Personalized medicine in cancer

Data Formats | Genome Variation | Techniques | Resources | Sharing

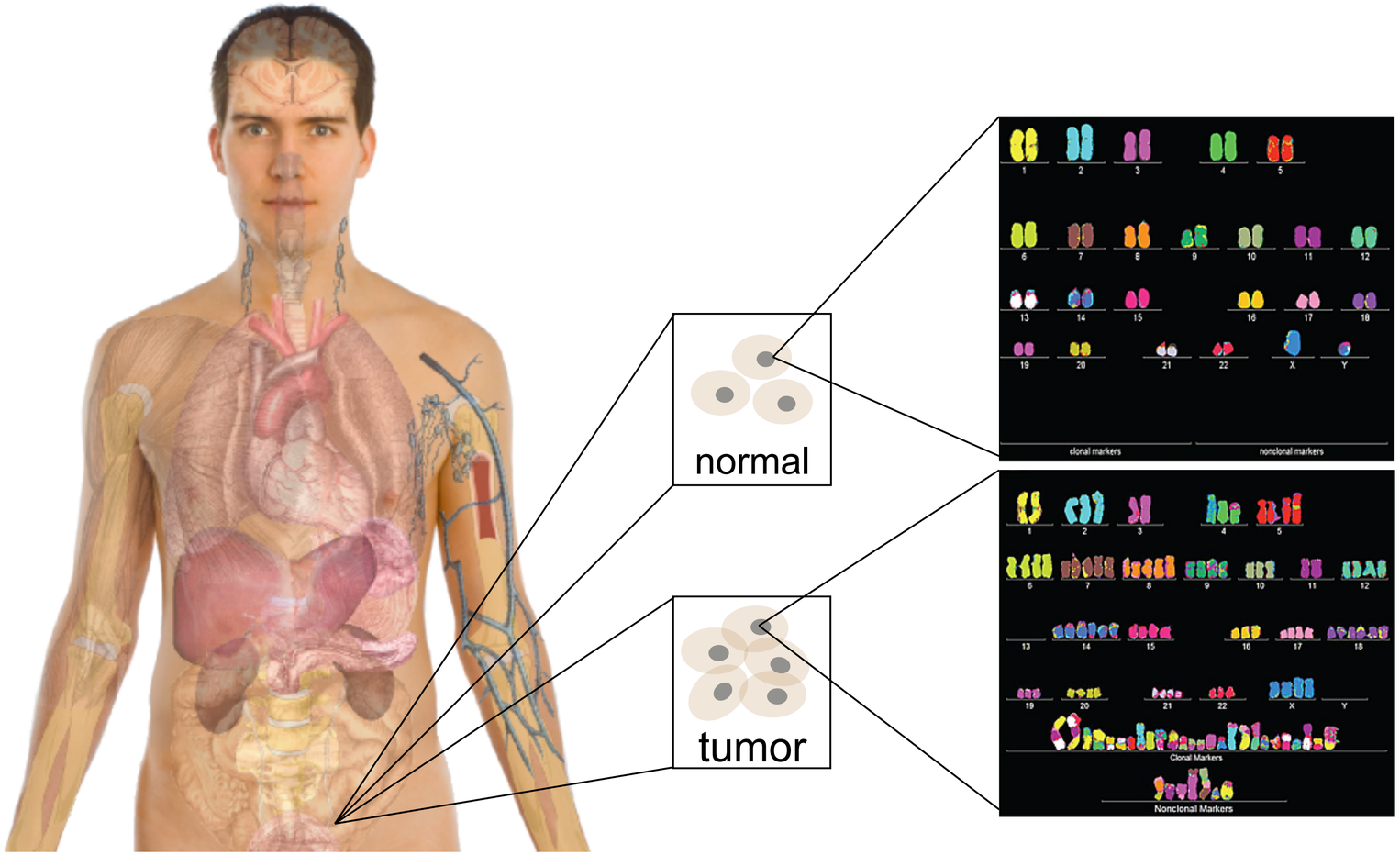
Michael Baudis **UZH SIB**
Computational Oncogenomics

University of Zurich[UZH]

Swiss Institute of Bioinformatics

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™

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

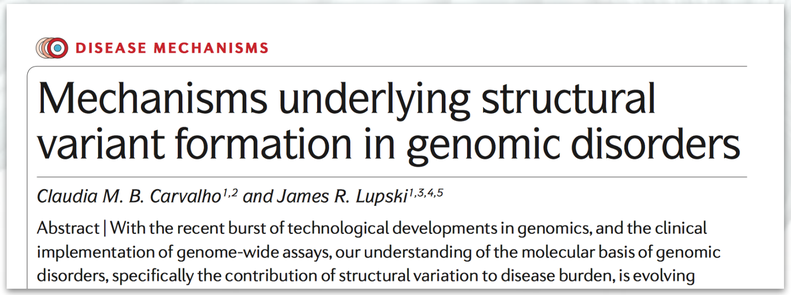

doi:10.1038/nature19057

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

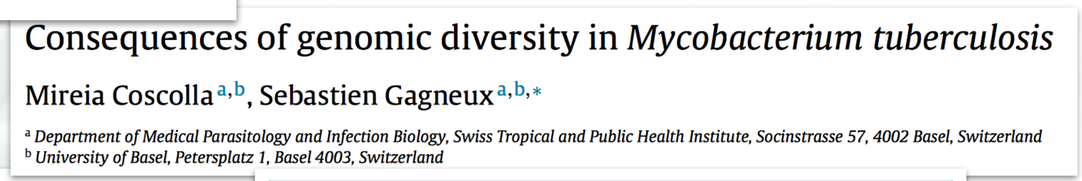
Genomic Classification of Cutaneous Melanoma

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

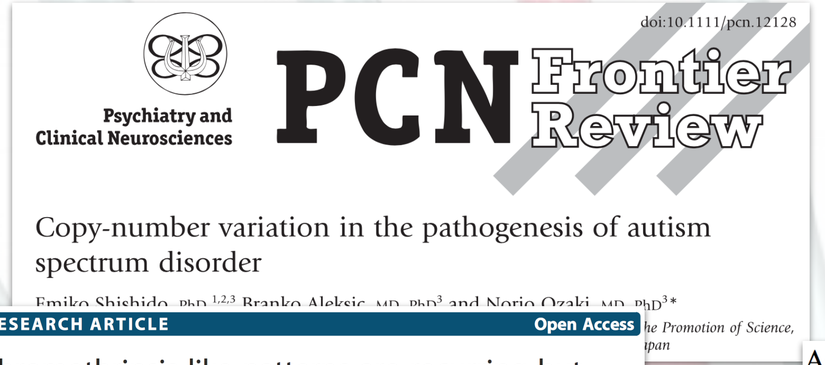


Consequences of genomic diversity in *Mycobacterium tuberculosis*

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

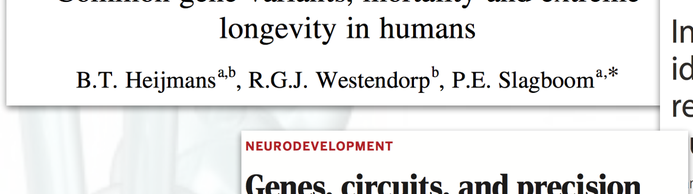
Psychiatry and Clinical Neurosciences

PCN Frontier Review

Copy-number variation in the pathogenesis of autism spectrum disorder

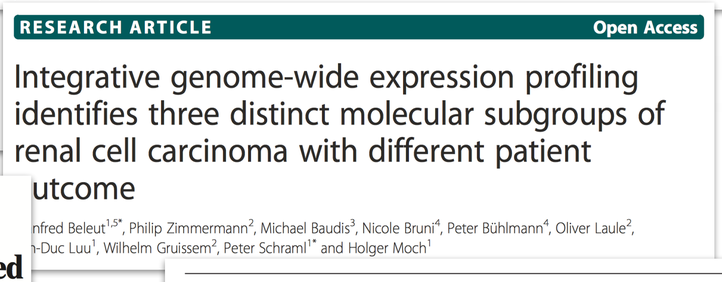
Emiko Shishido, PhD[1,2,3], Branko Aleksic, MD, PhD[3] and Norio Ozaki, MD, PhD[3]*

he Promotion of Science, apan



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

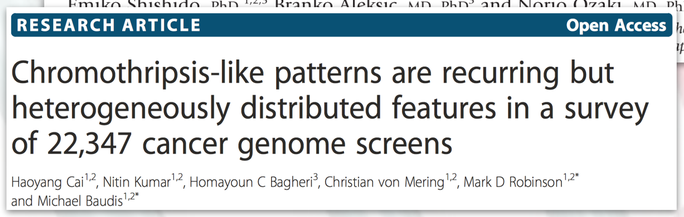
Integrative genome-wide expression profiling identifies three distinct molecular subgroups of renal cell carcinoma with different patient utcome

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

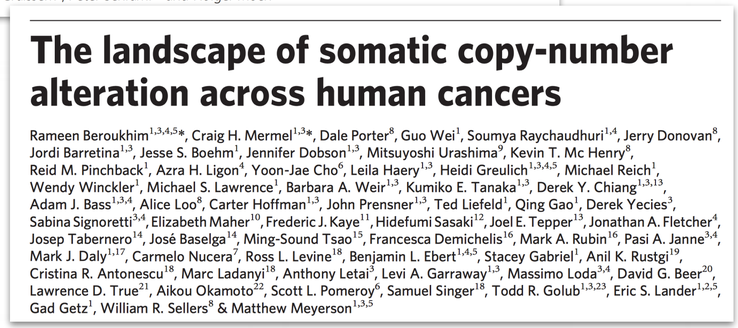
Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens

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 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

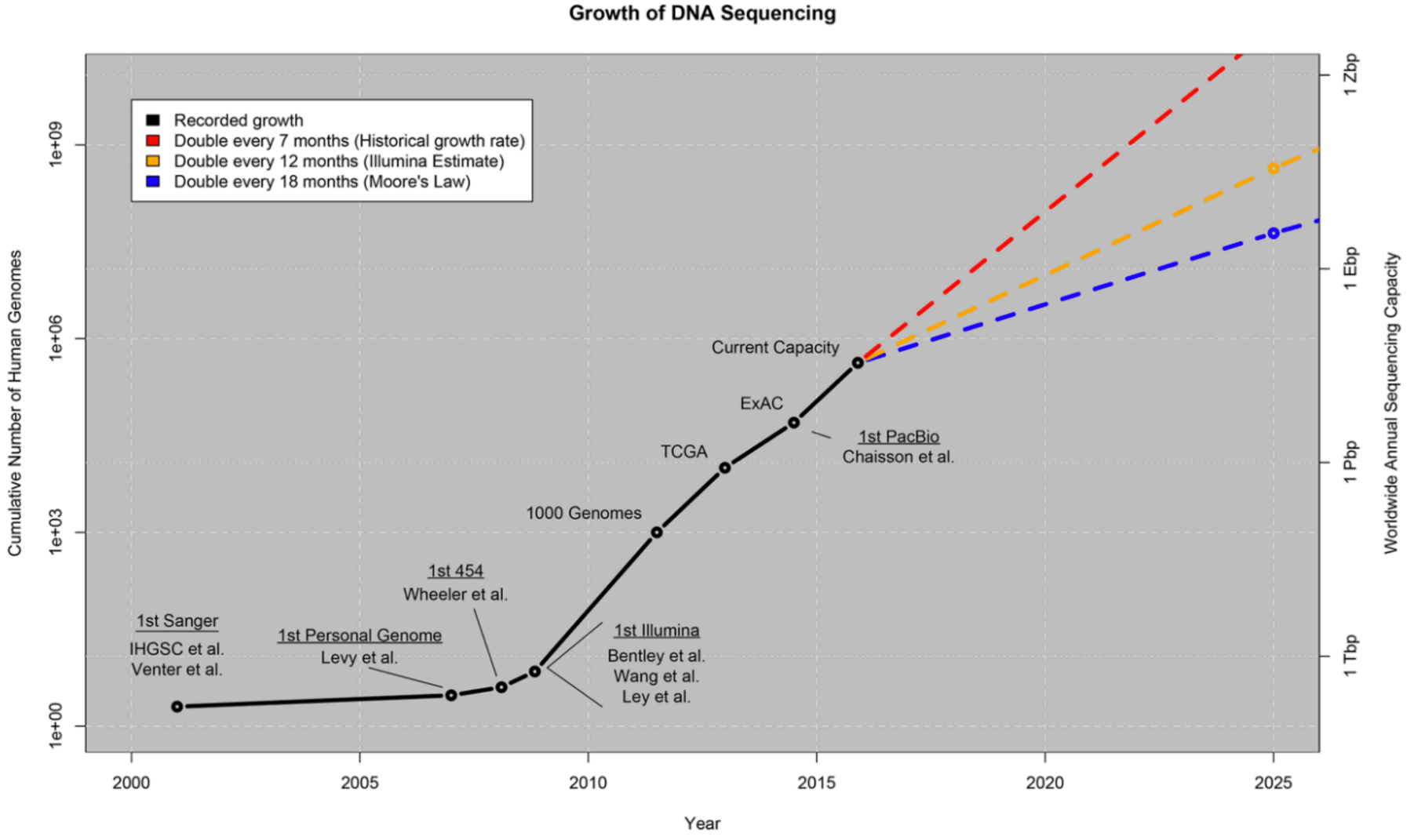


Fig 1. Growth of DNA sequencing. The plot shows the growth of DNA sequencing both in the total number of human genomes sequenced (left axis) as well as the worldwide annual sequencing capacity (right axis: Tera-basepairs (Tbp), Peta-basepairs (Pbp), Exa-basepairs (Ebp), Zetta-basepairs (Zbps)). The values 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.

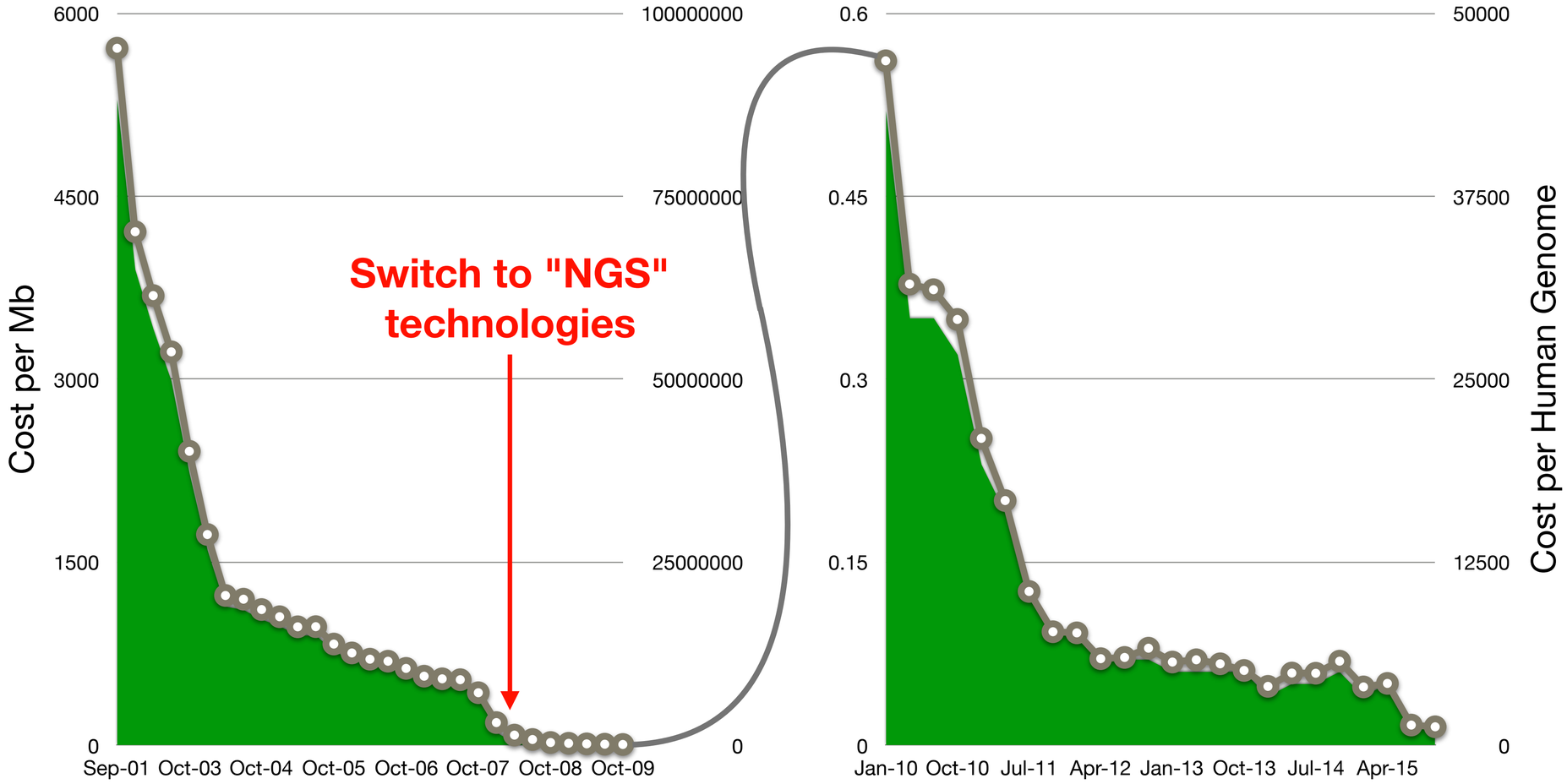
Citation: Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195

