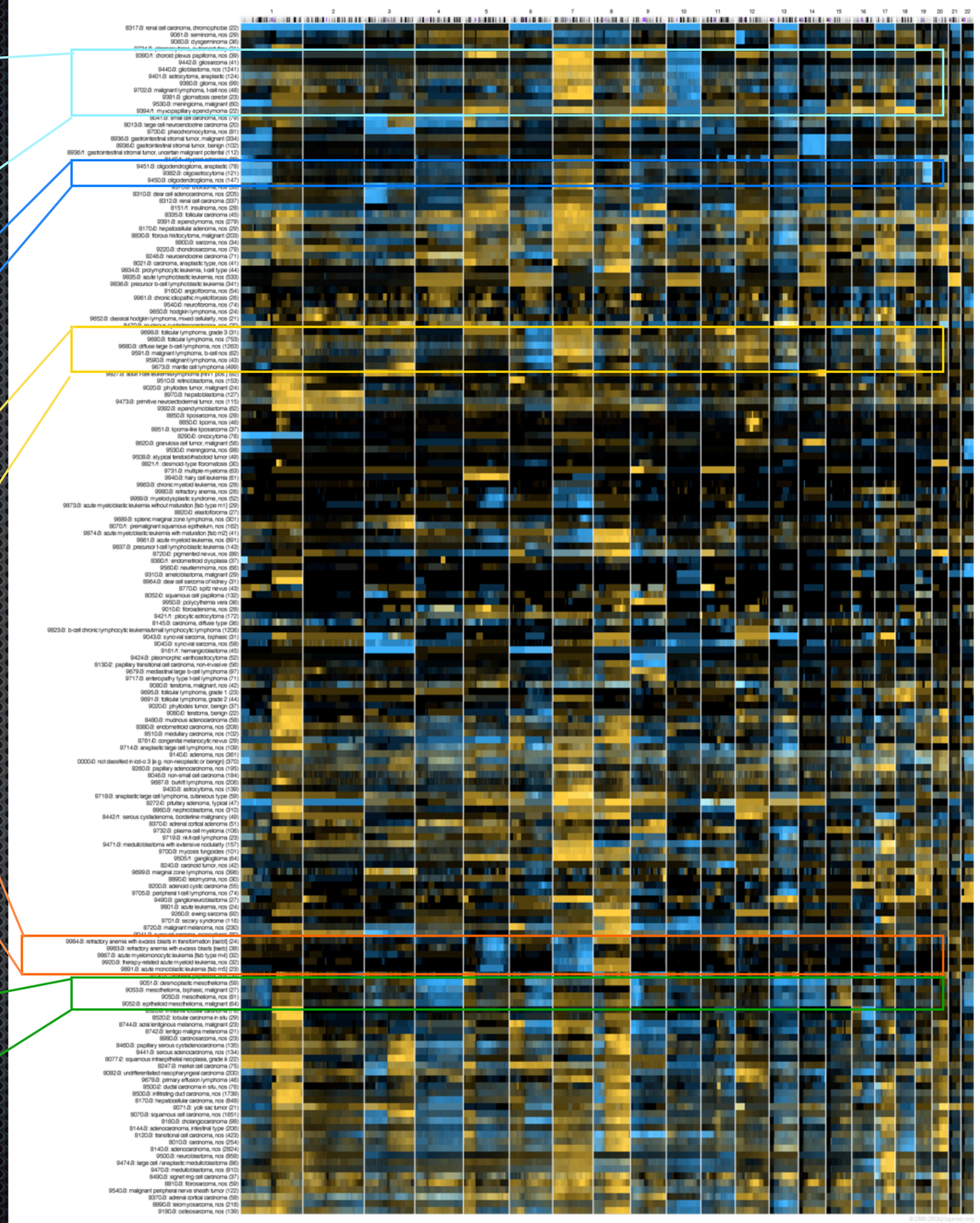
9390/1: choroid plexus papilloma, nos (39)
9442/3: gliosarcoma (41)
9440/3: glioblastoma, nos (1241)
9401/3: astrocytoma, anaplastic (124)
9380/3: glioma, nos (99)
9702/3: malignant lymphoma, t-cell nos (48)
9381/3: gliomatosis cerebri (23)
9530/3: meningioma, malignant (60)
9394/1: myxopapillary ependymoma (22)

9451/3: oligodendroglioma, anaplastic (78)
9382/3: oligoastrocytoma (121)
9450/3: oligodendroglioma, nos (147)

9698/3: follicular lymphoma, grade 3 (31)
9690/3: follicular lymphoma, nos (753)
9680/3: diffuse large b-cell lymphoma, nos (1263)
9591/3: malignant lymphoma, b-cell nos (62)
9590/3: malignant lymphoma, nos (43)
9673/3: mantle cell lymphoma (499)

9984/3: refractory anemia with excess blasts in transformation [raebt] (24)
9983/3: refractory anemia with excess blasts [raeb] (38)
9867/3: acute myelomonocytic leukemia [fab type m4] (32)
9920/3: therapy-related acute myeloid leukemia, nos (32)
9891/3: acute monoblastic leukemia [fab m5] (23)

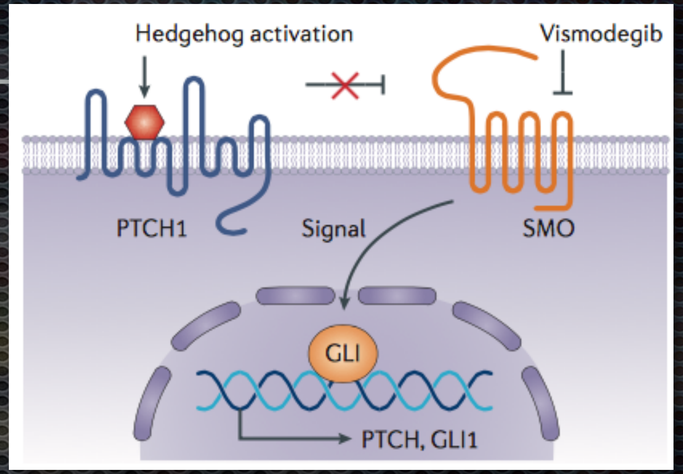
9051/3: desmoplastic mesothelioma (59)
9053/3: mesothelioma, biphasic, malignant (27)
9050/3: mesothelioma, nos (81)
9052/3: epithelioid mesothelioma, malignant (64)



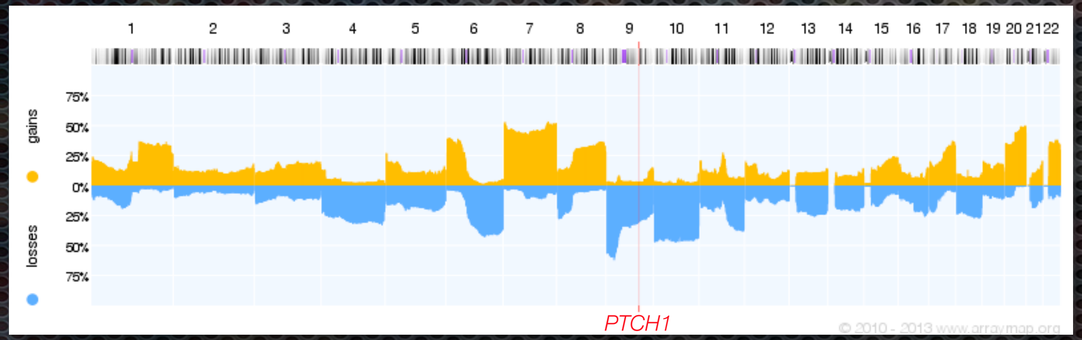
arrayMap

Large datasets for rare cancer gene events

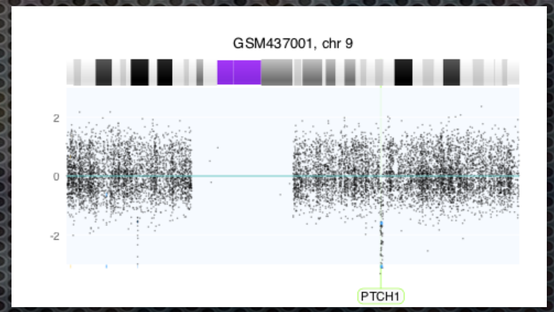
- The Sonic Hedgehog (SHH) pathway has become a “druggable” target in the therapy of syndromic and/or advanced basaliomas (e.g. in Gorlin syndrome).
- In the pathway, PTCH1 acts as “tumor suppressor” counteracting SMO=>GLI mediated transcriptional activation.
- We were interested if the gene also could be involved in subsets of malignant melanomas ...

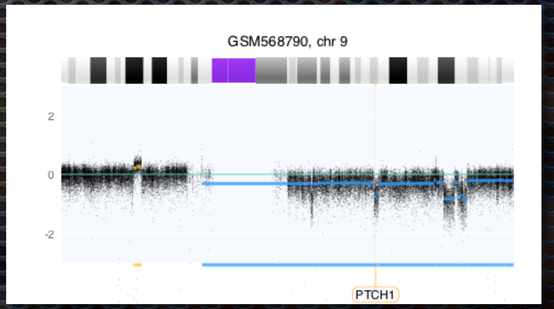


Dlugosz, A., Agrawal, S., & Kirkpatrick, P. (2012, June). Nature Reviews Drug Discovery, pp. 437–438



no “hot spot” (but 30% deletions)



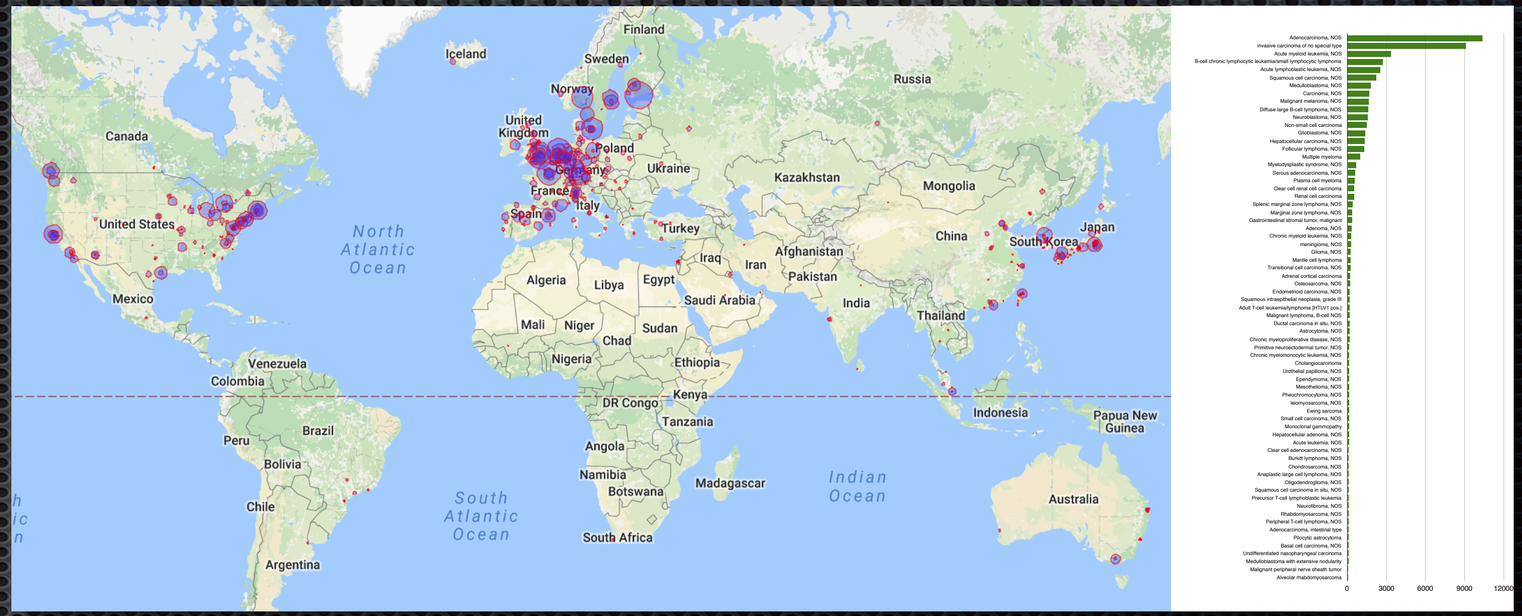


probably pathogenic homozygous deletions in few cases (3/~700): large datasets needed

arrayMap

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

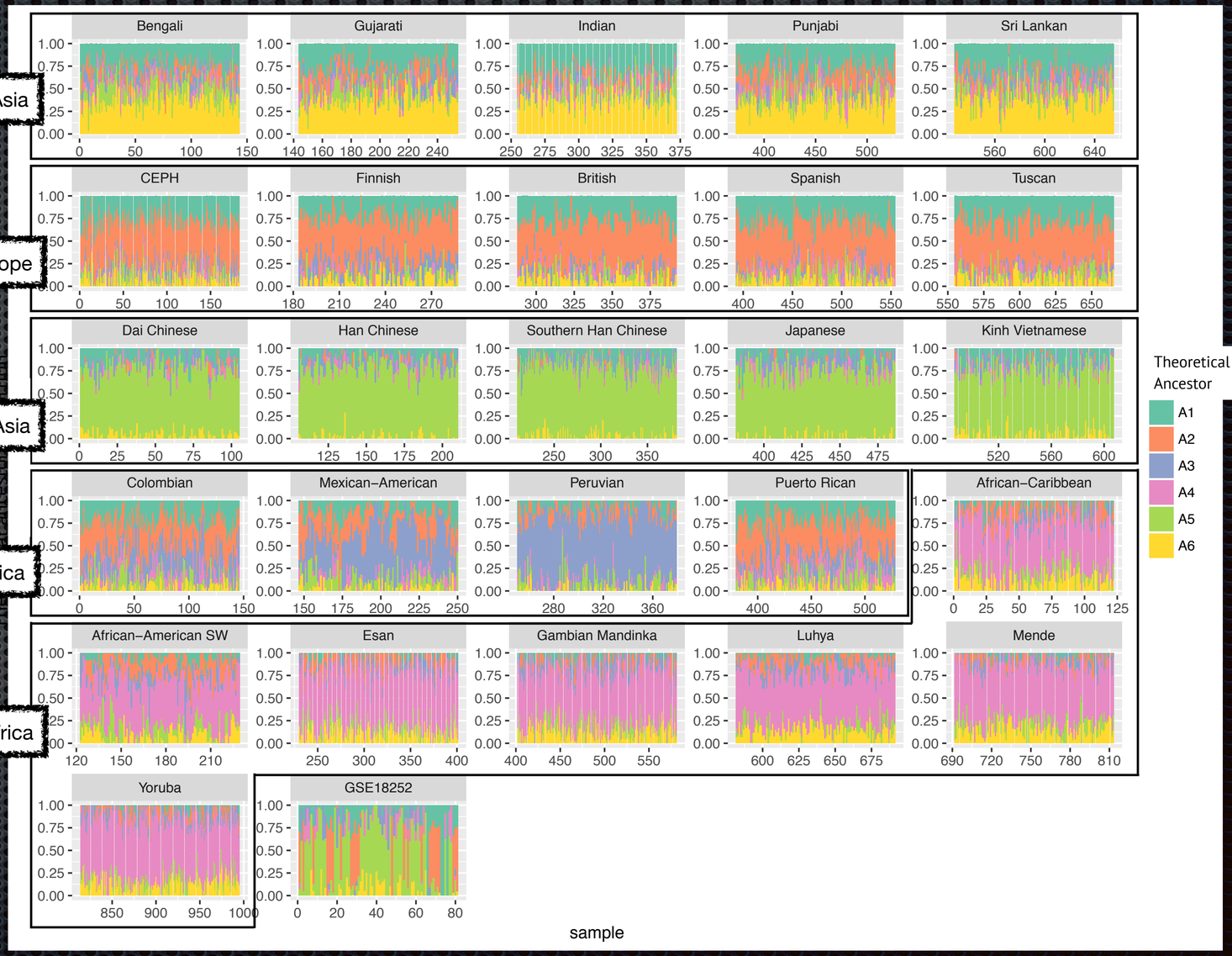


Geographic distribution of >140'000 cancer genome profiles reported in the literature. The numbers are derived from the 2947 publications registered in the Progenetix database.

progenetix

Population stratification in cancer samples based on SNP array data

- 2504 genome profiles from 1000 Genome project phase 1 as reference
- 5 superpopulations: South Asia, Europe, South America, East Asia and Africa.
- SNP positions used in 9 Affymetrix SNP arrays are extracted to train a population admixture model.



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 →

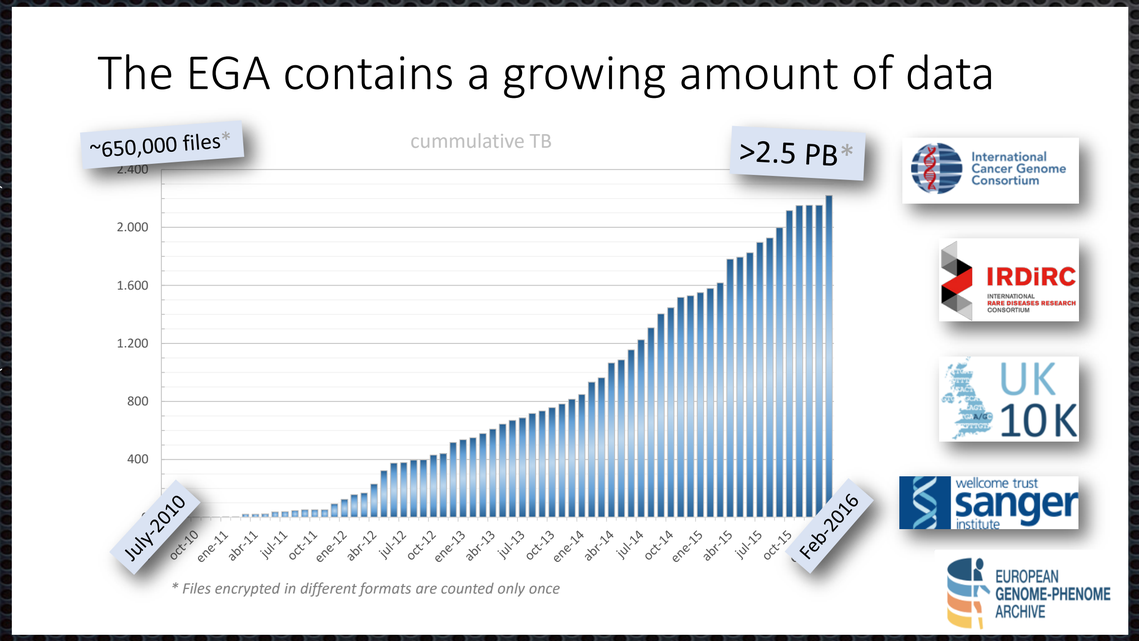
VIEW OUR LEADERSHIP

MORE ABOUT US

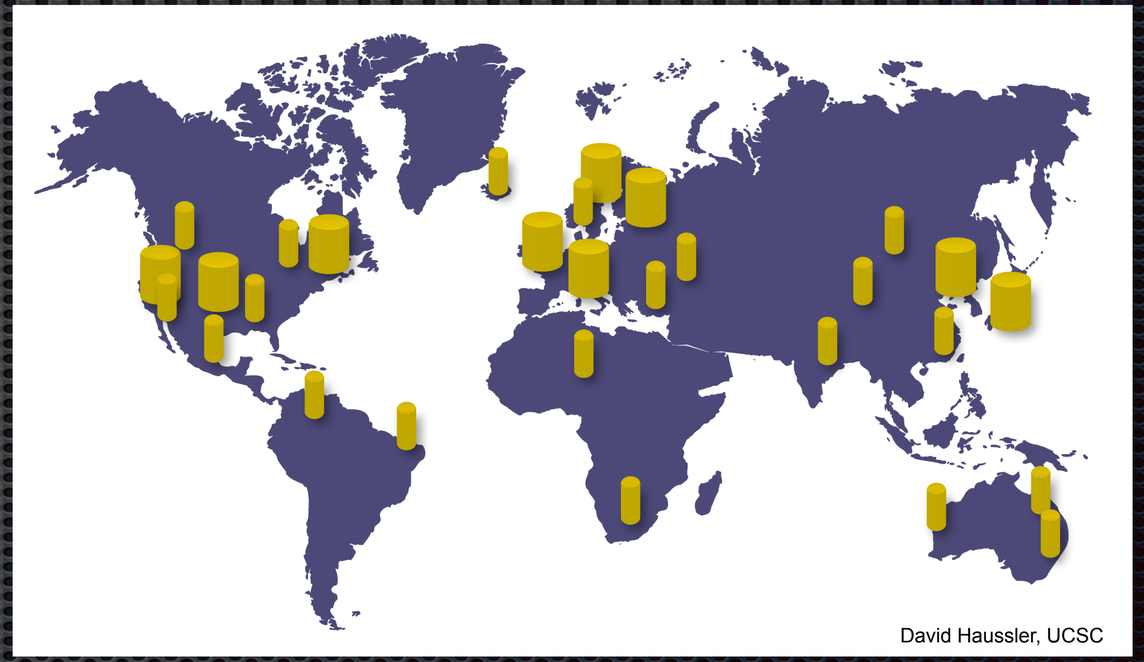
BECOME A MEMBER

Genome Datasets: Rapid Growth, Limited Access

population based and cancer research studies produce a rapidly increasing amount of genome sequence data



Source: Jordi Rambla (CRG Barcelona) 2016



genome data is stored in an increasing number of institutional and core repositories, with **incompatible data** structures and **access** policies

GA4GH API promotes sharing

GENOMICS

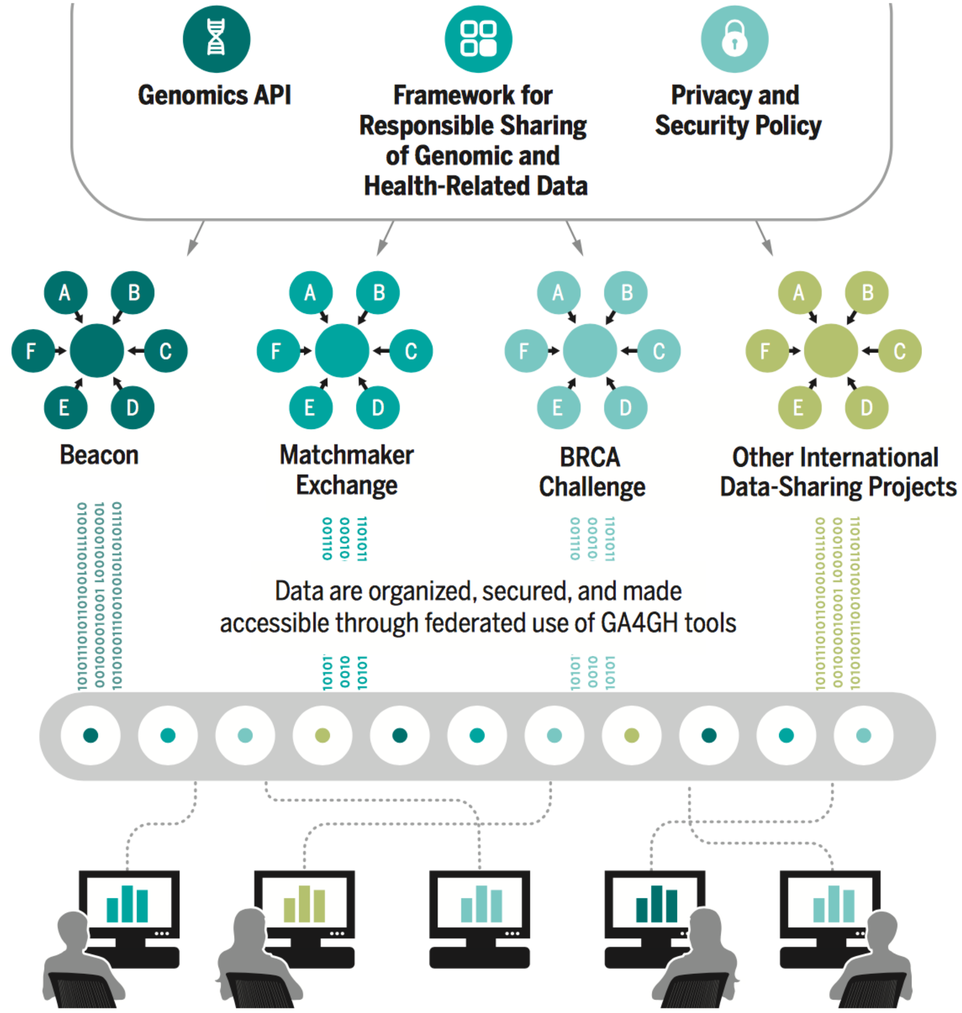
A federated ecosystem for sharing genomic, clinical data

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.