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

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

DNA Sequencing Costs: Data

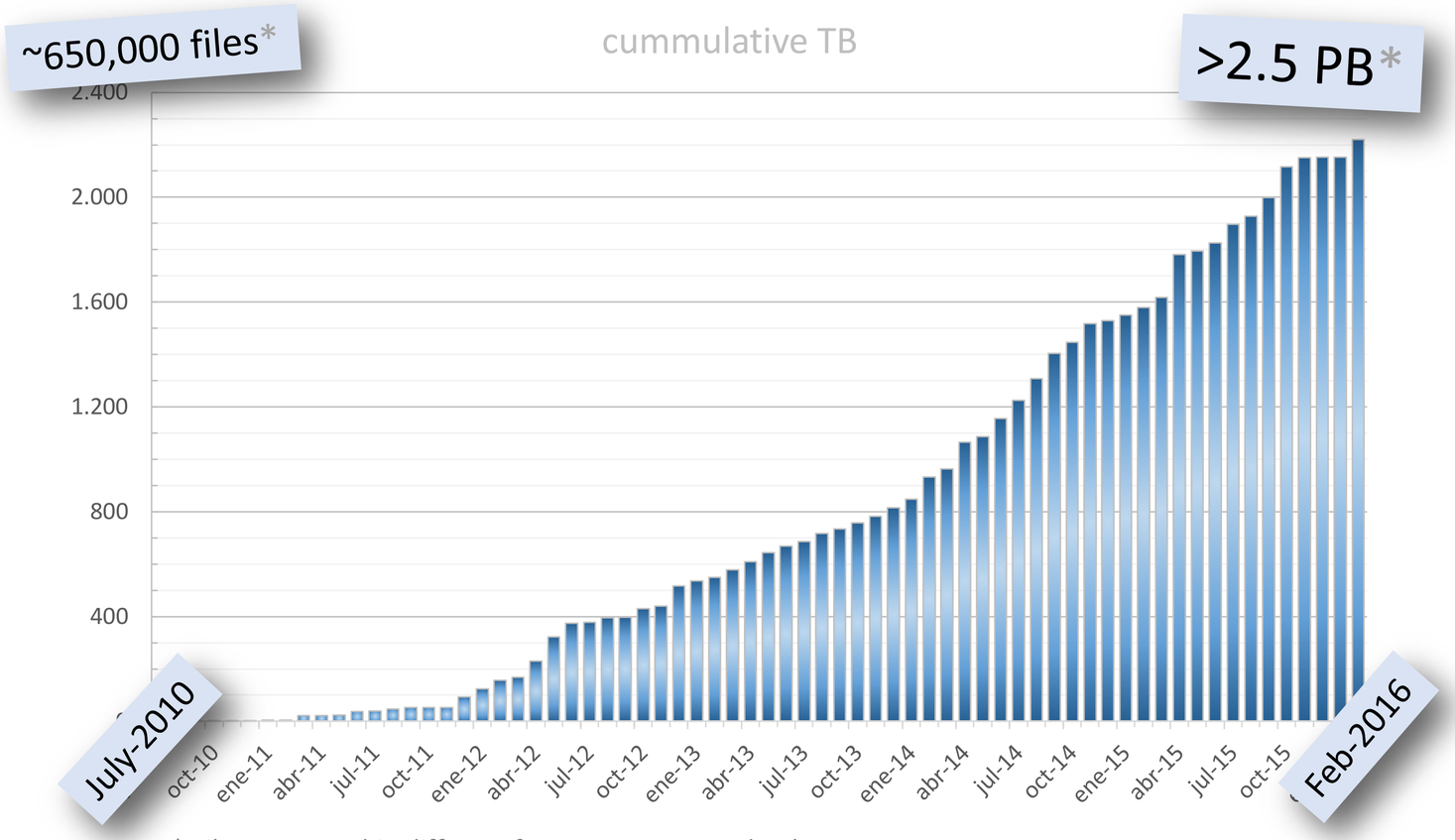
Data from the NHGRI Genome Sequencing Program (GSP)



- 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

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

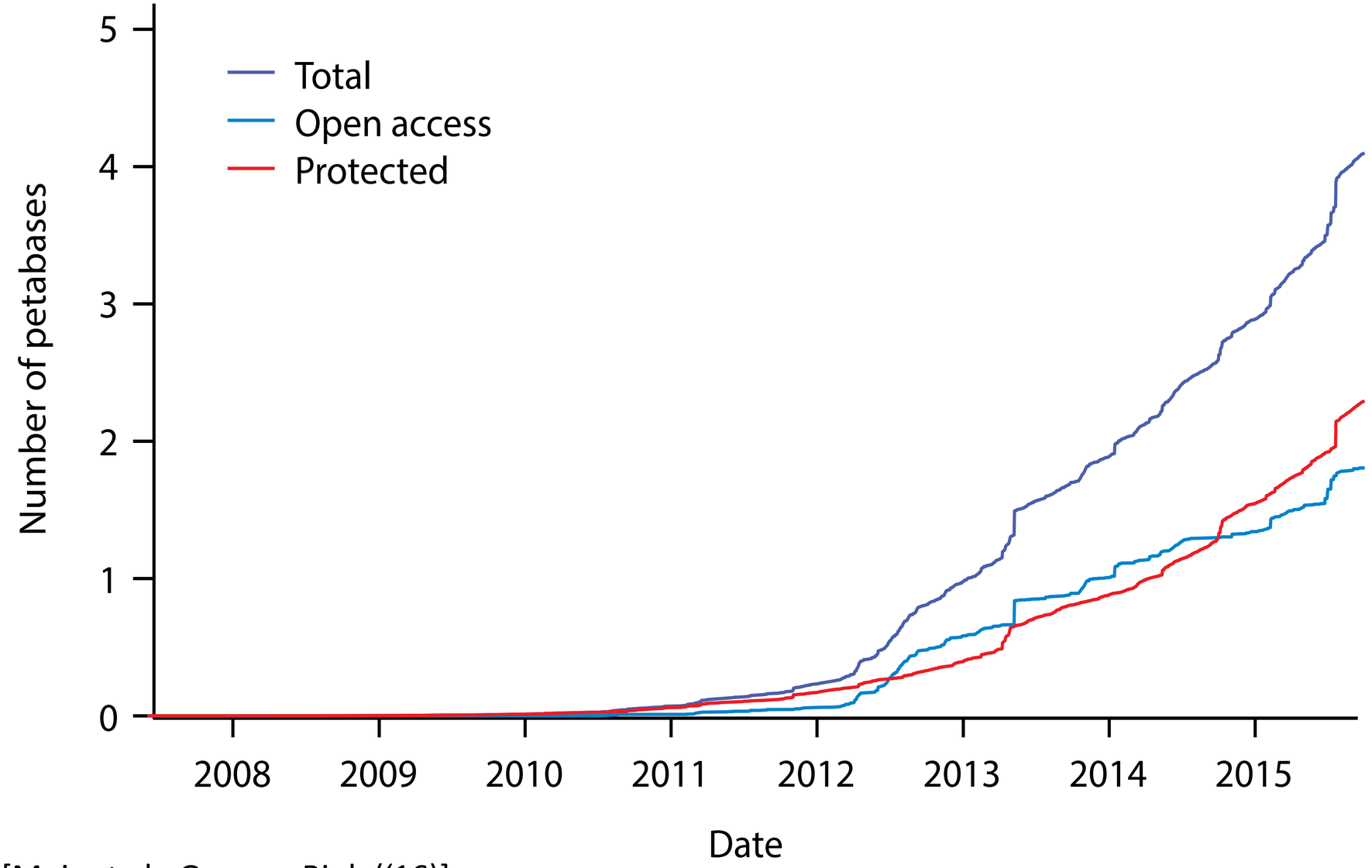
Source: Jordi Rambla (CRG Barcelona) 2016

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 digitec.ch (Dec 2016) ~42'000CHF (100x10TB disks)
- this is less than a PhD position per year in Switzerland …
- (real costs are 2x that, + duplication, facilities, service … => ~500’000CHF
- **However: 30x BAM file => 100GB**
- Still: 500’000CHF => 1PB => 10’000 genomes => 50CHF/genome (BAM format)

419.–
Seagate IronWolf
(10000GB, 3.5",

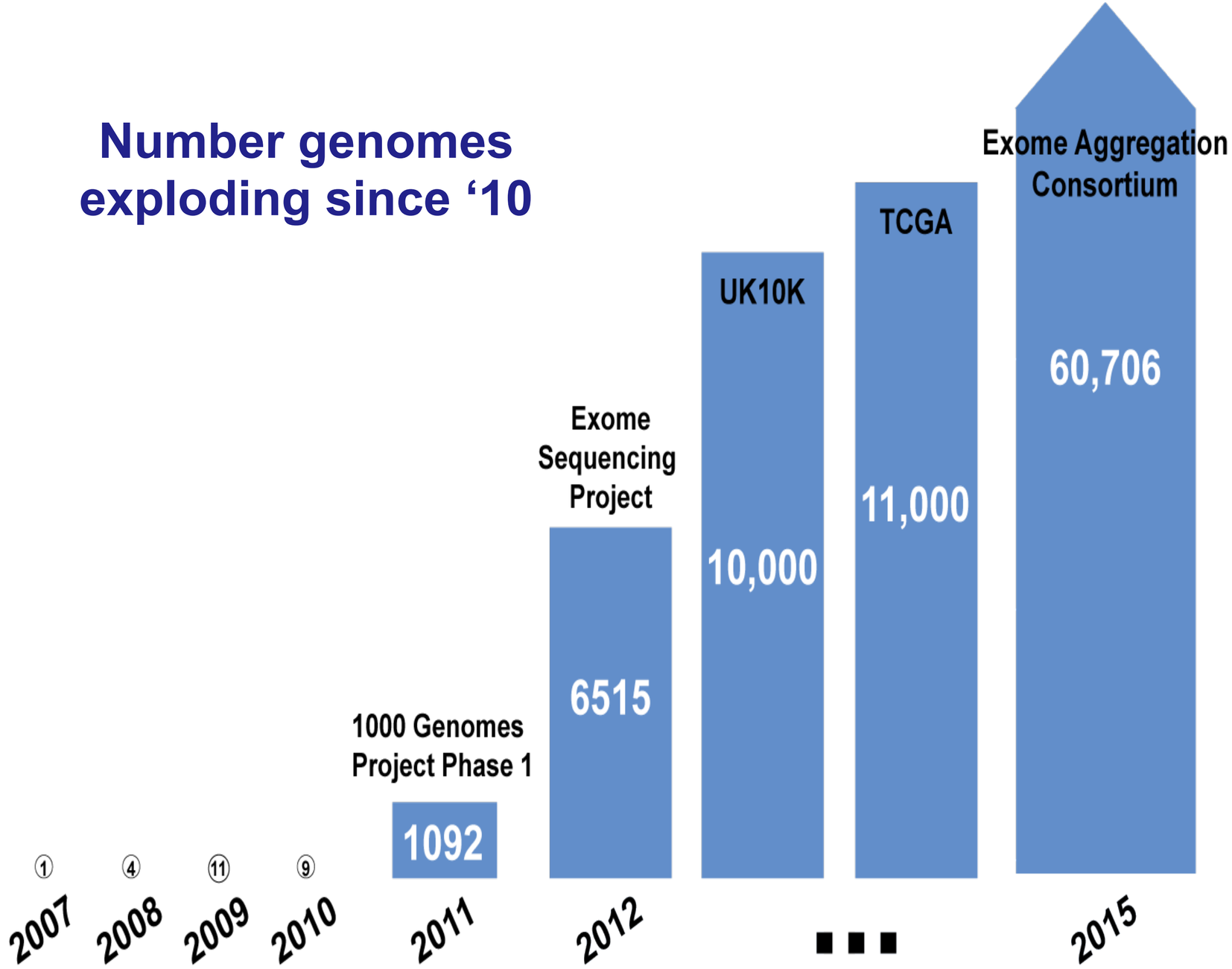
Increase in number of bases in SRA, Peta-scale after ‘10



[Muir et al., GenomeBiol. (‘16)]