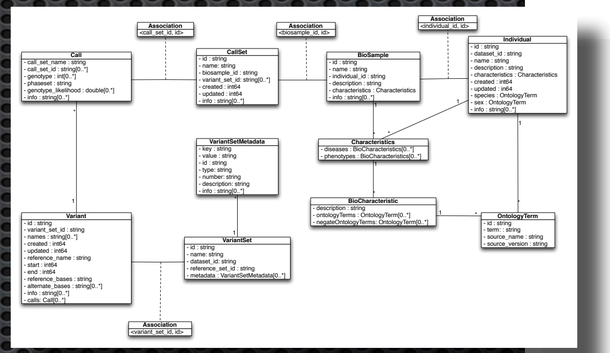
Global Alliance for Genomics & Health

GA4GH HISTORY & MILESTONES

- January 2013 - 50 participants from eight countries
- June 2013 - White Paper, over next year signed by 70 “founding” member institutions (e.g. SIB, UZH)
- March 2014 - Working group meeting in Hinxton & 1st plenary in London
- October 2014 - Plenary meeting, San Diego; interaction with ASHG meeting
- June 2015 - 3rd Plenary meeting, Leiden
- September 2015 - GA4GH at ASHG, Baltimore
- October 2015 - DWG / New York Genome Centre
- April 2016 - Global Workshop @ ICHG 2016, Kyoto
- October 2016 - 4th Plenary Meeting, Vancouver
- May 2017 - Strategy retreat, Hinxton
- October 2017 - 5th plenary (Orlando): new structure

Developing the GA4GH **Metadata** Schema

- arrayMap for GA4GH
  - metadata schema development through implementation of arrayMap resource data
  - OntologyTerm objects for biodata
  - implementation w/ ontology services



Driving **Beacon** Development

- Beacon+
  - CNV/CNA as first type of structural variants
  - disease specific queries
  - quantitative reporting

Global Alliance for Genomics & Health

EMBL-EBI

arrayMap

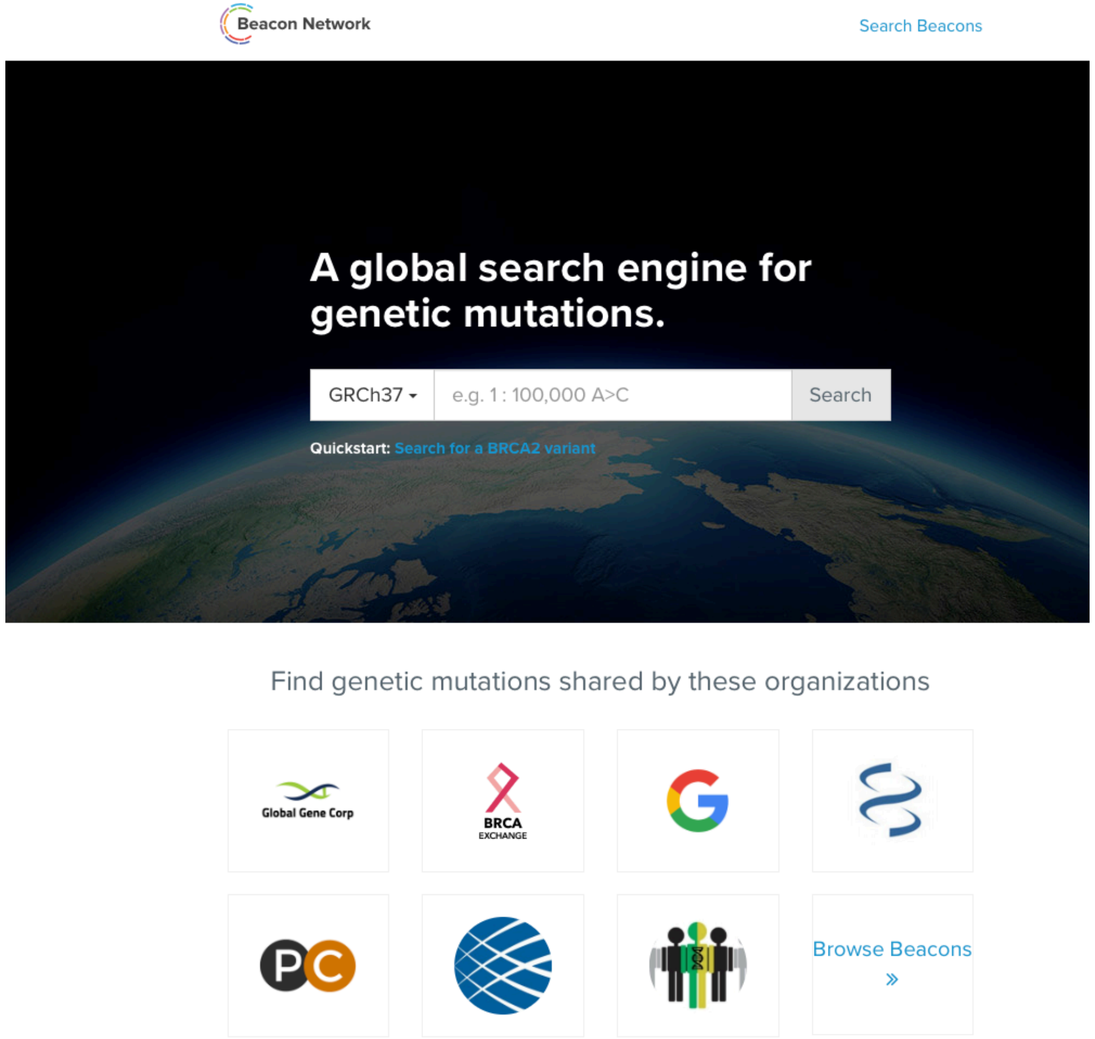
elixir

SIB Swiss Institute of Bioinformatics

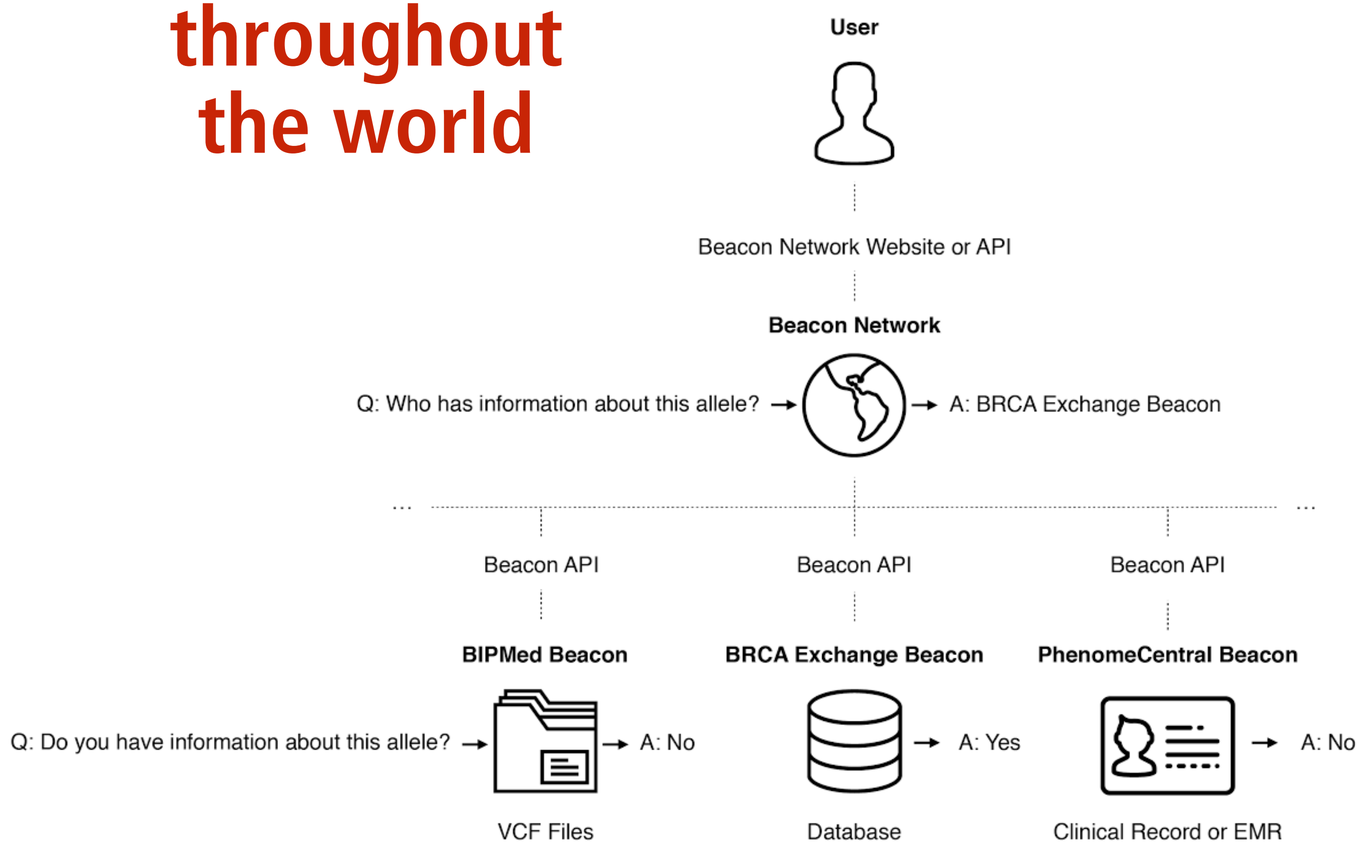

```
"_id" : ObjectId("58297ca32ca4591e5a0df054"),
"id" : "AM_V_1778741",
"variant_set_id" : "AM_VS_HG18",
"reference_name" : "10"
"start" : 579049,
"end" : 17236099,
"alternate_bases" : "DUP",
"reference_bases" : ".",
"info" : {
  "svlen":16657050,
  "cipos":[
      -1000,
      1000
  ],
  "ciend":[
      -1000,
      1000
  ]
},
"calls" : [
        {
                "genotype" : [
                        ".",
                        "."
                ],
                "call_set_id" : "AM_CS_TCGA-61-1917-01A-01D-0648-01",
                "info" : {
                        "segvalue" : 0.5491
                }
        }
],
"created" : ISODate("2016-11-14T08:33:58.202Z"),
"updated" : ISODate("2016-11-14T08:33:58.202Z"),
}
```