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

Number genomes exploding since ‘10

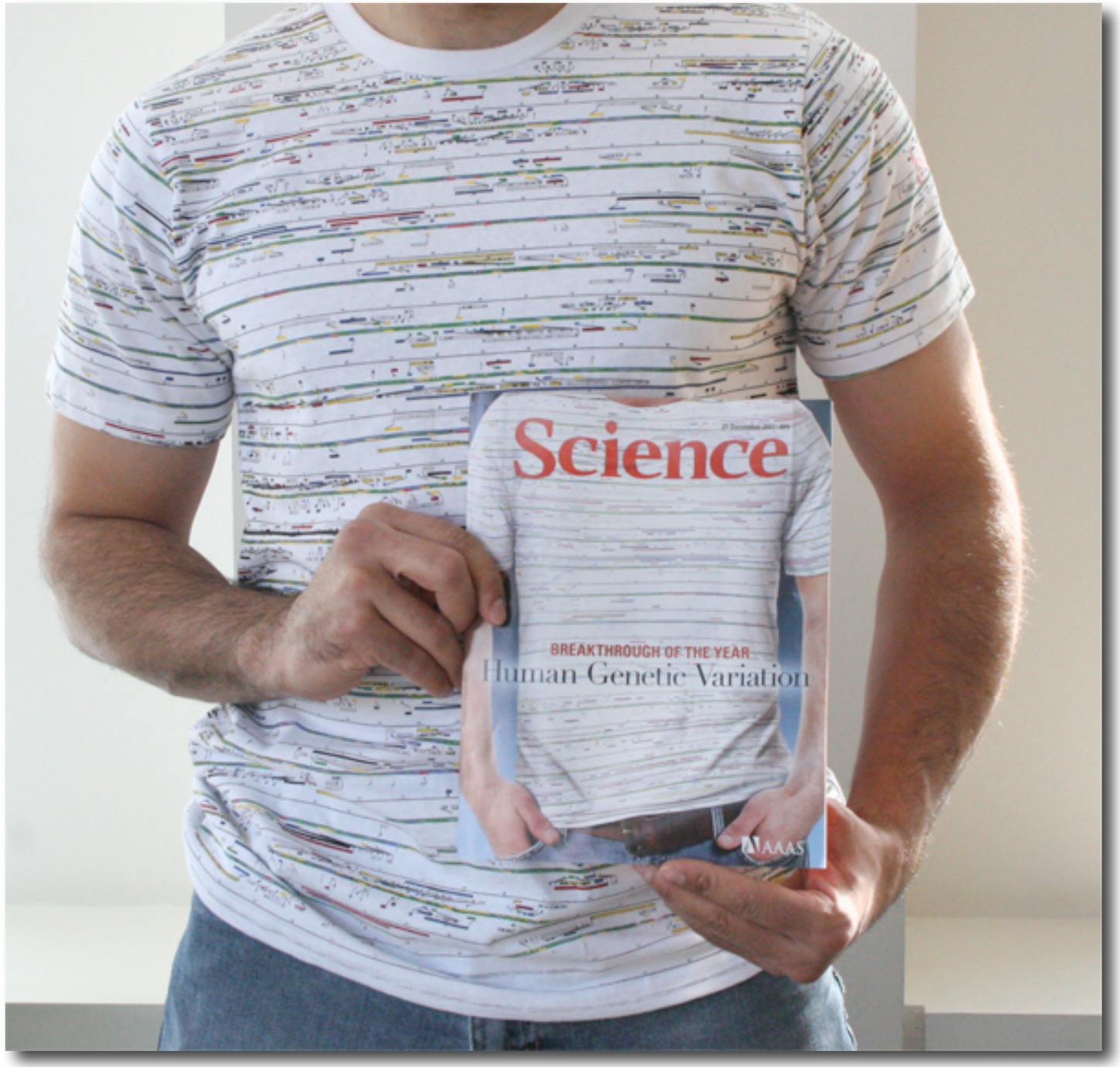


Genomes Everywhere

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

The trouble with human genome variation



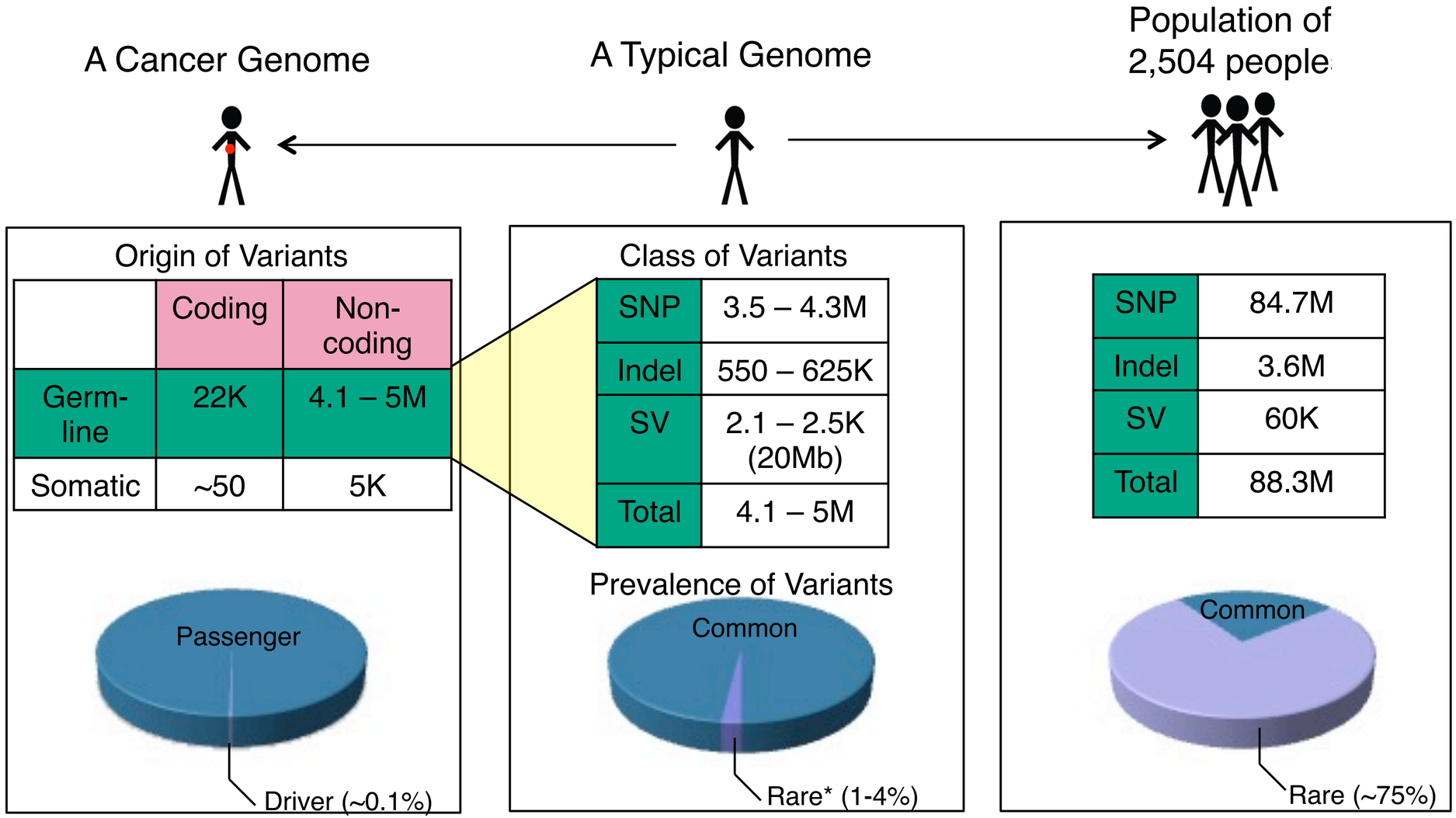
Conclusions from the analysis of variation in the human genome

- 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

Human Genetic Variation

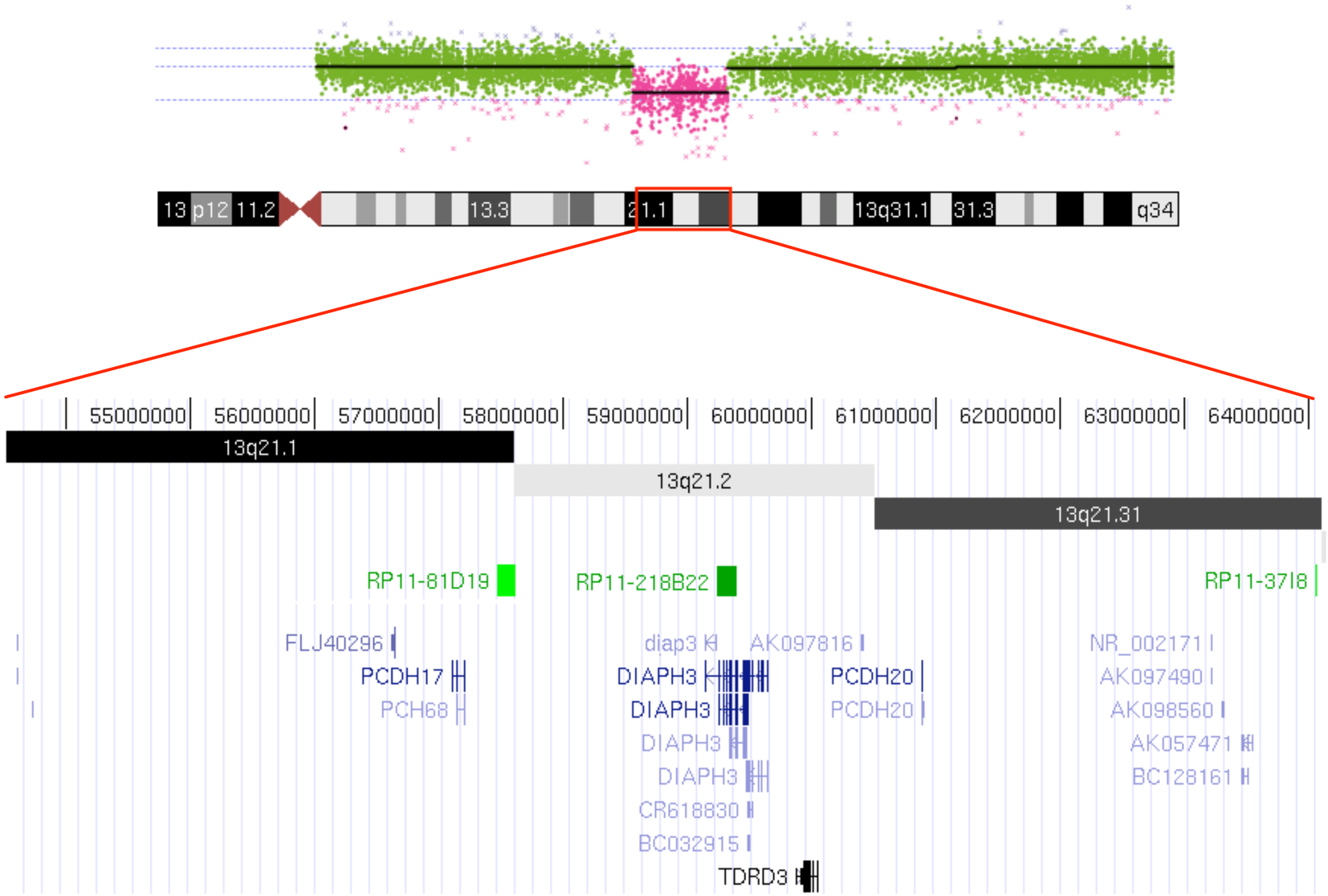


* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

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

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)

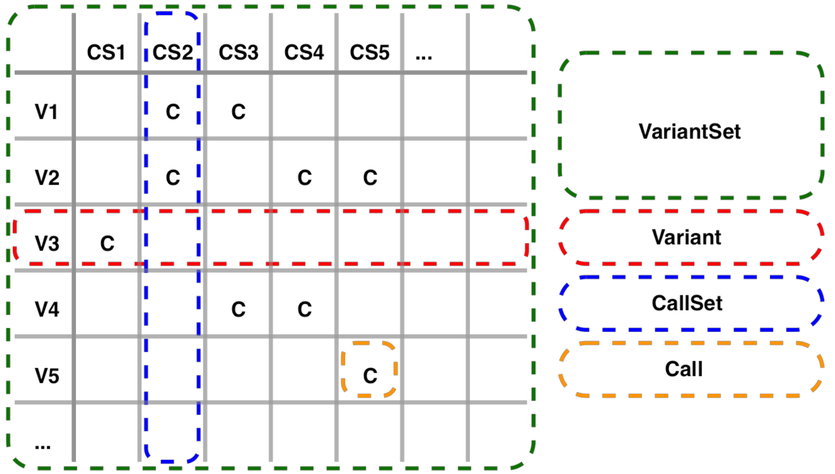


REFERENCE RESOURCES FOR HUMAN GENOME VARIANTS

- 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...

The Variant Call Format (VCF)