Beacon Project

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



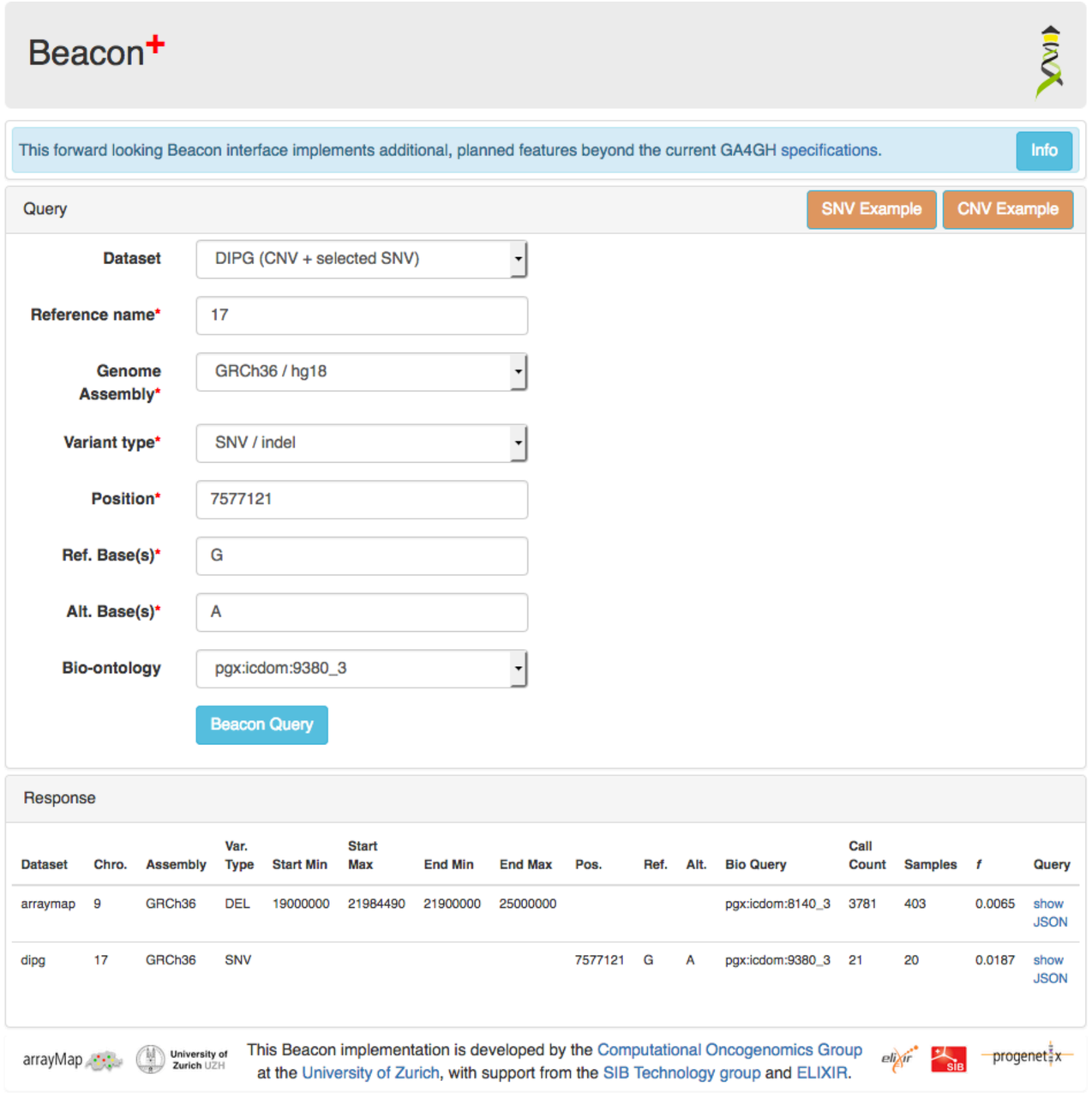
> 50 Beacons throughout the world



Global Alliance for Genomics & Health

Beacon+ Concept

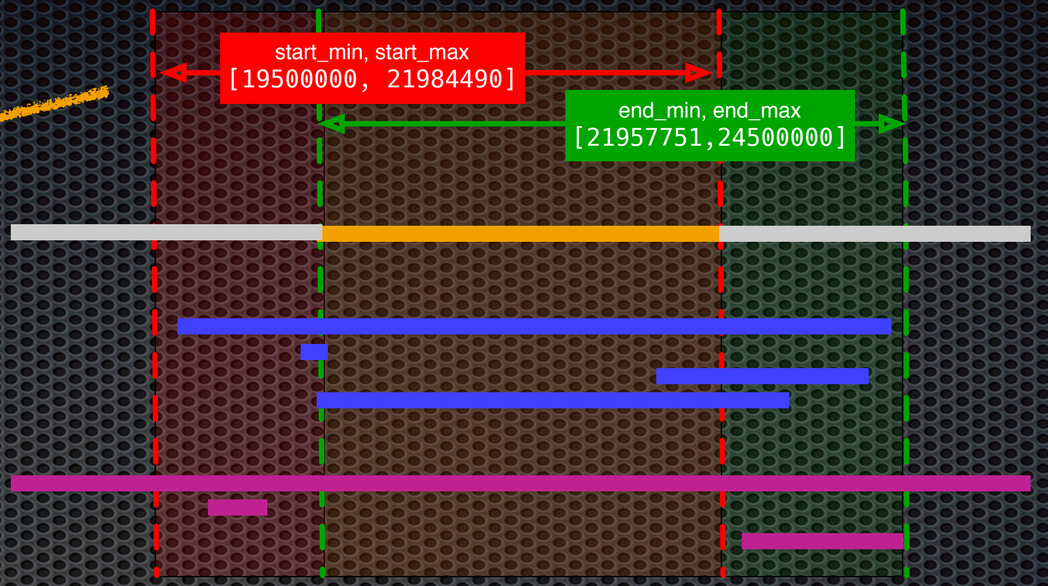
- Implementation of cancer beacon prototype, backed by arrayMap and DIPG data set (MacKay *et al.*, Cancer Cell 2017, in print)
- structural variations (DUP, DEL) in addition to SNV
- diagnosis queries using ontology codes (NCIT, ICD-O)
- quantitative responses
- current version uses **GA4GH schema compatible** database



```
{
   "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" : {
   "url" : “http://progenetix.org/beacon/info/“,
   "dataset_allele_responses" : [
      {
         “dataset_id" : "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

Metadata



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

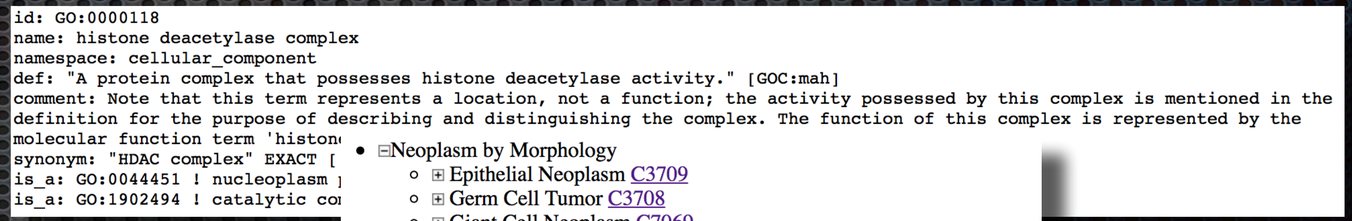
- Beacon+ **range queries** allow the definition of a genome region of interest, containing a specified variant (or other mappable feature)
- “fuzzy” matching of region ends is essential for features without base specific positions
- current Beacon implementation addresses CNV (<DUP>, <DEL>), as are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

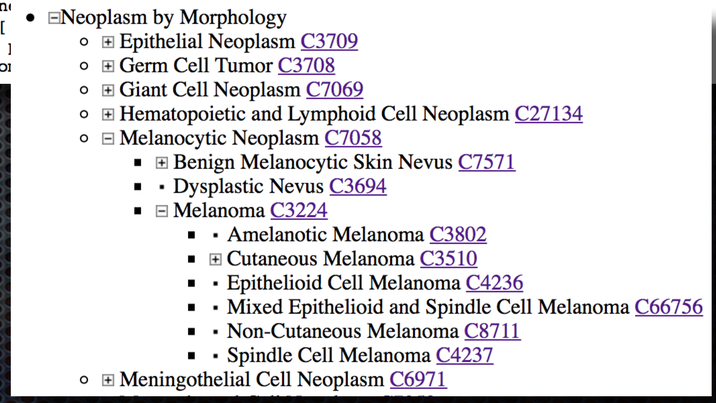
Bioinformatics: **Ontologies**

- ontologies in information sciences describe concrete and abstract **objects**, there precisely defined **hierarchies** and **relationships**
- ontologies in bioinformatics support the move from a descriptive towards an **analytical science** in describing biological data and relations among it

> "The widest use of ontologies within biology is for conceptual annotation – a representation of stored knowledge more computationally amenable than natural language."*

- Gene ontology (GO)
- NCIt Neoplasm Core
- Uberon anatomical structures
- Experimental Factor Ontology (EFO)
- Disease Ontology (DO)





Ontologies in bioinformatics. R Stevens, C Wroe, P Lord and C Goble, Department of Computer Science, University of Manchester. Retrieved from http://homepages.cs.ncl.ac.uk/phillip.lord/download/publications/handbook.pdf

ONTOLOGIES ARE RARELY USED IN CASE REPORTING

- Medical practice relies on established, slow moving classification systems.
- Medical diagnoses consist of an abundance of observations and classification items.
- We do not have (never will?) enough ontology concepts for detailed disease descriptions (Where to stop?)
- Relationships may help - but how to do them uniformly?

ONTOLOGIES ARE RARELY USED IN CASE REPORTING

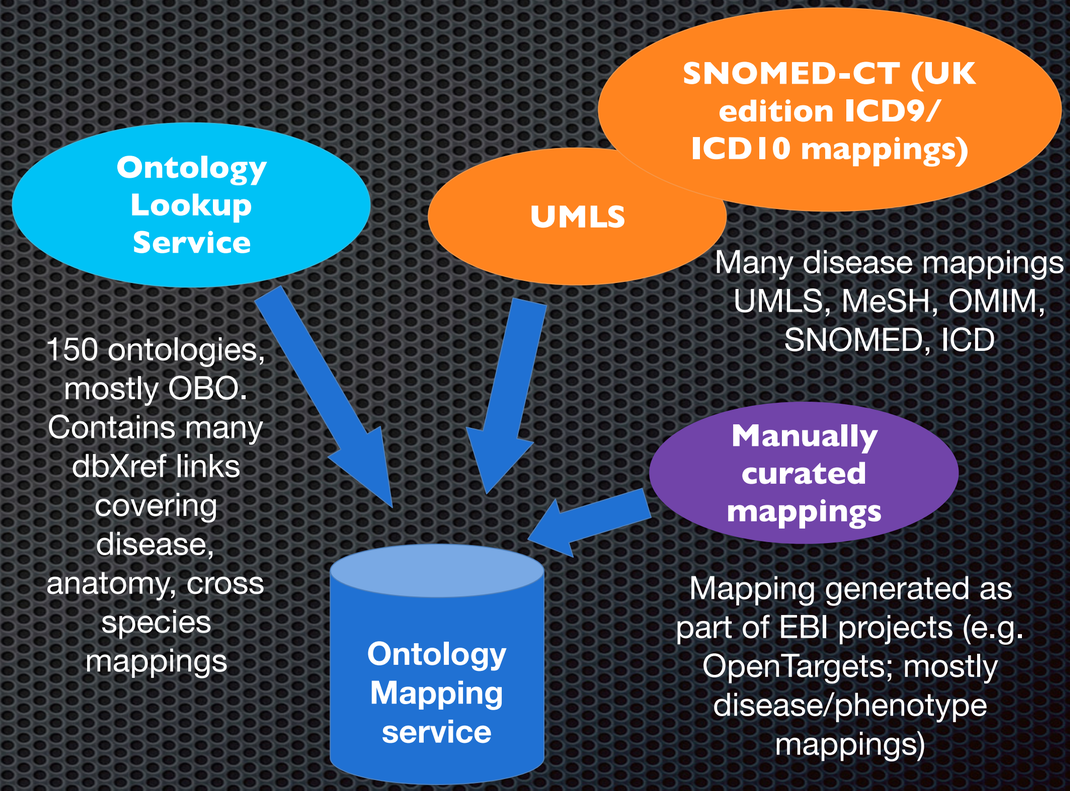
- Medical practice relies on established, slow-changing classification systems.
- Medical diagnoses consist of [illegible]undance of observations and classification items.
- We do not have (and never will?) enough ontology concepts for detailed disease descriptions (Where to stop?)
- [illegible]tionships may help - but how to do them uniformly?

THE GOLD STANDARD FOR A MEDICAL DIAGNOSIS IS STILL WELL WRITTEN PROSE.

Making Ontologies Work for GA4GH Implementation Studies

- biomedical "metadata" in different resources frequently follows incompatible classification systems
- medical coding systems are driven by different paradigms compared to biological ontologies (e.g. for cross-species comparisons)
- frequently used classifications (ICD, Snomed…) are either not “ontologised” or cannot be referenced in open resources

Federated queries across resources need **curated mappings** of classifications/ ontologies

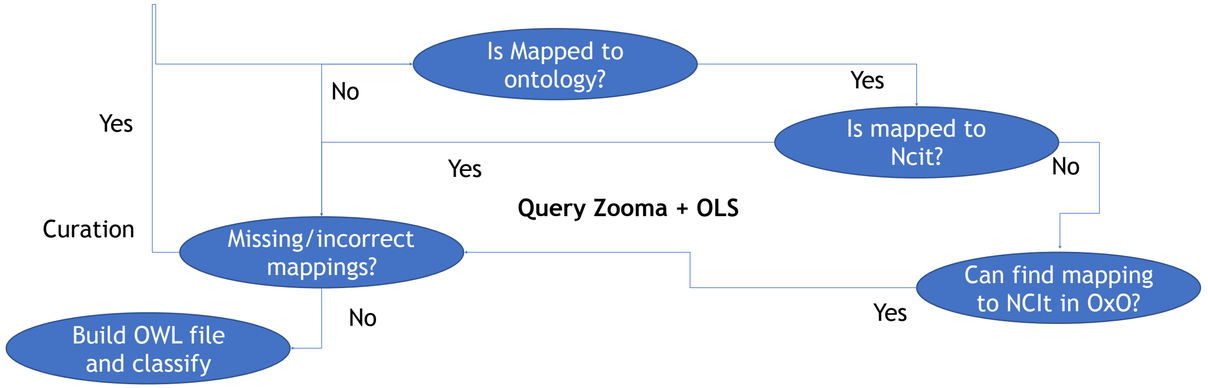


Ontology Mapping Service

Working towards ontologies w/ arrayMap: Mapping >55'000 samples from ICD-O to NCIt neoplasm core

ICDM	ICDMORPHOLOGY	NCItcode	NCItlabel	ICDT	ICDTOPOGRAPHY	NCItcode	NCItlabel	NCIt_mapp	NCIt_mapped_ICDM_T_label
8021/3	Carcinoma anaplastic type			C739	thyroid gland	C12400	thyroid gland	C3878	Thyroid Gland Undifferentiated (Anaplastic) Carcinoma
9451/3	Oligodendroglioma anaplastic	C4326	anaplastic oligodendroglioma	C719	Brain	C12439	brain	C4326	Anaplastic Oligodendroglioma
9051/3	Desmoplastic mesothelioma	C6747		C499	connective and soft tissue			C6747	Desmoplastic Mesothelioma
9732/3	Plasma cell myeloma	C3242	multiple myeloma	C42	hematopoietic and reticuloendothelial systems			C3242	Plasma Cell Myeloma
8070/3	Squamous cell carcinoma	C2926	non-small cell lung carcinoma	C140	pharynx			C102872	Pharyngeal Squamous Cell Carcinoma
8380/3	Endometrioid adenocarcinoma	C3769	endometrioid carcinoma	C54	corpus uteri	C12316		C6287	Endometrial Endometrioid Adenocarcinoma
8070/3	Squamous cell carcinoma	C2926	non-small cell lung carcinoma	C44	skin	C12470	zone of skin	C4819	Skin Squamous Cell Carcinoma
8430/3	Mucoepidermoid carcinoma	C45544	pulmonary mucoepidermoid carcinoma	C089	salivary gland	C12426	saliva-secreting gland	C5953	Minor Salivary Gland Mucoepidermoid Carcinoma
9680/3	Diffuse large B-cell lymphoma	C8851	diffuse large B-cell lymphoma	C42	hematopoietic and reticuloendothelial systems			C8851	Diffuse Large B-Cell Lymphoma
8800/3	Sarcoma	C9118	sarcoma	C559	uterus nos			C9306	Soft Tissue Sarcoma
8441/3	Serous adenocarcinoma	C7550	ovarian serous adenocarcinoma	C570	fallopian tube	C12403	fallopian tube	C40101	Serous Adenocarcinoma
9689/3	splenic marginal zone lymphoma nos			C422	spleen	C12432	spleen	C4663	Splenic Marginal Zone Lymphoma
8077/2	Squamous intraepithelial neoplasia grade III			C53	cervix uteri	C12311		C89476	Grade III Vaginal Intraepithelial Neoplasia
8140/0	Adenoma	C4196	adenoma	C189	large intestine excl. rectum and rectosigmoid junction			C4349	Colon Adenocarcinoma
8272/3	Pituitary carcinoma	C4536	Pituitary carcinoma	C751	pituitary gland	C12399	pituitary gland	C4536	Pituitary Gland Carcinoma
8500/2	Ductal carcinoma in situ	C3641	ductal carcinoma in situ	C50	breast	C12971	breast	C2924	Ductal Breast Carcinoma In Situ
8200/3	Adenoid cystic carcinoma	C2970	adenoid cystic carcinoma	C32	larynx	C12420	larynx	C2970	Adenoid Cystic Carcinoma
9370/3	Chordoma	C2947	Chordoma	C419	bone	C13076	bone tissue	C2947	Chordoma
9717/3	Enteropathy type T-cell lymphoma			C17	small intestine	C12386	small intestine	C4737	Enteropathy-Associated T-Cell Lymphoma
9698/3	Follicular lymphoma grade 3			C42	hematopoietic and reticuloendothelial systems			C3460	Grade 3 Follicular Lymphoma
9863/3	Chronic myeloid leukemia	C3177	chronic myelogenous leukemia	C42	hematopoietic and reticuloendothelial systems			C3174	Chronic Myelogenous Leukemia BCR-ABL1 Positive