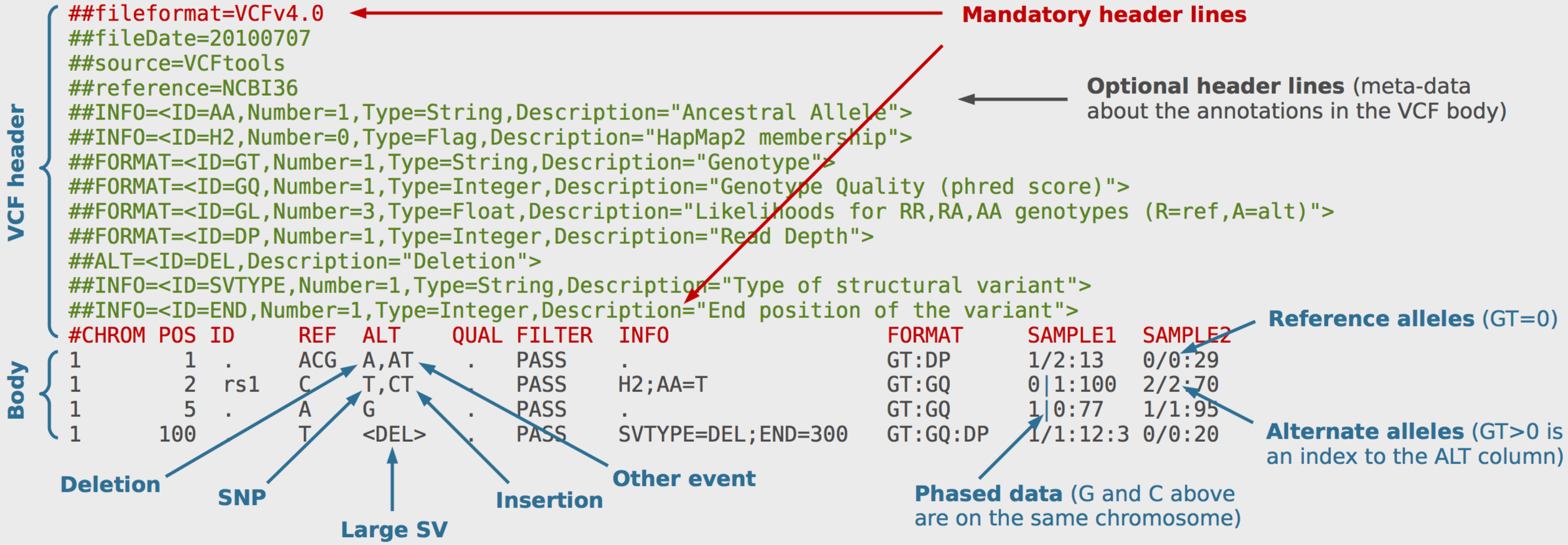
```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```



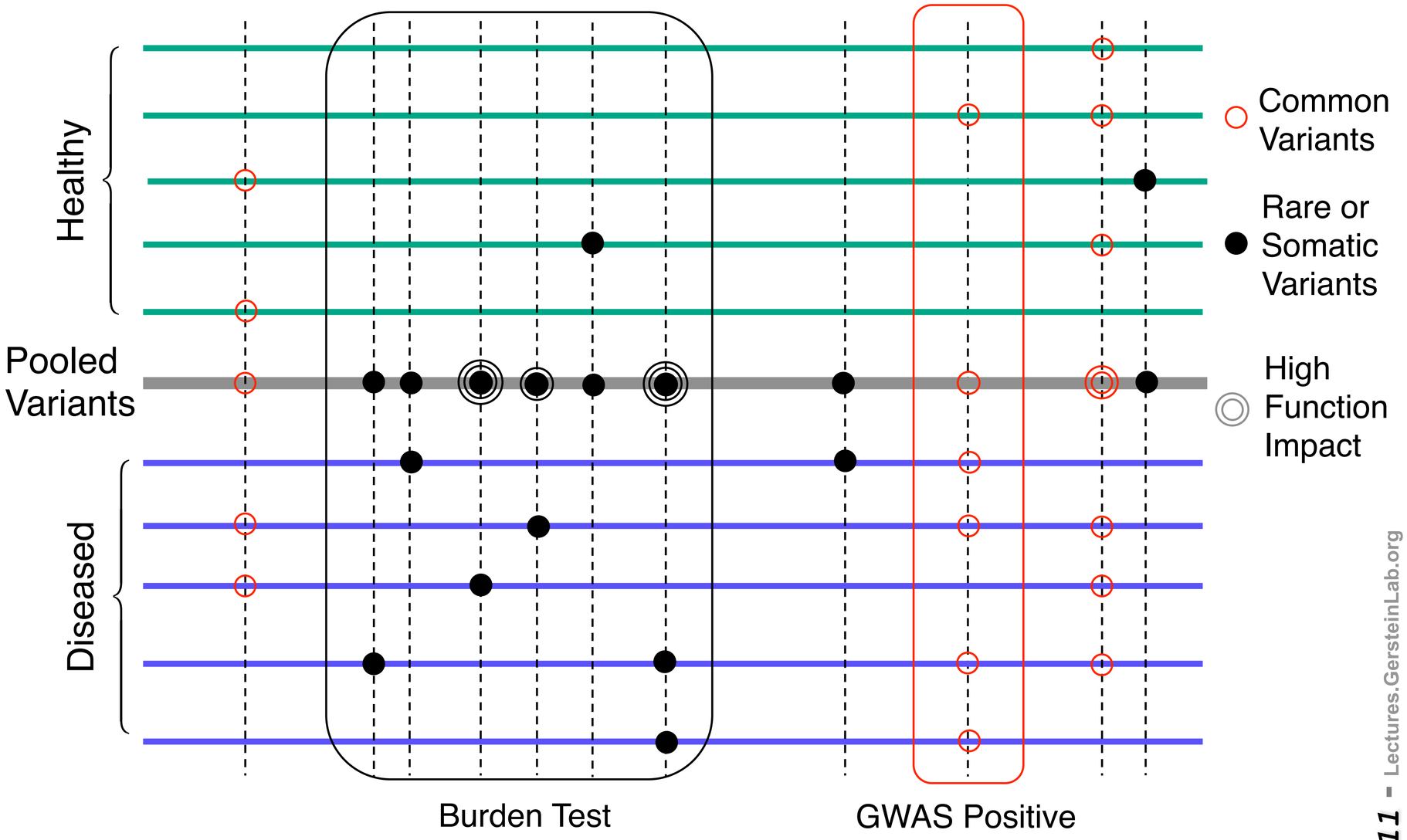
- stores the results of a single or multiple interpretations of genome sequencing datasets, in comparison to a reference genome

The VCF file format: Standard for variant calling

Example



Association of Variants with Diseases



i.e., directly observed in the diseased biosample/individual

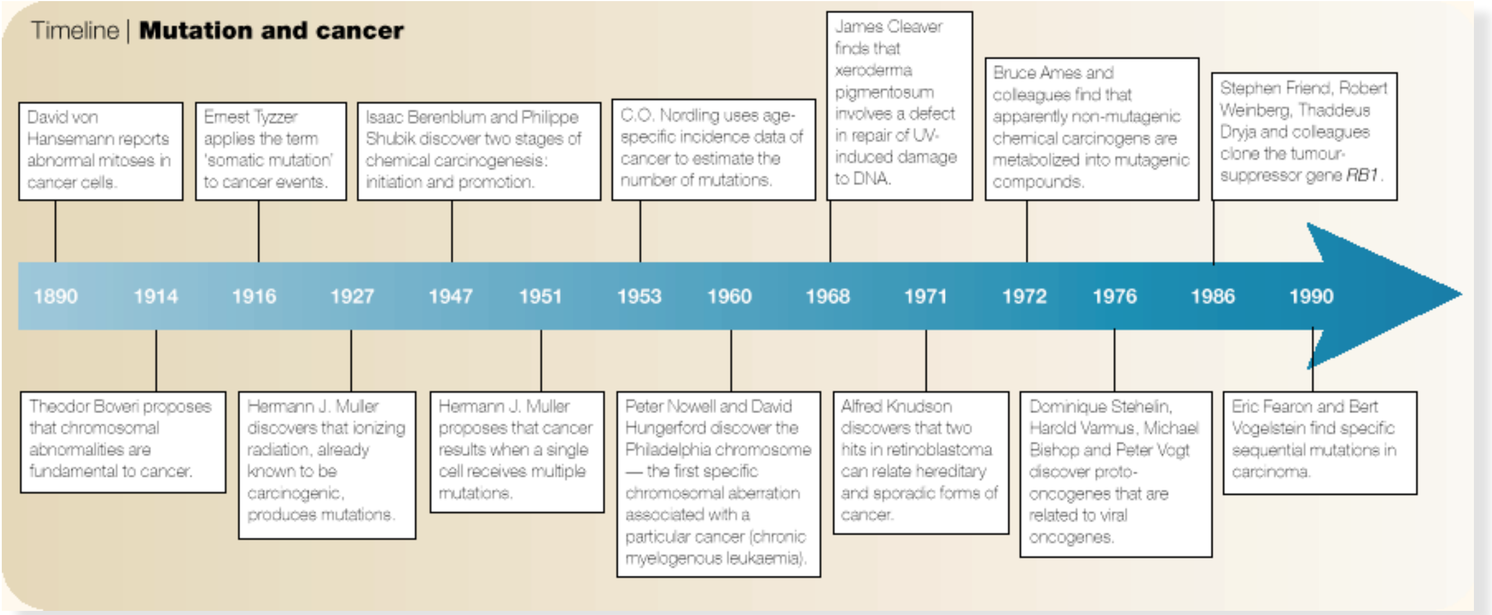
Genome Data for Cancer Research and Personalized Health™

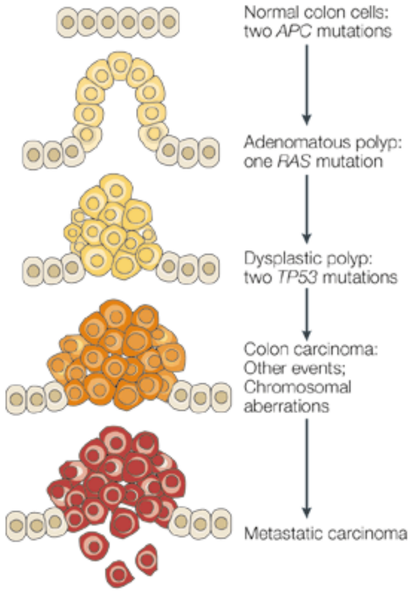
Repositories, Biocuration, Standards, Metadata, Accessibility

Michael Baudis **UZH SIB**
Computational Oncogenomics

University of Zurich[UZH]

SIB Swiss Institute of Bioinformatics

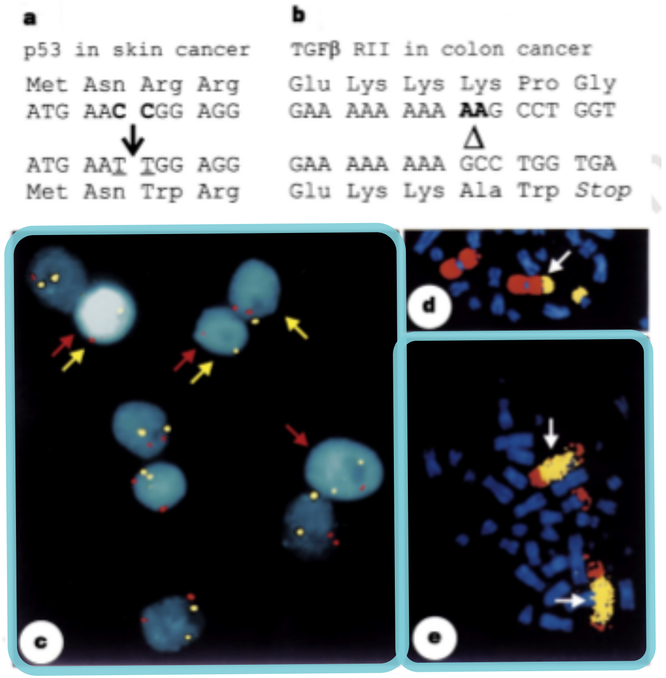




Cancers are based on acquired and inherited genomic mutations

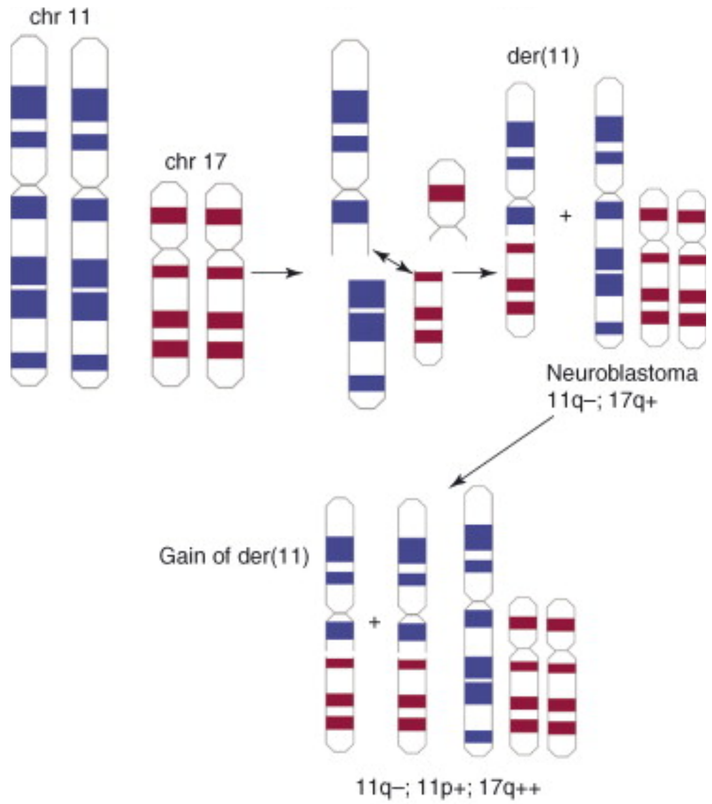
Knudson, A. G. (2001). Two genetic hits (more or less) to cancer. Nature Reviews Cancer, 1(2), 157–162.

Mutations & genomic rearrangements in cancer



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

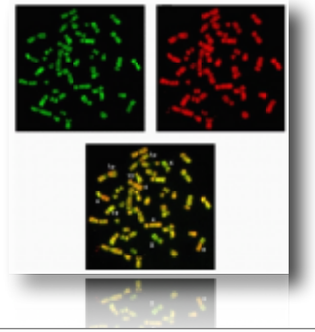
a. small mutation (di-pyrimidine exchange at p53 in Xeroderma pigmentosum patient)
b. two-base deletion in *TGFB* in a colorectal cancer patient with mismatch repair deficiency
c. chromosomal losses (FISH; red=3, yellow=12) in CRC
d. t(1;17) in neuroblastoma, whole-chromosomal painting
e. *MYCN* gene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

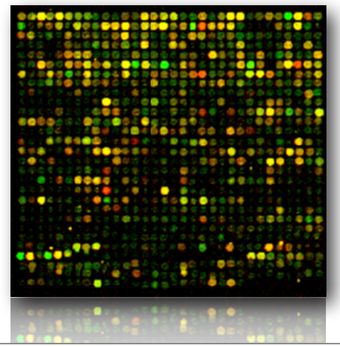


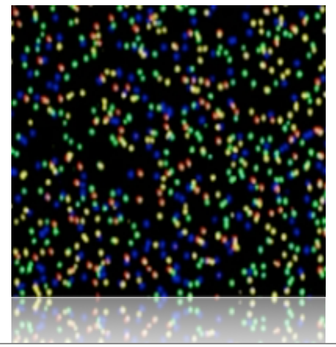
Generation of copy number imbalances in cancer through imbalanced cytogenetic rearrangements - partial deletion of 11q, gain of 11pterq21 and 2 addl. copies of 17q

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

WHOLE GENOME SCREENING IN CANCER

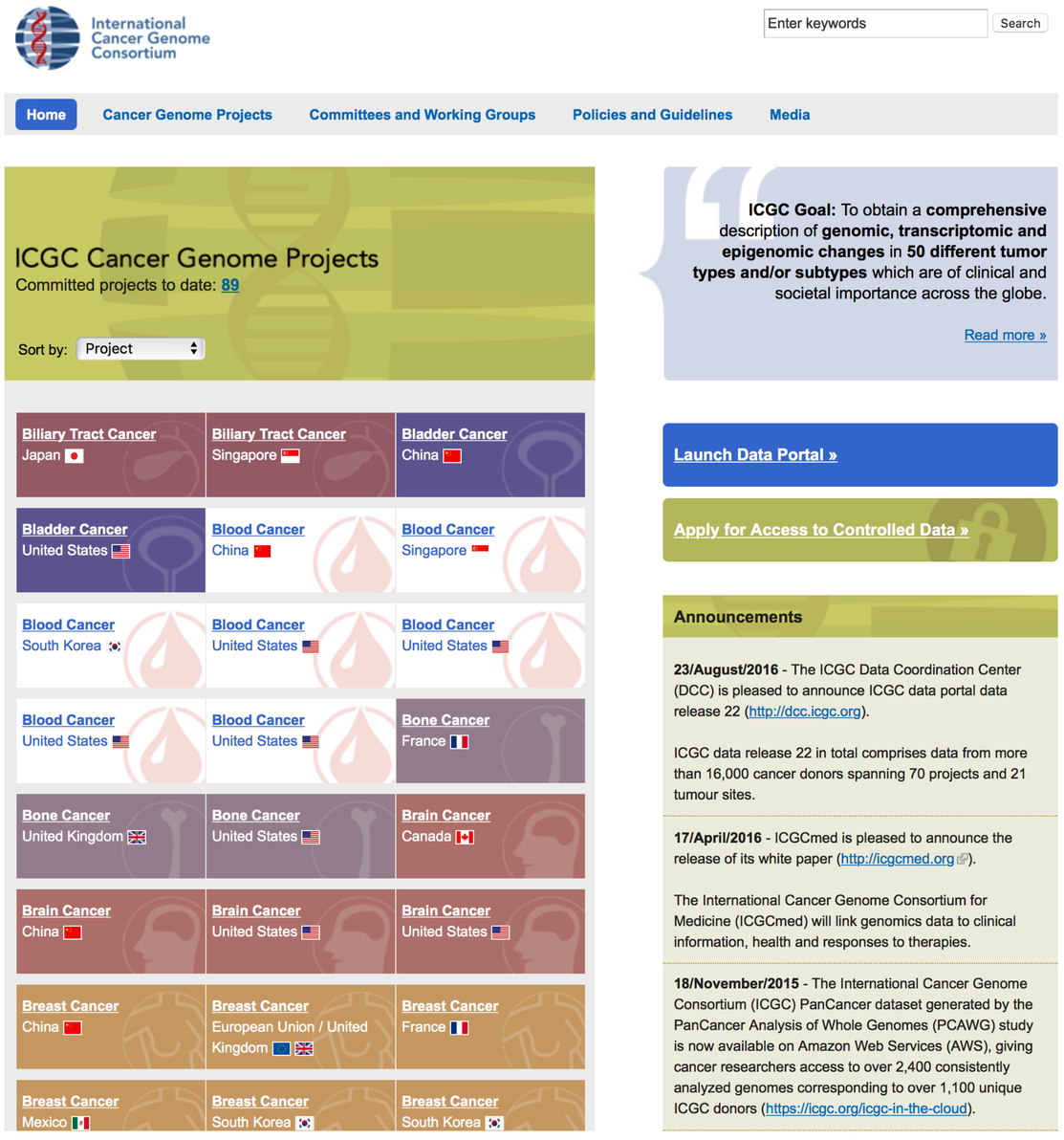






	chromosomal CGH	genomic arrays	"NGS" genome sequencing
1st application report	**1992**	**1997**	**2010**
source	DNA (paraffin, micro-dissected ...)	DNA (paraffin, micro-dissected ...)	DNA (paraffin, micro-dissected ...)
main source problems	mixed/degraded source tissue	mixed/degraded source tissue	mixed/degraded source tissue
resolution	chromosomal bands = few megabases	mostly in the 100kb range, but tiling possible	single bases
target identification	surrogate (position)	"semidirect" (segmentation spanning probes)	direct quantitative and qualitative
structural	no	depending on type	yes
available data	>24,000 cases (57%) through **Progenetix**	raw data repositories (GEO, EMBL, SMD), **arrayMap**	limited (few entities, study consortia…); variant call data in dbgap, clinvar …
predominant data format	ISCN = static	raw => depends on bioinformatics	mostly selected variant calls

RESOURCES FOR CANCER GENOMICS



RESOURCES FOR CANCER GENOMICS

COSMIC
Catalogue of somatic mutations in cancer

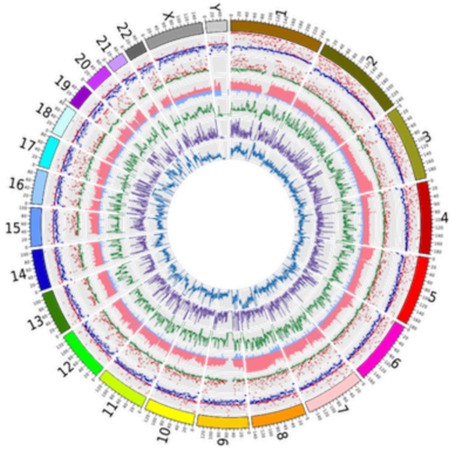
Home ▼ | Resources ▼ | Curation ▼ | Tools ▼ | Data ▼ | News ▼ | Help ▼ | About ▼ | Search COSMIC... | Login ▼

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.

Bioinformatics 2001

progenetix | arrayMap

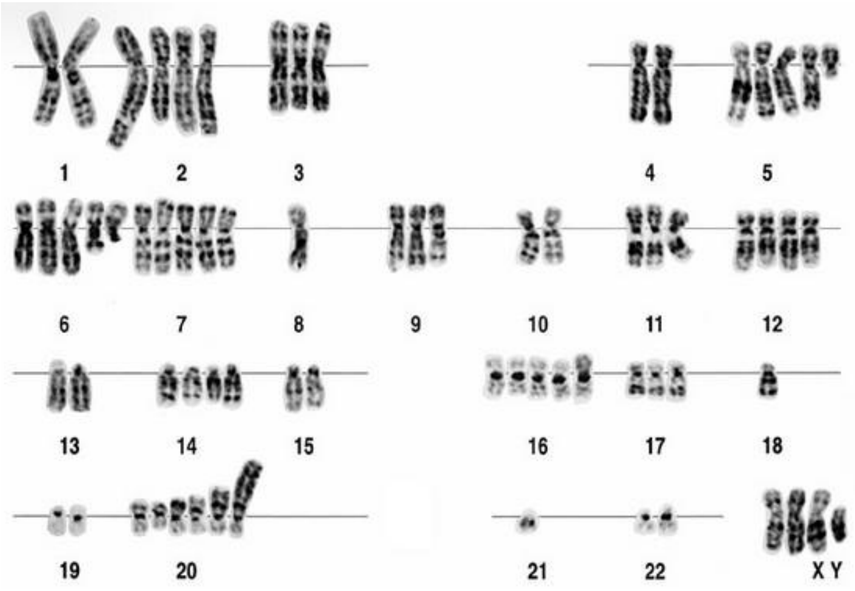
PLOS ONE 2012

	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 ...)

Genomic changes at the DNA level are hallmarks of cancer

We inherited 23 paternal and 23 maternal chromosomes, mostly identical.

Normal karyotype



Tumor karyotype

Our goal: identify CN changes to improve characterization, classification, and treatment of cancers

Chromosomal events can be the basis for regional copy number imbalances

Raymond R. Stallings. Are chromosomal imbalances important in cancer?
Trends Genet (2007) vol. 23 (6) pp. 278-83

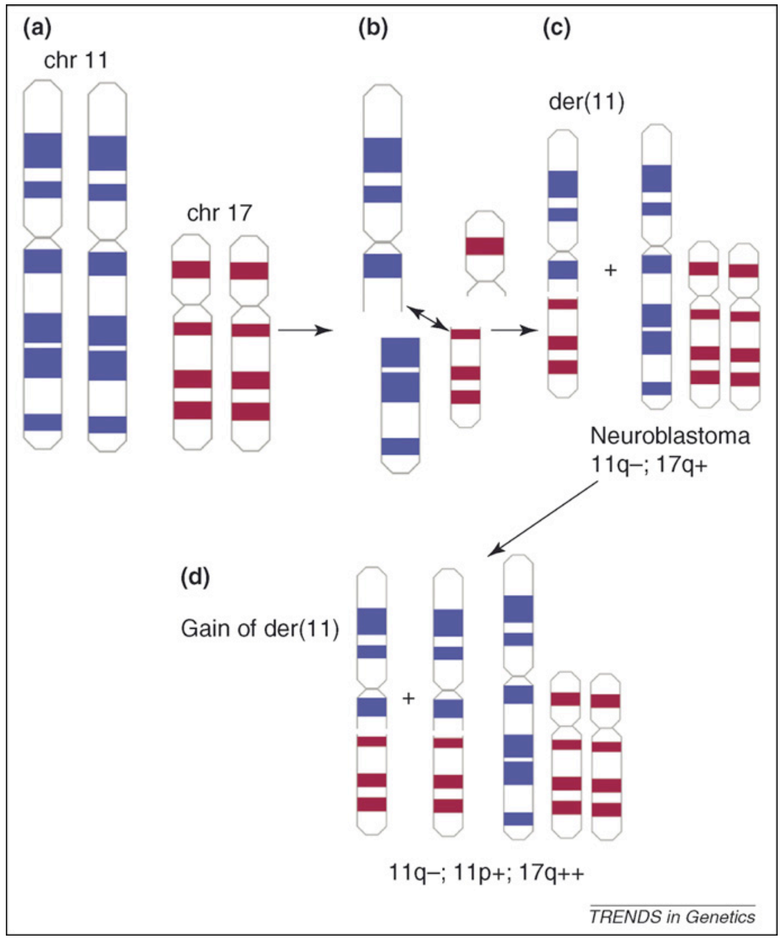
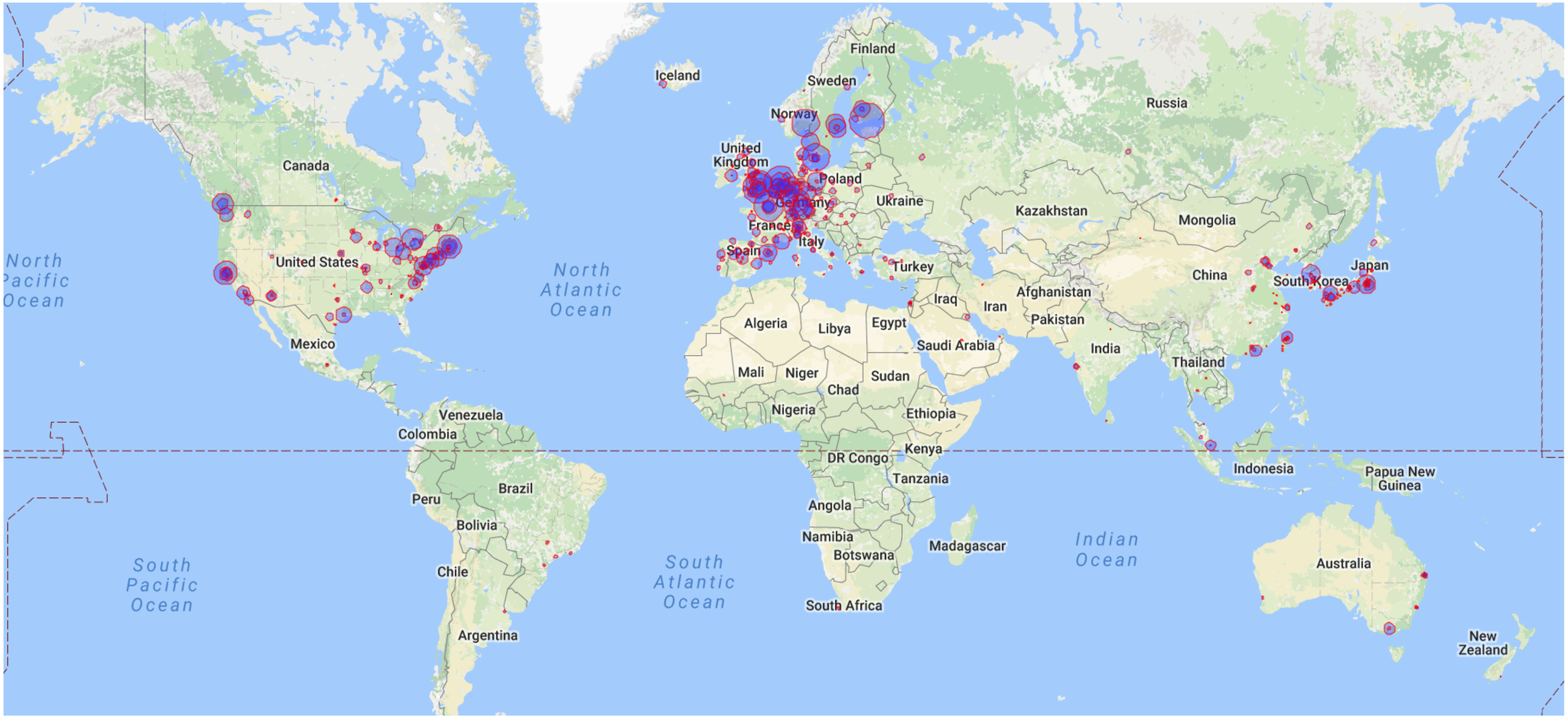


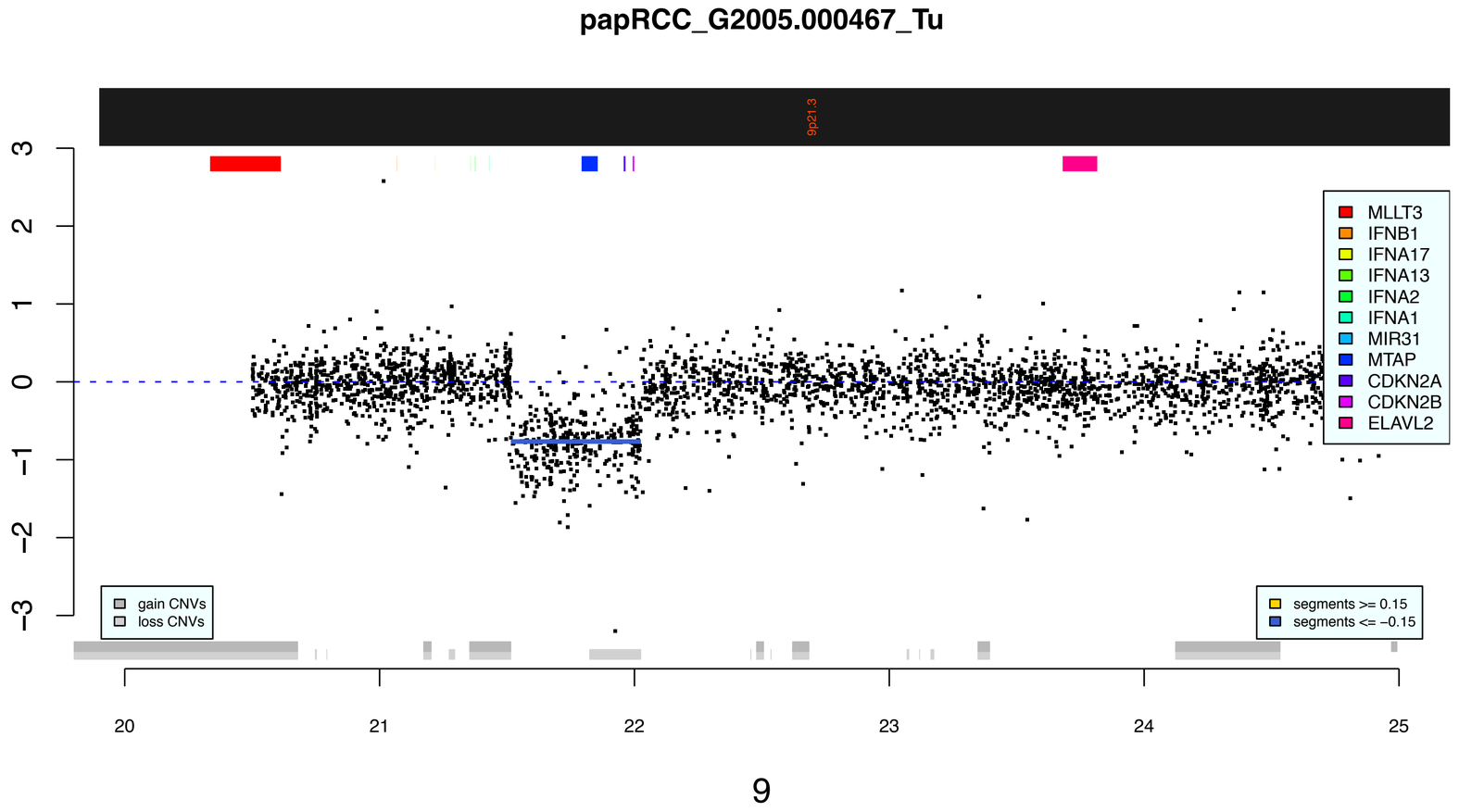
Figure 1. The recurrent unbalanced t(11;17), which generates loss of 11q and gain of 17q in neuroblastoma. Normal cells have two copies of chromosomes 11 and 17 **(a)**. In neuroblastoma cells, breaks occur in bands 11q14 and 17q11.2, followed by the inappropriate rejoining of 11q with 17q **(b)**. The chromosomes segregate so that der(17) is lost, whereas der(11), along with one normal chromosome 11 and two normal chromosomes 