8852/3	Liposarcoma myxoid	C3735	myxoid liposarcoma	C499	connective and soft tissue			C27781	Myxoid Liposarcoma
9080/3	Teratoma malignant			C809	unknown	C35882	Hereditary elliptocytosis	C3403	Tetratoma
8530/3	Inflammatory carcinoma	C4872	breast carcinoma	C50	breast	C12971	breast	C4001	Inflammatory Breast Carcinoma
8140/3	Adenocarcinoma	C27745	lung adenocarcinoma	C809	unknown	C35882	Hereditary elliptocytosis	C2852	Adenocarcinoma
						C12762	oropharynx	C2970	Adenoid Cystic Carcinoma
						C12415	kidney	C3158	Leiomyosarcoma
						C12499	internal ear	C2970	Adenoid Cystic Carcinoma
						C12683	bronchus	C2923	Bronchioloalveolar Carcinoma
						C12343	retina	C3224	Melanoma
					ial systems			C8459	Hepatosplenic T-Cell Lymphoma
						C12393	pancreas	C8294	Pancreatic Adenocarcinoma
					ial systems			C3996	Monoclonal Gammopathy of Undetermined Significance
								C4817	Ewing Sarcoma
								C3288	Oligodendroglioma
						C12422	tongue	C4648	Tongue Squamous Cell Carcinoma
					ial systems			C2862	Primary Myelofibrosis
								C4833	Oral Cavity Squamous Cell Carcinoma
						C12390	rectum	C9383	Rectal Adenocarcinoma
								C3158	Leiomyosarcoma
					ial systems			C3898	Extranodal Marginal Zone Lymphoma of Mucosa-Associate
						C12404	female gonad	C4512	Ovarian Mucinous Cystadenoma

- From 456 pairs of ICD-O terms Morphology and Topography representative of cancer entities in arrayMap
- Develop Python script to take ICD-O Morphology and Topography labels separately QUERY ZOOMA, OxO and OLS to find mapping to NCIt



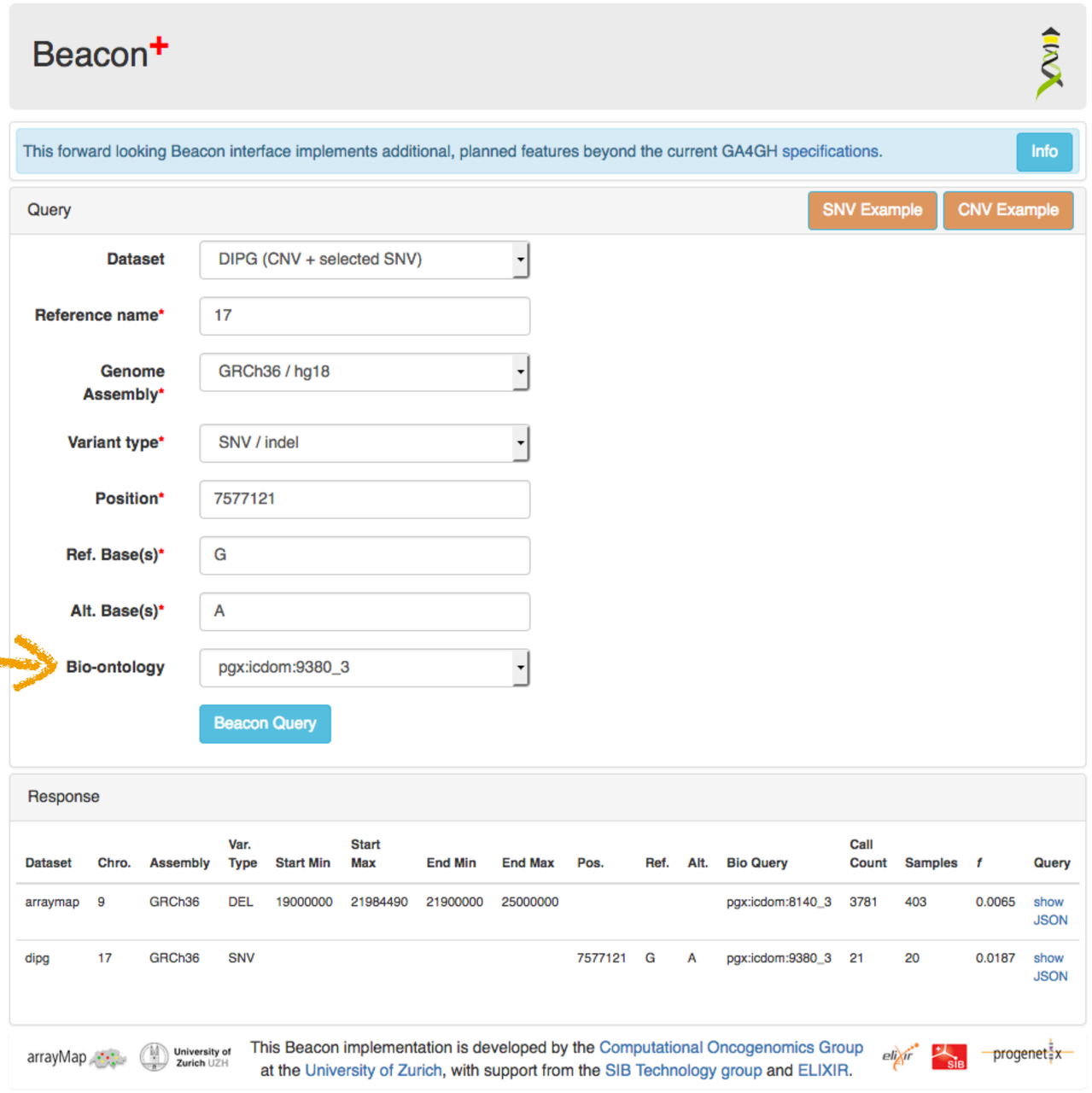
From 456 pairs of ICD-O
70% ICD-O Morphology - NCIt
65% ICD-O Topography - NCIt

45% ICD-O-3 Pairs mapped to NCIt terms

=> MANUAL CURATION of >50%

Beacon+ Concept

- Implementation of cancer beacon prototype, backed by arrayMap and DIPG data set (MacKay *et al.*, Cancer Cell 2017, in print)
- structural variations (DUP, DEL) in addition to SNV
- diagnosis queries using ontology codes (NCIT, ICD-O)
- quantitative responses
- current version uses **GA4GH schema compatible** database



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	arraymap
Reference name*	9
Genome Assembly*	GRCh36 / hg18
Variant type*	DEL (Deletion)
Start *min* Position*	19,500,000
Start *max* Position	21,964,826
End *min* Position	21,958,228
End *max* Position	24,500,000
Bio-ontology	ncit:c3224: Melanoma (1098)

Beacon Query

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

Response

Dataset	Assembly	Chro	Var Type	Start Range	End Range	Pos	Ref Alt	Bio Query	Variants Calls Samples	$f_{alleles}$	Response Context
arraymap	hg18	9	DEL	19,500,000 21,964,826	21,958,228 24,500,000			ncit:c3224	157 171 171	0.1557	JSON UCSC Handover

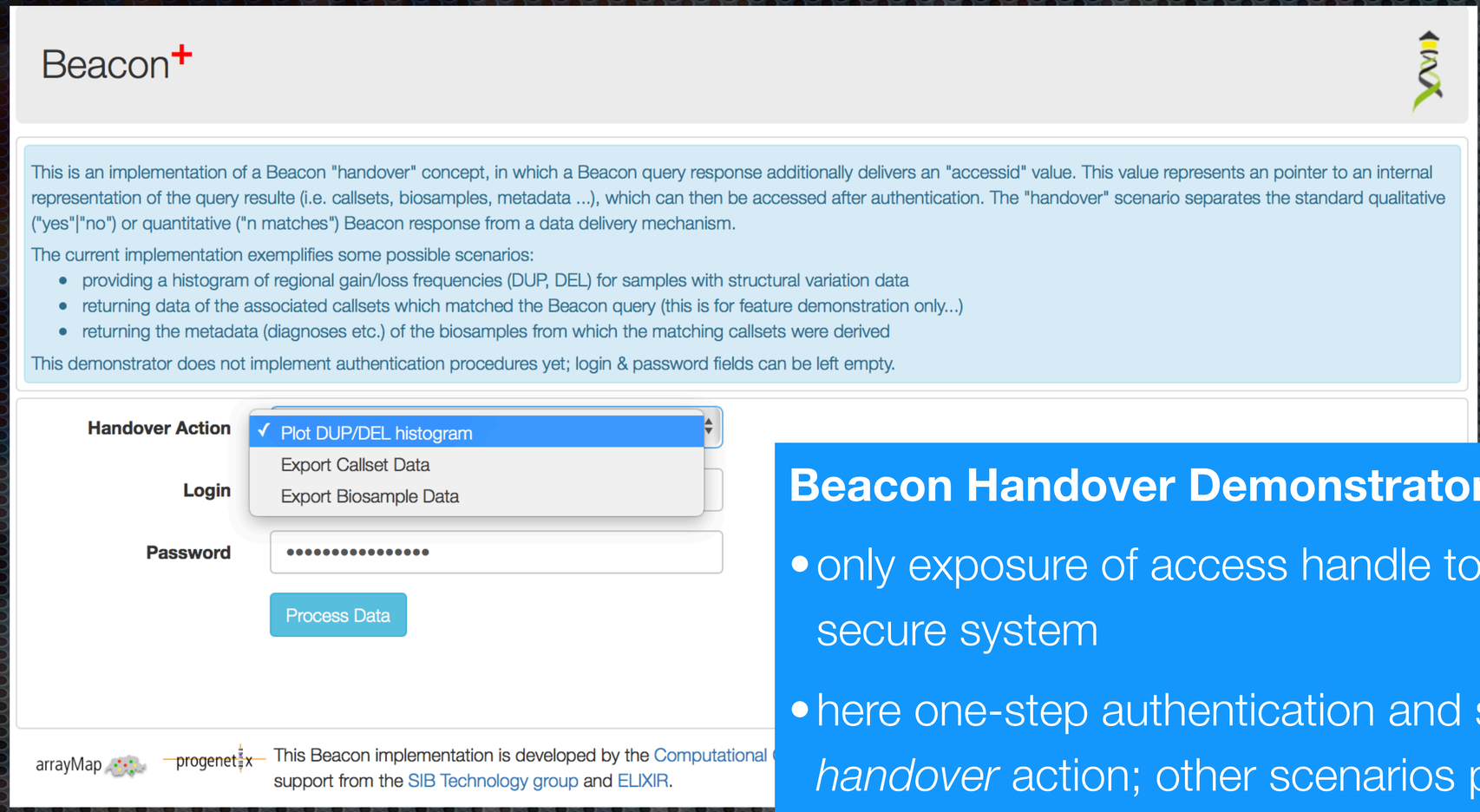
arrayMap progenetix This Beacon implementation is developed by the Computational Oncogenomics Group at the University of Zurich, with support from the SIB Technology group and ELIXIR.

University of Zurich UZH elixir SIB

Beacon Handover Demonstrator

- only exposure of access handle to data stored in secure system
- here one-step authentication and selection of *handover* action; other scenarios possible / likely
- *handover* response outside of Beacon protocol / system