17, is retained **(c)**. The resulting cell is $11q^{-}$ (one copy) and $17q^{+}$ (three copies). During tumor progression, some tumors gain an extra copy of the der(11) chromosome **(d)**, becoming $11p^{+}$ (three copies) and $17q^{++}$ (four copies). Gain of additional copies of 17q is probably providing a selective advantage, given that gain of 17q occurs by other cytogenetic mechanisms and is associated with a poor clinical outcome. A gain of 11p might have no selective advantage, occurring only as a consequence of selection for additional gain of 17q. It seems likely that many recurrent chromosomal imbalances in cancer could occur as a consequence of a common chromosomal mechanism affecting other genomic regions that are the genuine selective targets.

RESOURCES FOR CANCER GENOMICS



- The map displays the geographic distribution (by corresponding author) of the 95073 genomic array, 36747 chromosomal CGH and 5012 whole genome/exome based cancer genome datasets. The numbers are derived from the 2942 publications registered in the Progenetix database.

CDKN2A deletion in a case of papillary RCC



- 500kb deletion detected by high resolution genomic array (Affymetrix SNP6)

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 ...

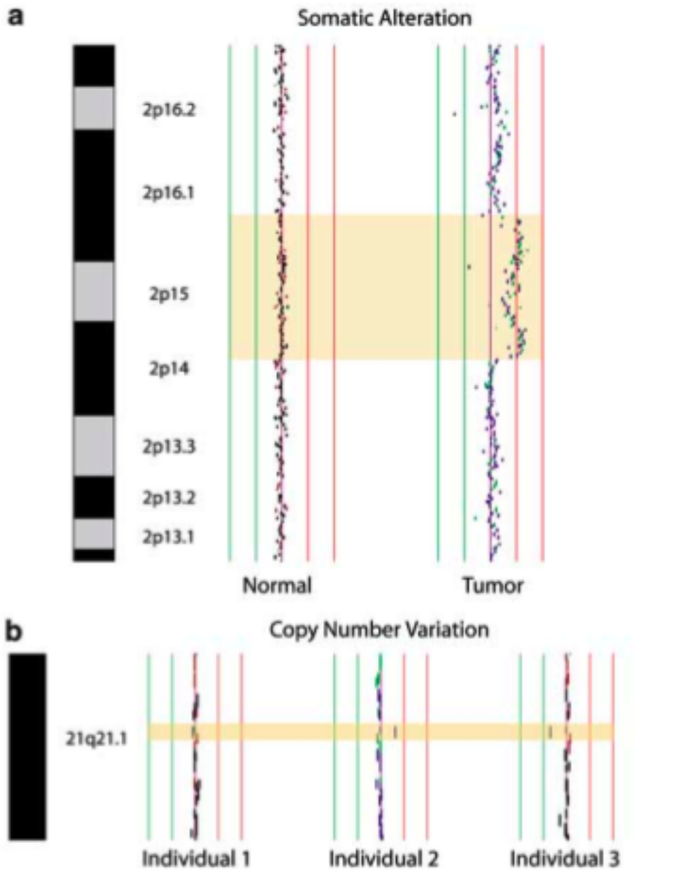


Figure 3 Somatic alterations and copy number variations. (**a**) Example of a segmental duplication observed at chromosome arm 2p present in the cancer cells but absent in the normal cells from the same individual. Each black dot represents a single BAC clone spotted on the array. The purple line represents 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)

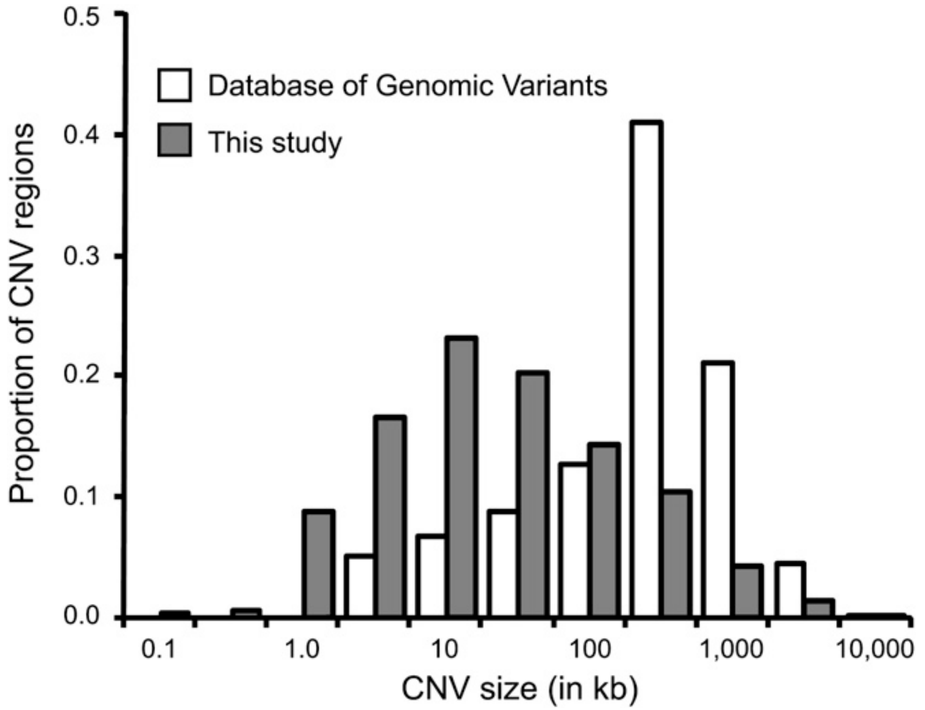
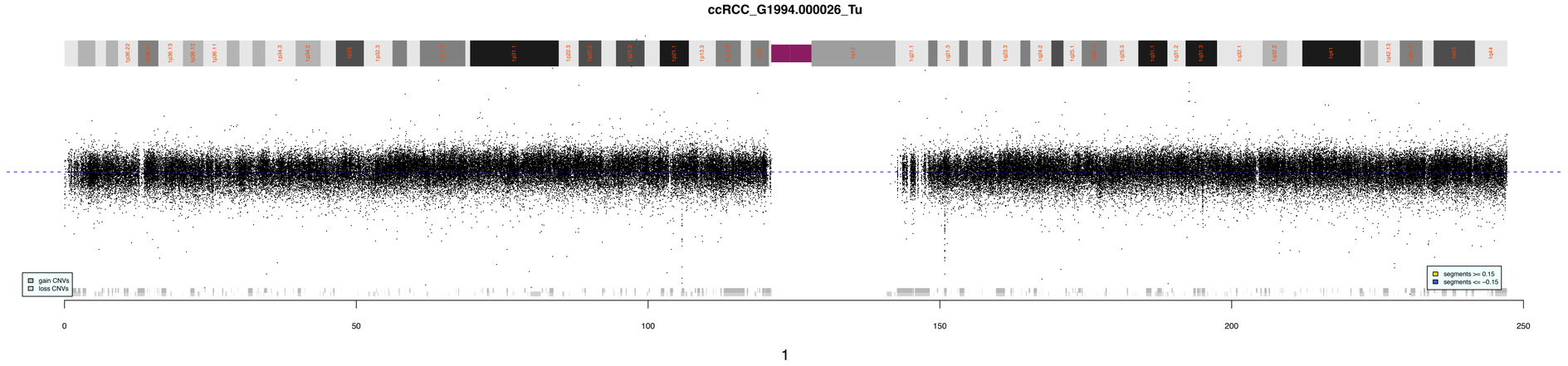
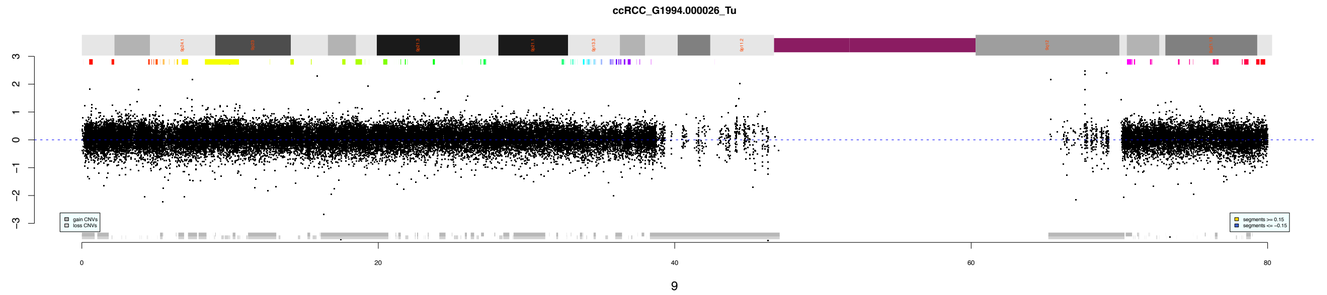


Figure 1. Size Distribution of CNVs from the Database of Genomic Variants, with Corresponding CNVs from This Study
We identified CNVs in at least one individual for 1153 of 2191 putative CNV regions annotated in the Database of Genomic Variants (DGV) as of 30 November 2006. Size distributions for these 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%).

Challenges in CGH data processing: repetitive elements - Alu, LINEs, SINEs



Significant genomic regions have a high content of repetitive DNA sequences, which make regional copy number estimations impossible and may lead to segmentation errors in border areas.



Copy number status data: Segmentation (CBS)

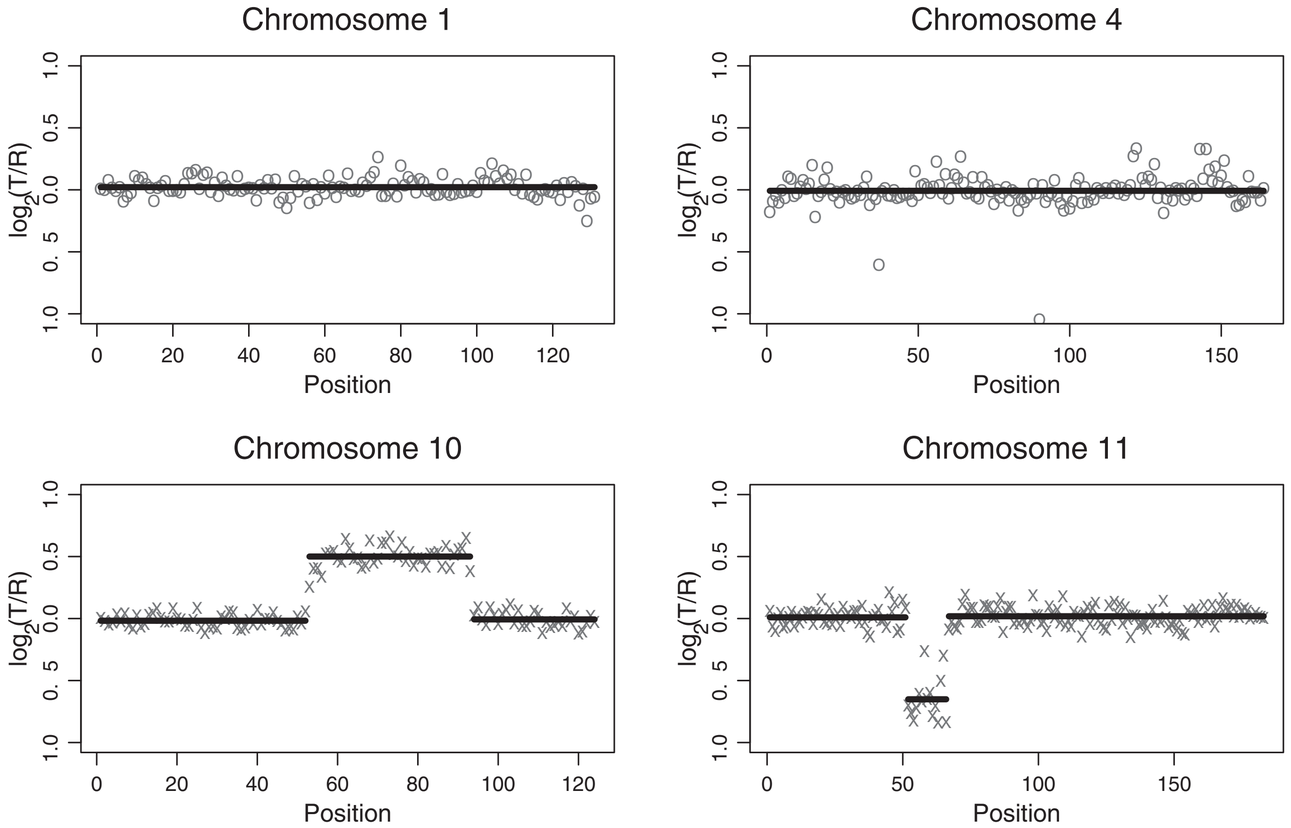
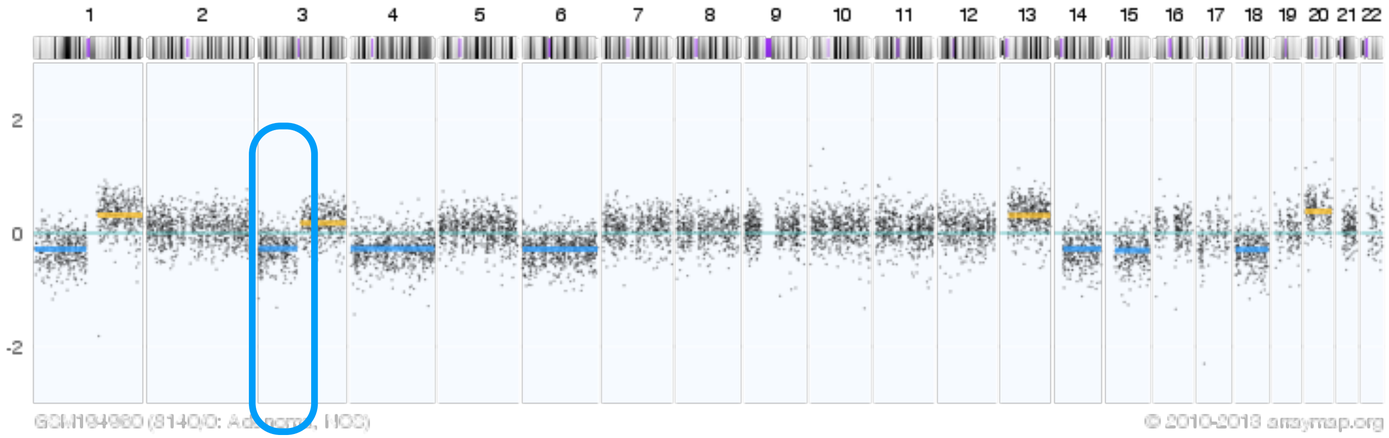


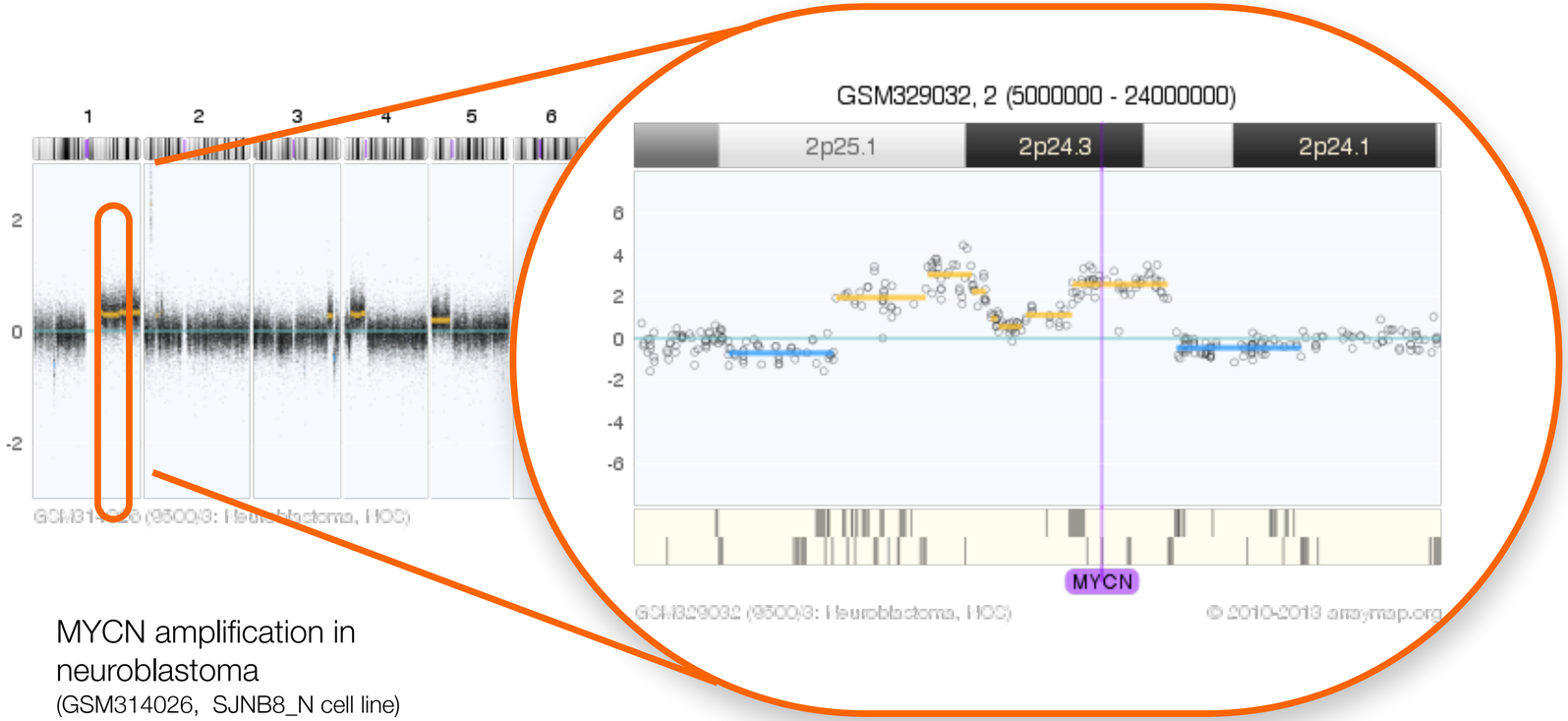
Fig. 1. A CBS analysis of the fibroblast cell line GM05296, which has known alterations only on chromosomes 10 and 11. The points are normalized log ratios, and the lines are the mean values among points in segments obtained by CBS.

Olshen et al. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics (2004) vol. 5 (4) pp. 557-72

Genomic arrays: Many probes + bioinformatics determine copy number aberrations



Gain of chromosome arm 3q in colorectal carcinoma

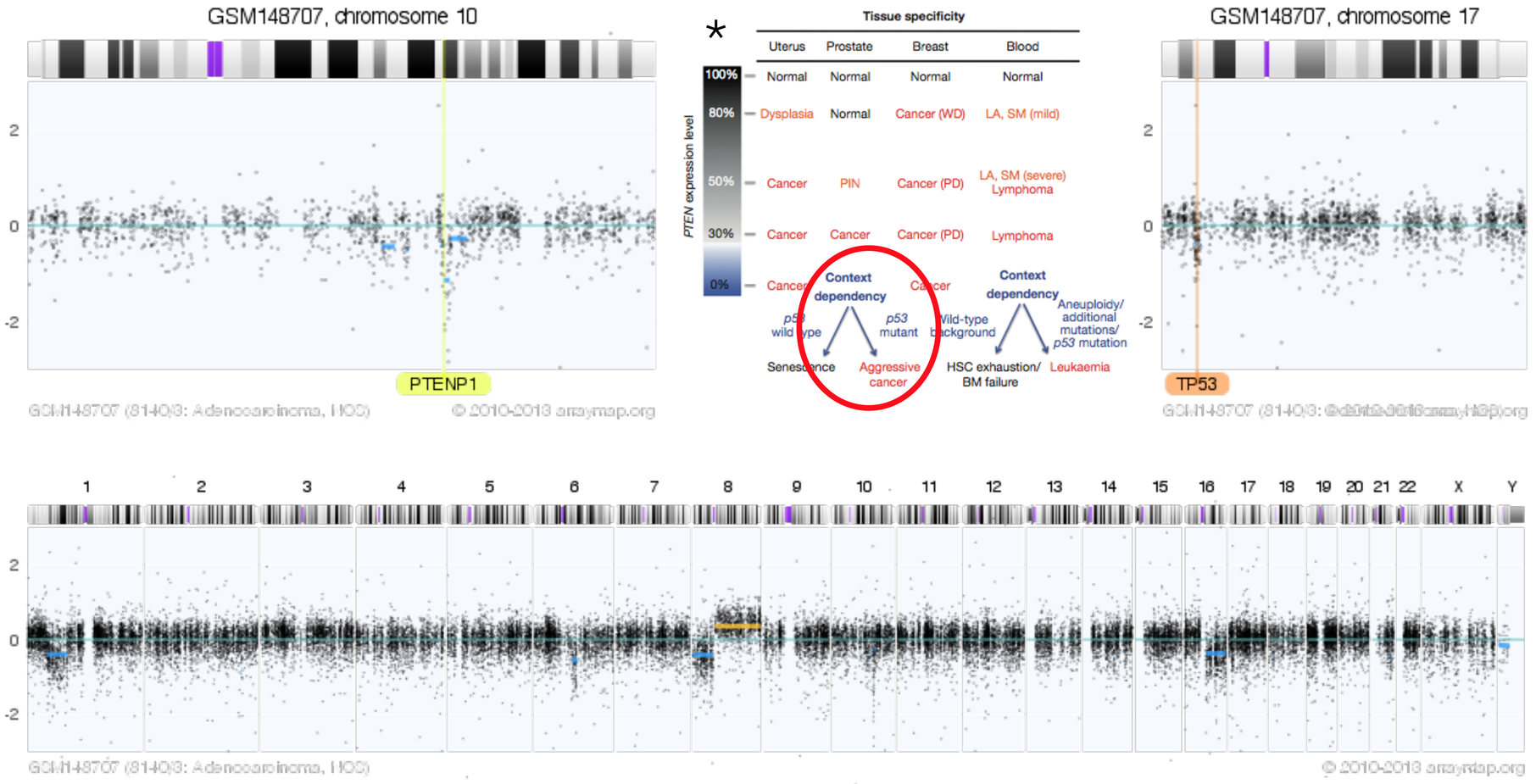


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

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

arrayMap

Gene dosage phenomena beyond simple on/off effects

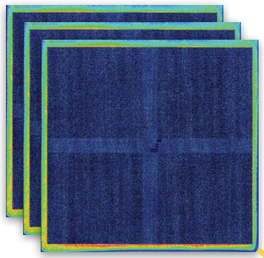


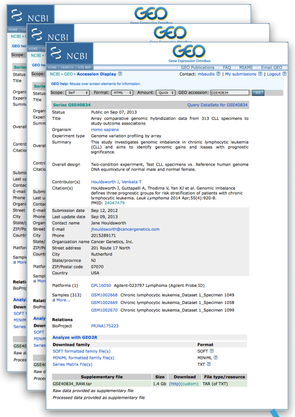
Combined heterozygous deletions involving *PTEN* and TP53 loci in a case of prostate adenocarcinoma
(GSM148707, PMID 17875689, Lapointe *et al*., CancRes 2007)

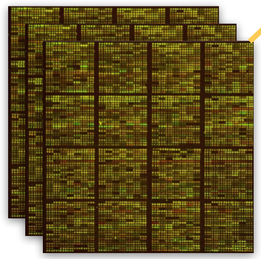
* A. H. Berger, A. G. Knudson, and P. P. Pandolfi, “A continuum model for tumour suppression,” *Nature*, vol. 476, no. 7359, pp. 163–169, Aug. 2011.

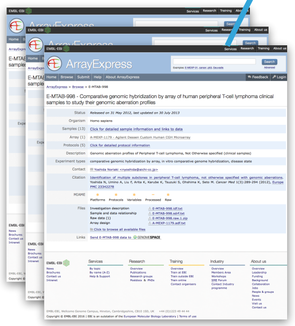
arrayMap

ARRAYMAP DATA PIPELINE

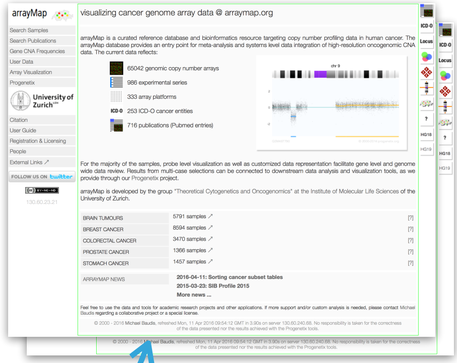


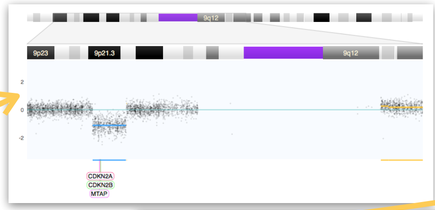


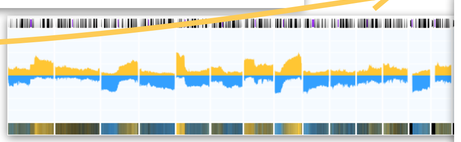


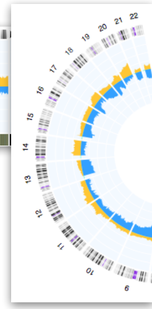


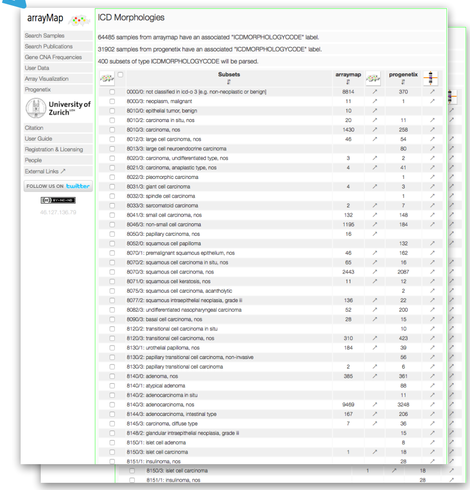
Automagical Processing Engine™





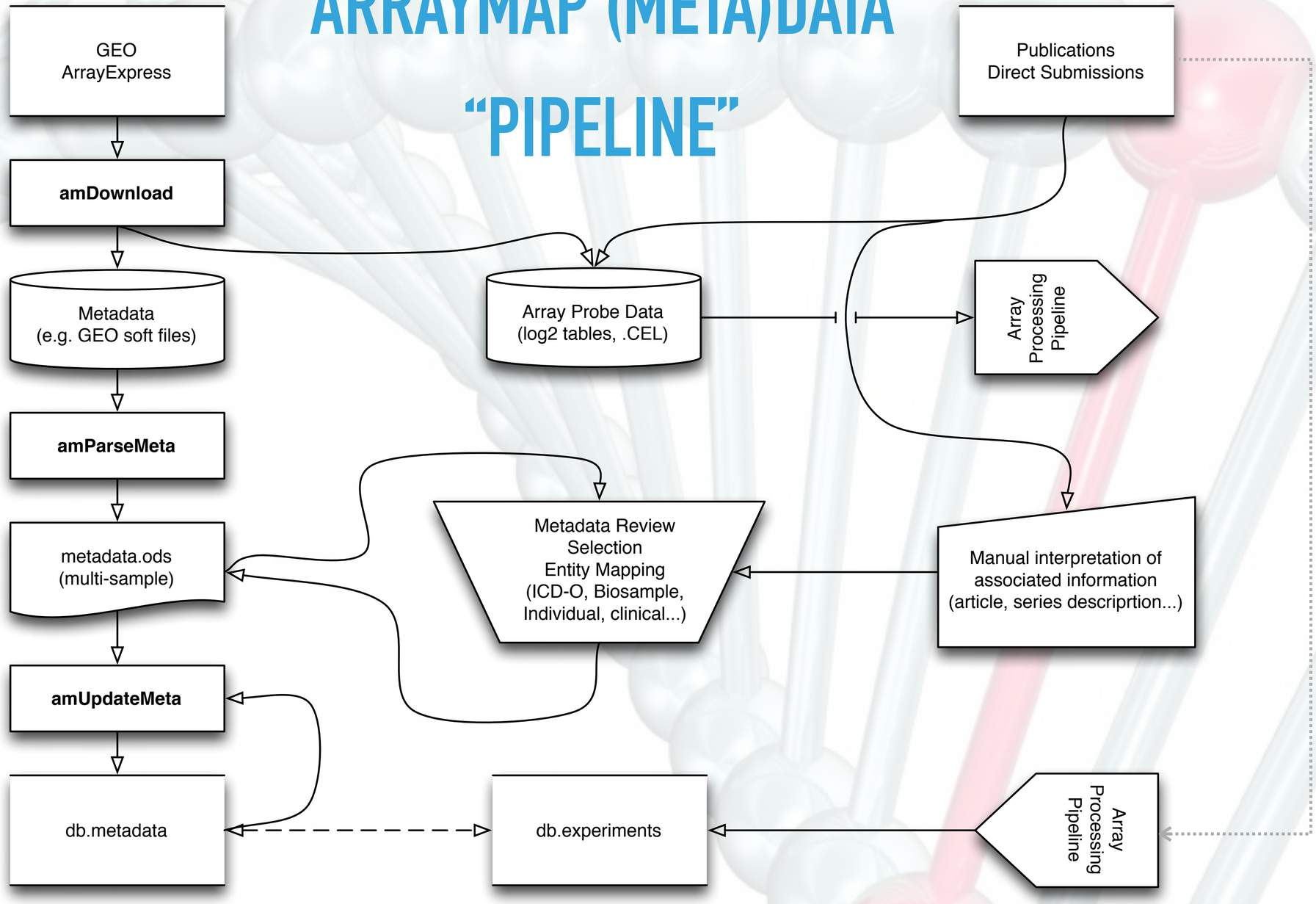




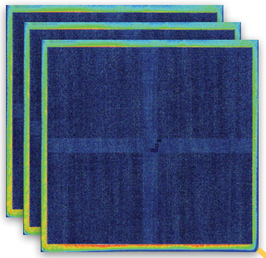


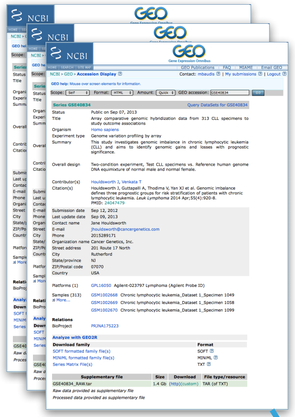
arrayMap

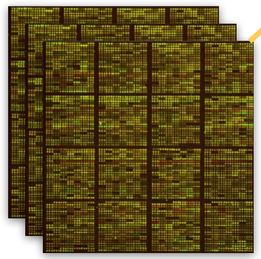
ARRAYMAP (META)DATA "PIPELINE"

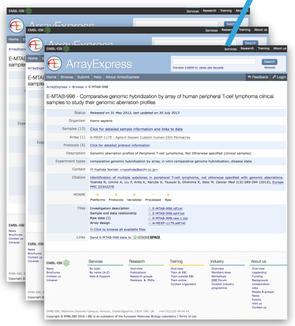