```
  "dataset_allele_responses" : [
    {
      "dataset_id" : "arraymap",
      "call_count" : "171",
      "sample_count" : "171",
      "variant_count" : "157",
     "error" : null,
      "exists" : true,
      "external_url" : "http://beacon.arraymap.org",
      "frequency" : "0.1557",
      "info" : {
        "callset_access_handle" : "d5850347-d411-11e7-8c89-ec436516cb41",
        "description" : "The query was against database \"arraymap_ga4gh\",
variant collection \"variants_cnv_grch36\". 171 matched calls for 157 distinct
variants. Out of 62033 biosamples in the database, 1098 matched the biosample
query; of those, 171 had the variant.",
      },
      "note" : "",
    }
  ],
```

Beacon Handover Demonstrator

- only exposure of access handle to data stored in secure system
- here one-step authentication and selection of *handover* action; other scenarios possible / likely
- *handover* response not managed by Beacon protocol / system => "Discovery" protocol?

Beacon+

This is an implementation of a Beacon "handover" concept, in which a Beacon query response additionally delivers an "accessid" representation of the query resulte (i.e. callsets, biosamples, metadata ...), which can then be accessed after authentication. The ("yes"|"no") or quantitative ("n matches") Beacon response from a data delivery mechanism.

The current implementation exemplifies some possible scenarios:

- providing a histogram of regional gain/loss frequencies (DUP, DEL) for samples with structural variation data
- returning data of the associated callsets which matched the Beacon query (this is for feature demonstration only...)
- returning the metadata (diagnoses etc.) of the biosamples from which the matching callsets were derived

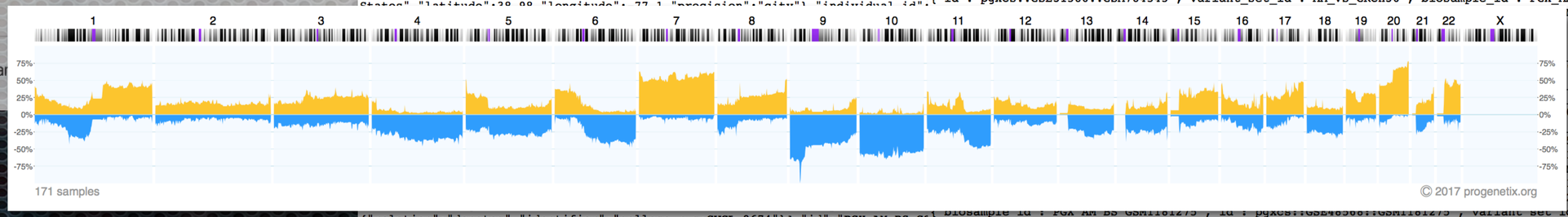
This demonstrator does not implement authentication procedures yet; login & password fields can be left empty.

Handover Action ✓ Plot DUP/DEL his
Export Callset Dat
Export Biosample

Login

Password ••••••••••••••

Process Data



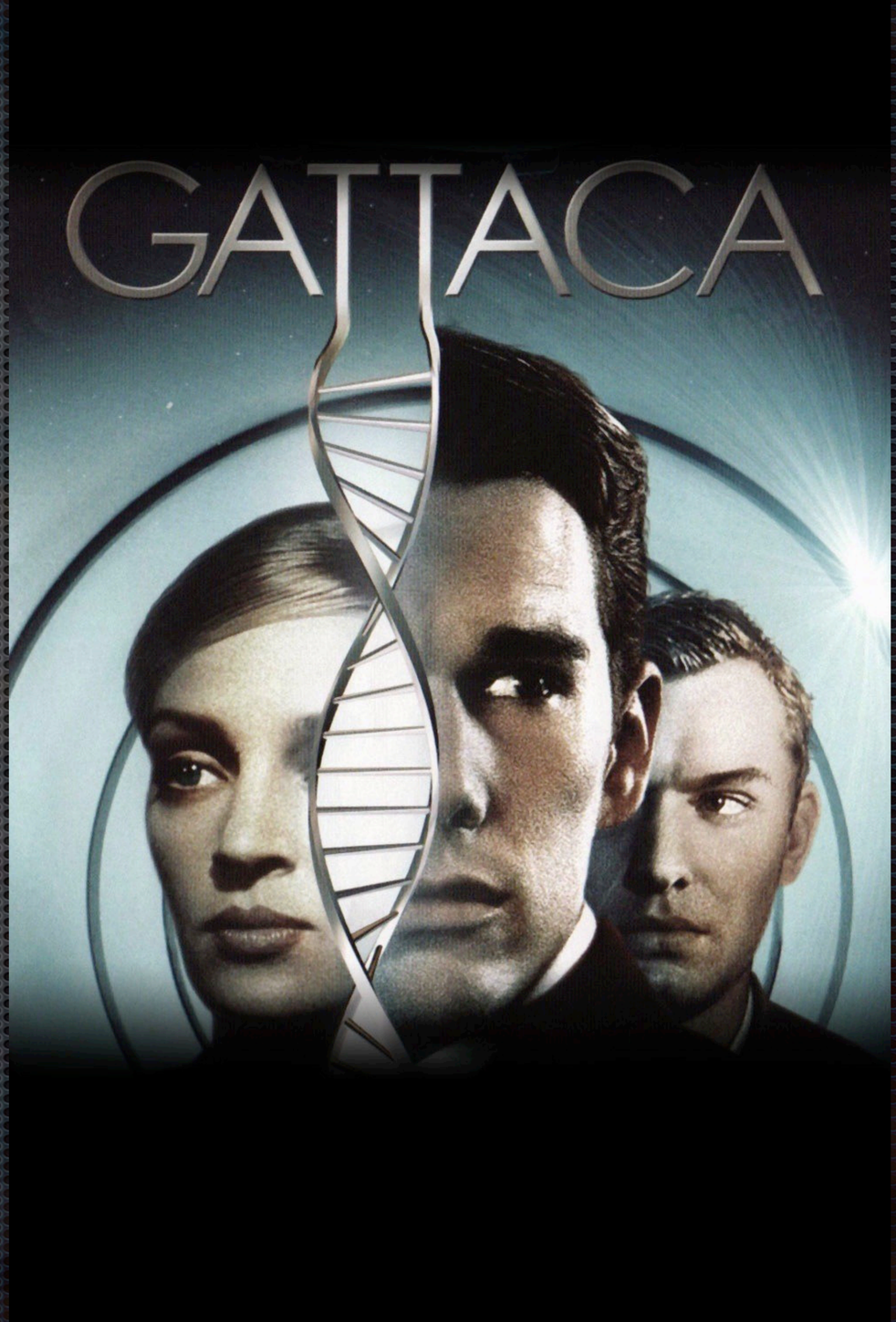
[{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE7822::GSM188260","biosample_id":"PGX_AM_BS_GSM188260"},
{"id":"pgxcs::GSE17534::GSM436995","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM436995"},
{"id":"pgxcs::GSE9003::GSM217839","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM217839"},
{"id":"pgxcs::GSE6779::GSM156007","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM156007"},
{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE7822::GSM188218","biosample_id":"PGX_AM_BS_GSM188218"},
{"biosample_id":"PGX_AM_BS_GSM750822","id":"pgxcs::GSE30291::GSM750822","variant_set_id":"AM_VS_GRCH36"},
{"biosample_id":"PGX_AM_BS_GSM952589","variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE38946::GSM952589"},
{"id":"pgxcs::GSE7822::GSM188283","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM188283"},
{"id":"pgxcs::GSE9003::GSM219026","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM219026"},
{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE44019::GSM1076714","biosample_id":"PGX_AM_BS_GSM1076714"},
{"biosample_id":"PGX_AM_BS_GSM188259","variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE7822::GSM188259"},
{"biosample_id":"PGX_AM_BS_GSM188282","variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE7822::GSM188282"},
{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE9003::GSM228382","biosample_id":"PGX_AM_BS_GSM228382"},
{"biosample_id":"PGX_AM_BS_GSM226864","id":"pgxcs::GSE9003::GSM226864","variant_set_id":"AM_VS_GRCH36"},
{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE22305::GSM555179","biosample_id":"PGX_AM_BS_GSM555179"},
{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE20393::GSM511482","biosample_id":"PGX_AM_BS_GSM511482"},
{"biosample_id":"PGX_AM_BS_GSM437007","variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE17534::GSM437007"},
{"biosample_id":"PGX_AM_BS_GSM188206","id":"pgxcs::GSE7822::GSM188206","variant_set_id":"AM_VS_GRCH36"},
{"biosample_id":"PGX_AM_BS_GSM156027","id":"pgxcs::GSE6779::GSM156027","variant_set_id":"AM_VS_GRCH36"},
{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE20393::GSM511512","biosample_id":"PGX_AM_BS_GSM511512"},
{"id":"pgxcs::GSE41592::GSM1019890","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM1019890"},
{"biosample_id":"PGX_AM_BS_GSM437008","variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE17534::GSM437008"},
{"biosample_id":"PGX_AM_BS_GSM639747","variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE26053::GSM639747"},
{"id":"pgxcs::GSE31586::GSM784559","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM784559"},
{"biosample_id":"PGX_AM_BS_GSM511479","variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE20393::GSM511479"},
{"biosample_id":"PGX_AM_BS_GSM50271","id":"pgxcs::GSE2520::GSM50271","variant_set_id":"AM_VS_GRCH36"},
{"id":"pgxcs::GSE7822::GSM188208","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM188208"},
{"id":"pgxcs::GSE7822::GSM188253","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM188253"},
{"biosample_id":"PGX_AM_BS_GSM952587","id":"pgxcs::GSE38946::GSM952587","variant_set_id":"AM_VS_GRCH36"},
{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE22305::GSM555185","biosample_id":"PGX_AM_BS_GSM555185"},
{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE7822::GSM188236","biosample_id":"PGX_AM_BS_GSM188236"},
{"biosample_id":"PGX_AM_BS_GSM1181277","variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE48568::GSM1181277"},
{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE31586::GSM784512","biosample_id":"PGX_AM_BS_GSM784512"},
{"biosample_id":"PGX_AM_BS_GSM511478","variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE20393::GSM511478"},
{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE7822::GSM188223","biosample_id":"PGX_AM_BS_GSM188223"},
{"id":"pgxcs::GSE19399::GSM482489","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM482489"},
{"biosample_id":"PGX_AM_BS_GSM1076715","id":"pgxcs::GSE44019::GSM1076715","variant_set_id":"AM_VS_GRCH36"},
{"id":"pgxcs::GSE31586::GSM784545","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM784545"},

[{"location":{"longitude":-118.49,"precision":"city","label":"Santa Monica, Uni
States","latitude":34.02},"individual_id":"PGX_IND_GSM1076715","name":"PGX_AM_
[{"term_label":"Melanoma","term_id":"ncit:C3224"},{"term_id":"pgx:icdom:8720_3
{"term_label":"Melanoma of the Skin","term_id":"pgx:seer:25010"}],"description
[{"identifier":"geo:GPL6801","relation":"denotes"},{"identifier":"geo:GSM10767
{"relation":"denotes","identifier":"geo:GSE44019"}],"individual_age_at_collecti
metastasis]","id":"PGX_AM_BS_GSM1076715"},{"bio_characteristics":[{"descriptior
[{"term_label":"Melanoma","term_id":"ncit:C3224"},{"term_label":"Malignant mela
{"term_id":"pgx:seer:25010","term_label":"Melanoma of the Skin"}]}],"name":"PGX
United States","latitude":39.95},"individual_id":"PGX_IND_GSM952606","id":"PGX_
{"age":"","age_class":"null"},"description":"malignant melanoma [cell line WM3
{"relation":"denotes","identifier":"geo:GSM952606"},{"identifier":"geo:GPL9777
{"id":"PGX_AM_BS_GSM750824","individual_age_at_collection":{"age_class":"null",
[{"identifier":"cellosaurus:CVCL_0069","relation":"denotes"},{"relation":"denot
{"relation":"denotes","identifier":"geo:GPL13786"}],"bio_characteristics":[{"de
[{"term_label":"Melanoma","term_id":"ncit:C3224"},{"term_id":"pgx:icdom:8720_3
{"term_id":"pgx:seer:25010","term_label":"Melanoma of the Skin"}]}],"name":"PGX

{"relation":"denotes","identifier":"cellosaurus:CVCL_9674"}],"id":"PGX_AM_BS_G
{"label":"Lund, Sweden","latitude":55.71,"longitude":13.19,"precision":"city"}
[{"term_id":"ncit:C3224","term_label":"Melanoma"},{"term_id":"pgx:icdom:8720_3
{"term_id":"pgx:seer:25010","term_label":"Melanoma of the Skin"}],"negated_ontc
{"latitude":48.21,"label":"Vienna, Austria","longitude":16.37,"precision":"city
[{"term_id":"ncit:C3224","term_label":"Melanoma"},{"term_label":"Malignant melanoma, NOS","term_id":"pgx:icdom:8720_3"},{"term_id":"pgx:icdot:C44","term_label":"skin"},
{"term_id":"pgx:seer:25010","term_label":"Melanoma of the Skin"}],"negated_ontology_terms":[],"description":"malignant melanoma [cell line MM1]"}],"external_identifiers":
[{"identifier":"geo:GSM557941","relation":"denotes"},{"identifier":"cellosaurus:CVCL_2075","relation":"denotes"},{"identifier":"pubmed:21584902","relation":"denotes"},
{"relation":"denotes","identifier":"geo:GPL6801"},{"relation":"denotes","identifier":"geo:GSE22461"}],"description":"malignant melanoma [cell line
MM1]","individual_age_at_collection":{"age":"","age_class":"null"},"id":"PGX_AM_BS_GSM557941"},{"bio_characteristics":[{"negated_ontology_terms":[],"ontology_terms":
[{"term_id":"ncit:C3224","term_label":"Melanoma"},{"term_label":"Malignant melanoma, NOS","term_id":"pgx:icdom:8720_3"},{"term_label":"skin","term_id":"pgx:icdot:C44"},
{"term_label":"Melanoma of the Skin","term_id":"pgx:seer:25010"}],"description":"malignant melanoma [cell line WM3438]"}],"location":
{"longitude":-75.16,"precision":"city","latitude":39.95,"label":"Philadelphia, United

{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE7822::GSM188230","biosample_id":"PGX_AM_BS_GSM188230"},
{"id":"pgxcs::GSE7822::GSM188264","variant_set_id":"AM_VS_GRCH36","biosample_id":"PGX_AM_BS_GSM188264"},
{"biosample_id":"PGX_AM_BS_GSM784542","id":"pgxcs::GSE31586::GSM784542","variant_set_id":"AM_VS_GRCH36"},
{"variant_set_id":"AM_VS_GRCH36","id":"pgxcs::GSE2520::GSM50202","biosample_id":"PGX_AM_BS_GSM50202"},

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

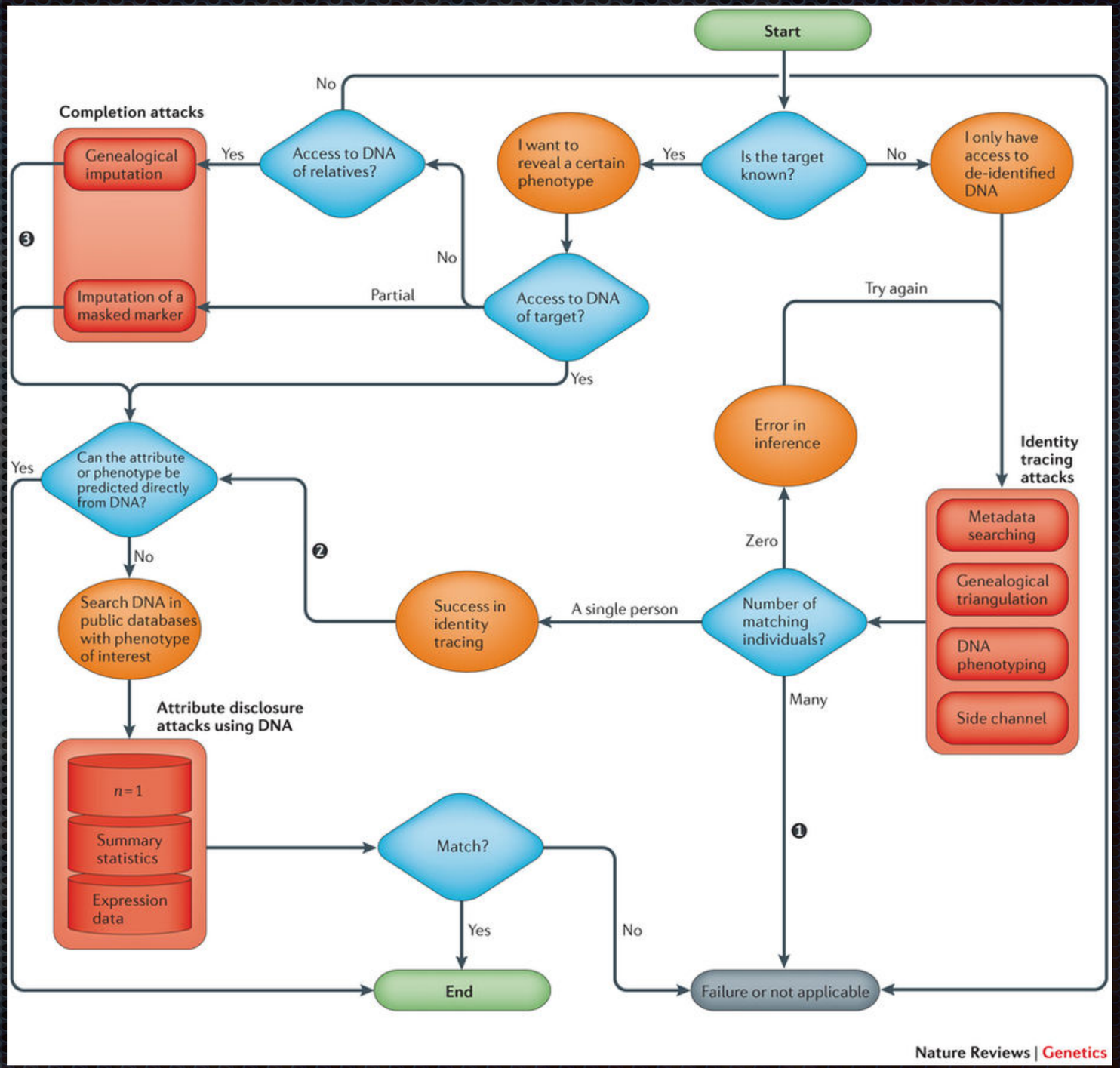
SAFETY

CRYPTOGRAPHY

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)



IDENTIFICATION OF INDIVIDUALS BASED ON "GENOMIC FINGERPRINTS"

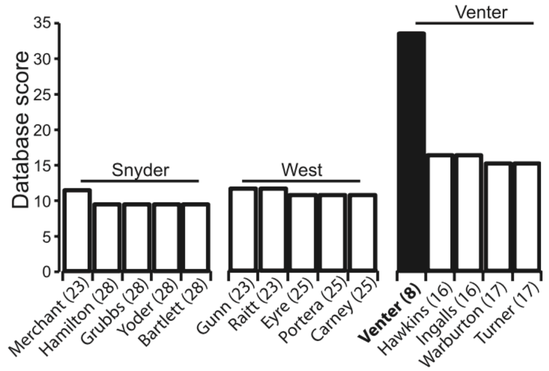
18 JANUARY 2013 VOL 339 SCIENCE www.sciencemag.org

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

Genome ***Beacons*** Compromise Security?

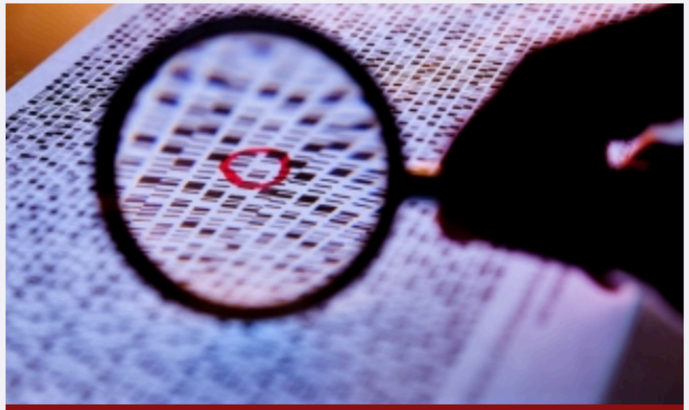
Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals

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.

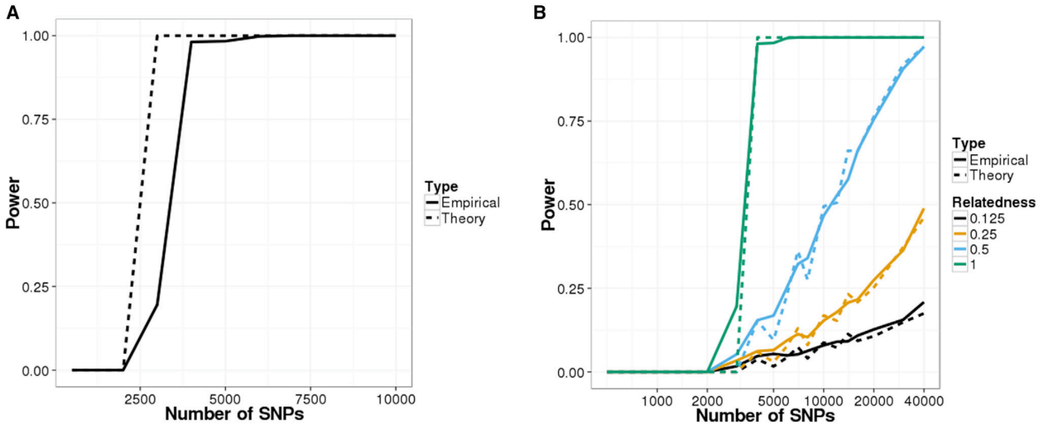


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

- 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

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

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

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

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)

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

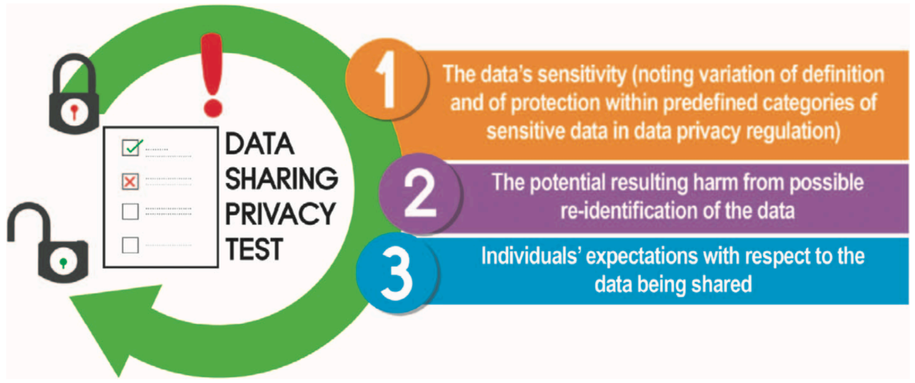
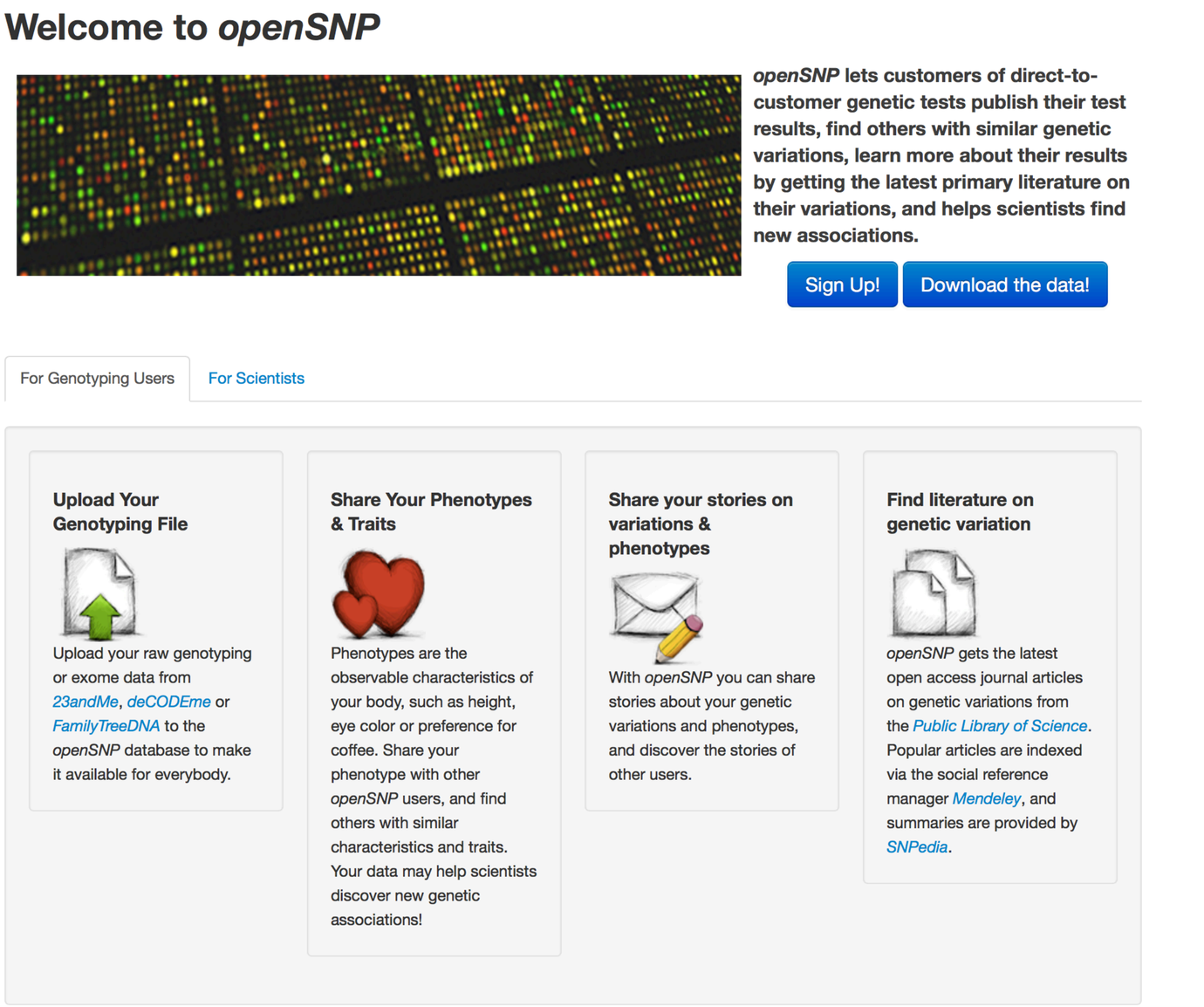


Figure 1. The three steps of a Data Sharing Privacy Test to distinguish degrees of data sensitivity within categories of data recognised as 'sensitive'.

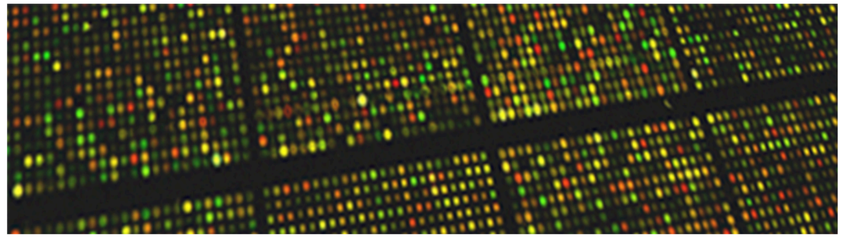
SHARE YOUR GENOME DATA?

- depositing genome data has the inherent risk of being identified and linked to your person
- What are the Risks?
- Would you contribute e.g. to OpenSNP?
- Discuss!



SHARE YOUR GENOME DATA?

Welcome to *openSNP*



openSNP **lets customers of direct-to-customer genetic tests publish their test results, find others with similar genetic variations, learn more about their results by getting the latest primary literature on their variations, and helps scientists find new associations.**

Sign Up! Download the data!

For Genotyping Users | For Scientists

Upload Your Genotyping File

Upload your raw genotyping or exome data from *23andMe*, *deCODEme* or *FamilyTreeDNA* to the *openSNP* database to make it available for everybody.

Share Your Phenotypes & Traits

Phenotypes are the observable characteristics of your body, such as height, eye color or preference for coffee. Share your phenotype with other *openSNP* users, and find others with similar characteristics and traits. Your data may help scientists discover new genetic associations!

Share your stories on variations & phenotypes

With *openSNP* you can share stories about your genetic variations and phenotypes, and discover the stories of other users.

Find literature on genetic variation

openSNP gets the latest open access journal articles on genetic variations from the *Public Library of Science*. Popular articles are indexed via the social reference manager *Mendeley*, and summaries are provided by *SNPedia*.

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

arraymap.org

progenetix.org

sib.swiss/baudis-michael

imls.uzh.ch/en/research/baudis