ARRAYMAP DATA PIPELINE





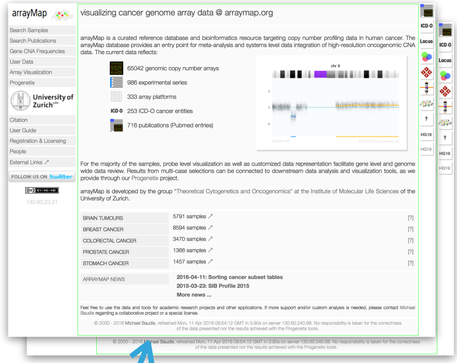


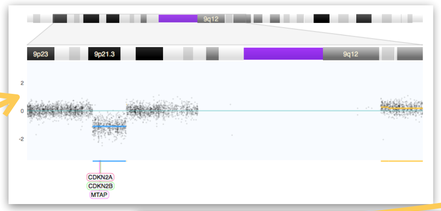


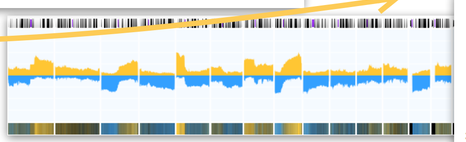
Automagical Processing Engine™

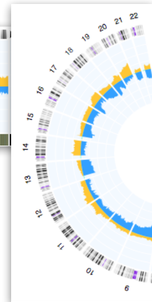
BIOCURATION

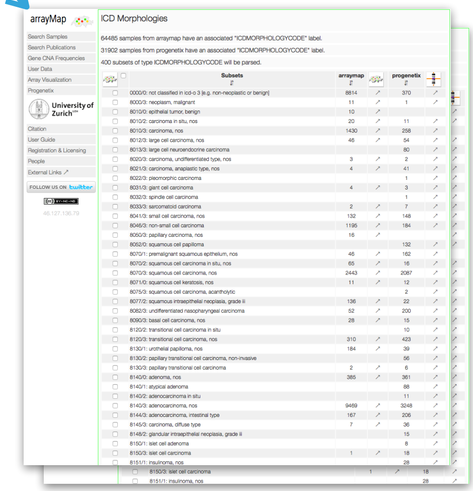
BIOINFORMATICS







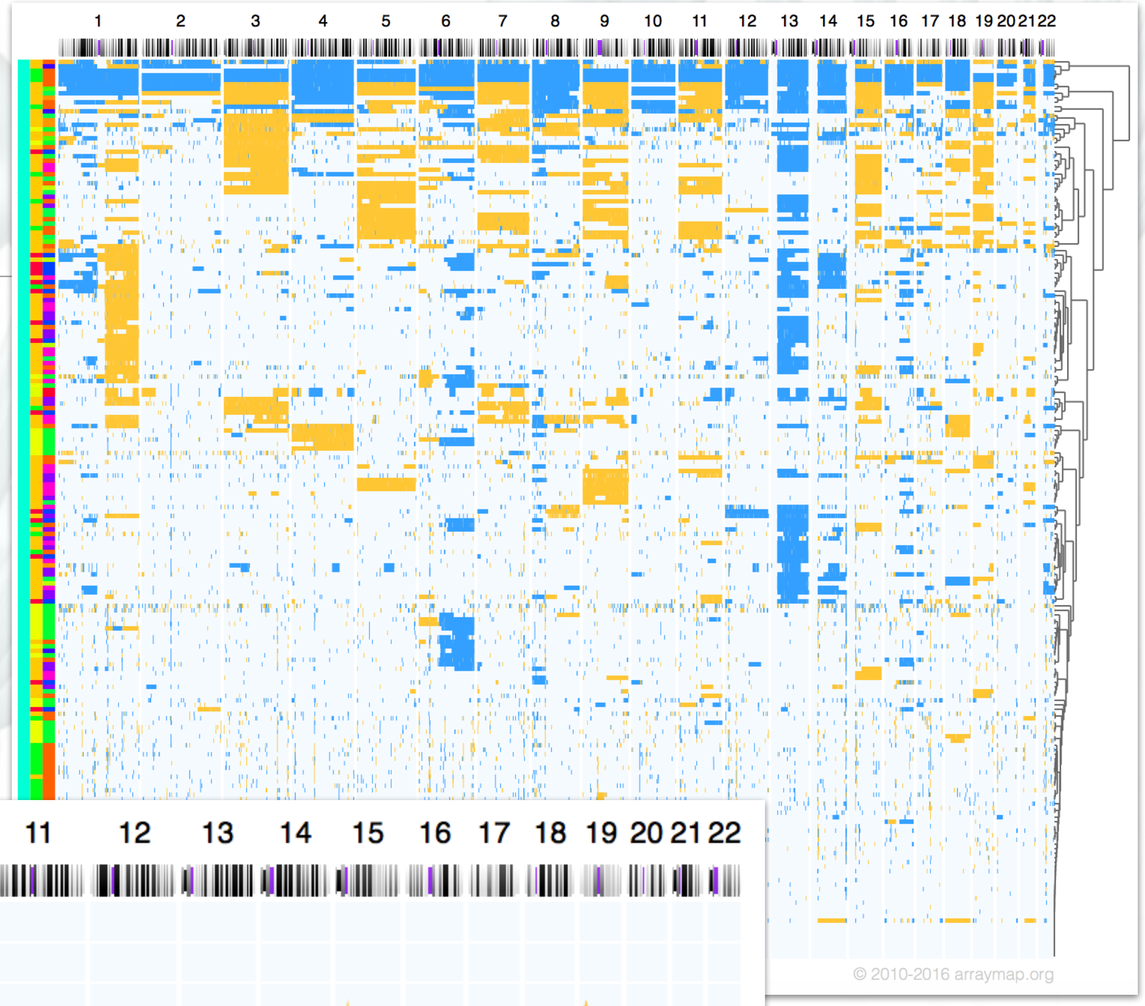


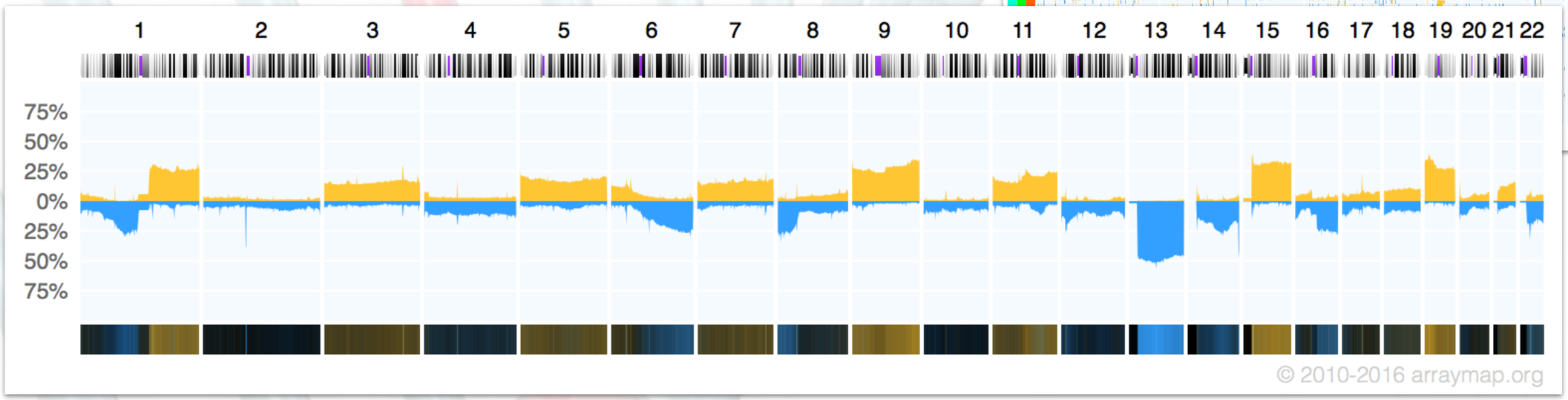


arrayMap

Genomic Heterogeneity in Cancer: Malignant Myeloma

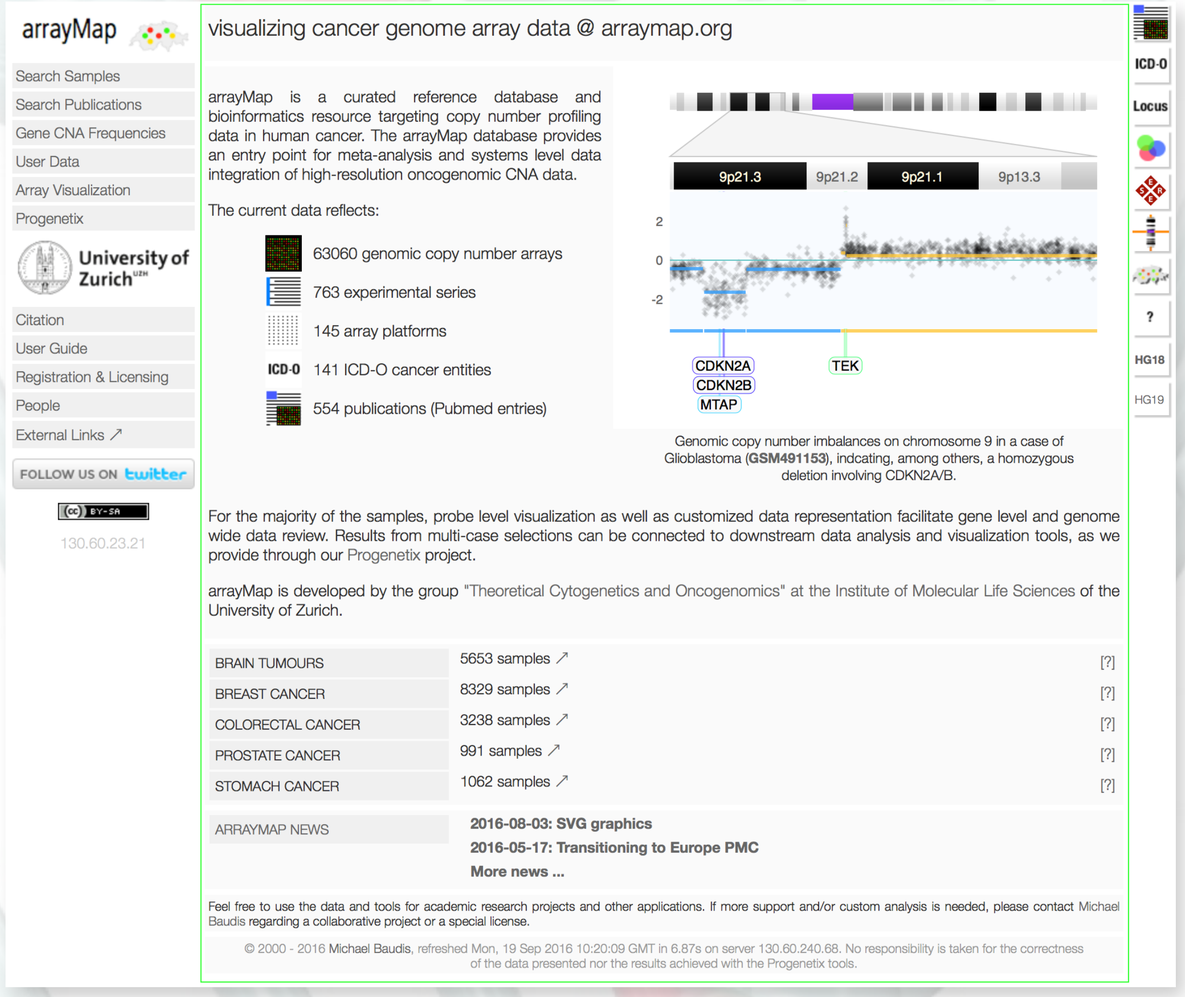
- variety of genomic copy number changes
- prevalence of whole chromosome CNA
- -13 with high frequency





arrayMap

The arrayMap Cancer Genome Resource

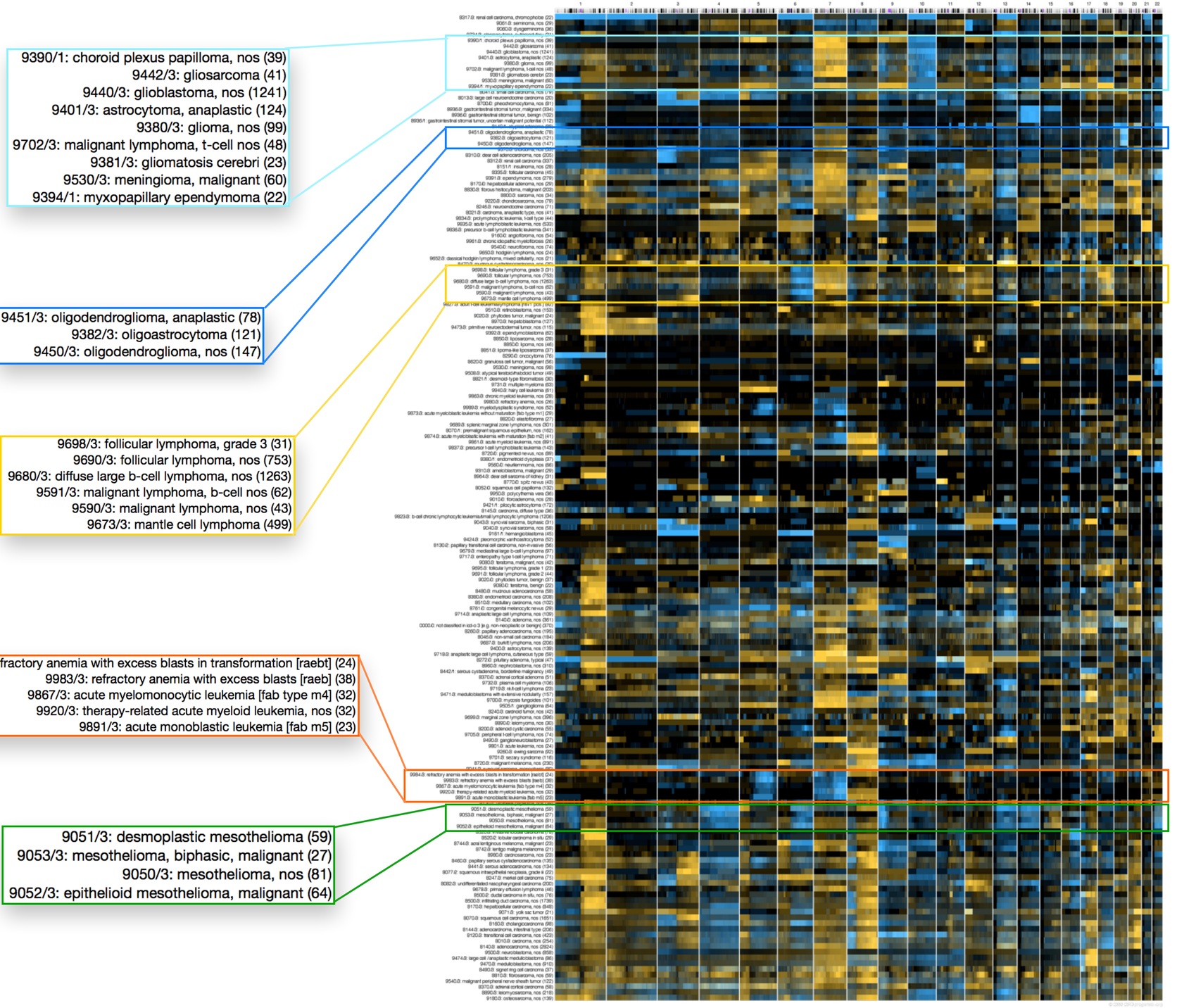


UID	SERIESID	PMID	ICDMORPHOLOGYCODE	ICDTOPOGRAPHYCODE
GSM1000061	GSE36942	23457519	8070/3	C10
GSM1000062	GSE36942	23457519	8070/3	C10
GSM1001316	GSE40777	23571474	8070/3	C53
GSM1001317	GSE40777	23571474	8010/3	C34
GSM1001318	GSE40777	23571474	8070/3	C09
GSM1001319	GSE40777	23571474	8010/3	C34
GSM1002668	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

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles



arrayMap

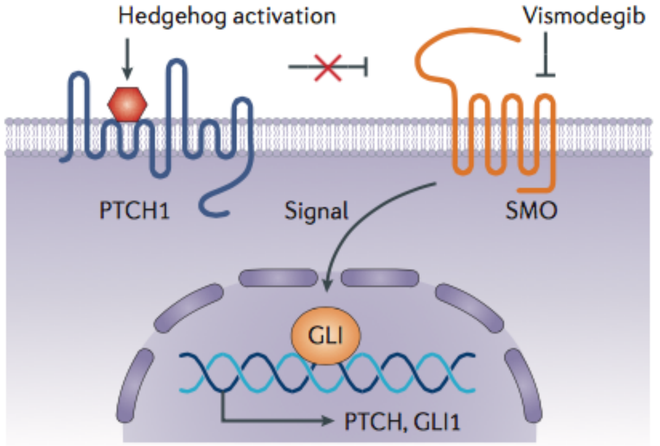
Rare focal events for “personalized” therapeutic options?

Example: PTCH1 deletions in malignant melanomas

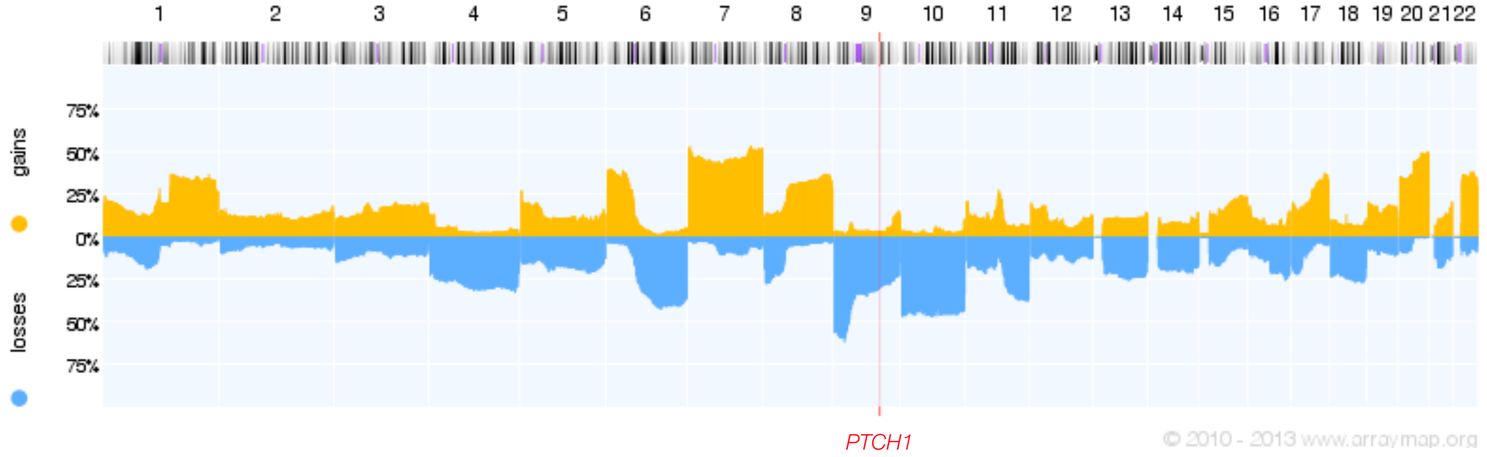
The Sonic Hedgehog (SHH) pathway has become a “druggable” target in the therapy of syndromic and/or advanced basaliomas (e.g. in Gorlin syndrome).

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 ...

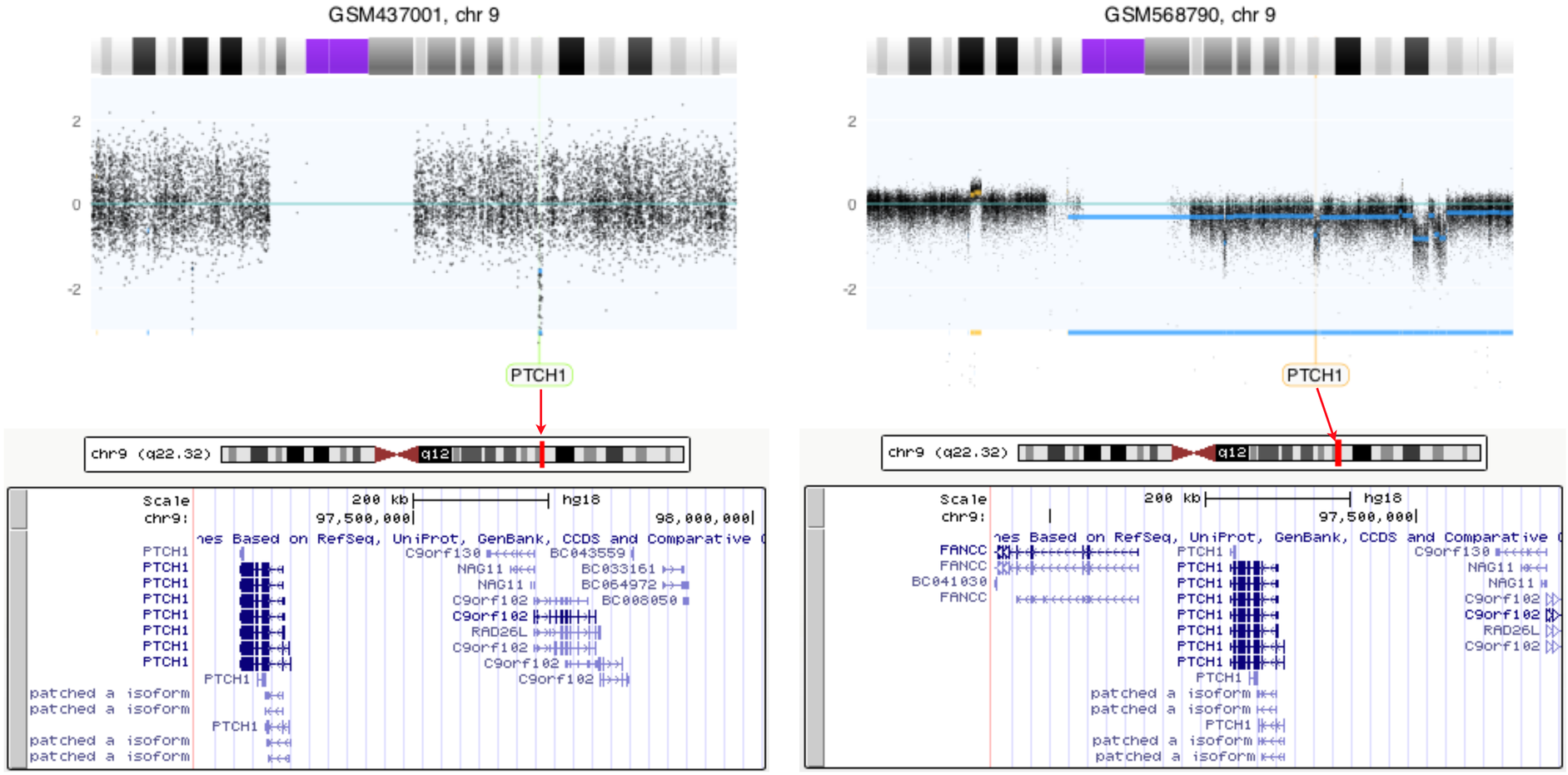


Dlugosz, A., Agrawal, S., & Kirkpatrick, P. (2012, June). Nature Reviews Drug Discovery, pp. 437–438



=> no “hot spot” (but 30% deletions)

Examples for malignant melanomas with focal / homozygous PTCH1 deletions



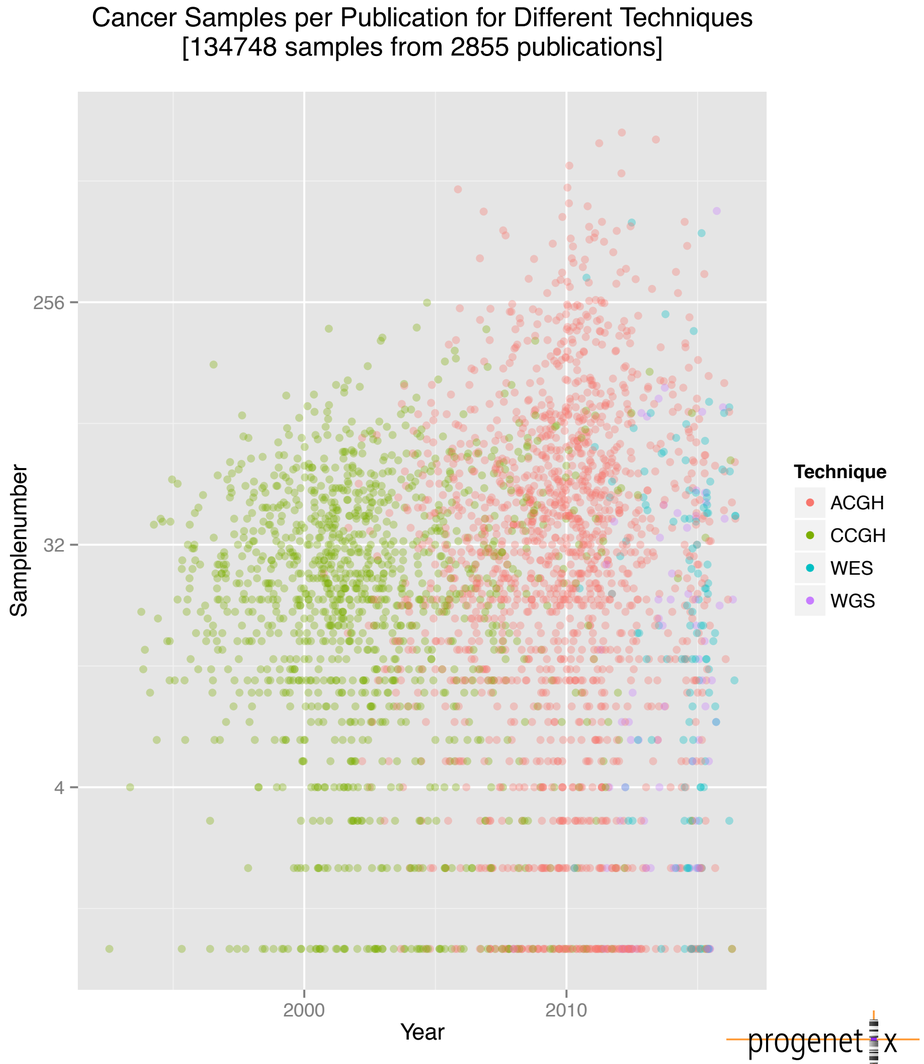
Systematic analysis of focal hits on “cancer genes” from large data collections can guide diagnostic and therapeutic procedures for personalized treatment options.

MOLECULAR CYTOGENETICS & SEQUENCING STUDIES FOR WHOLE GENOME PROFILING

The map displays the geographic distribution (by corresponding author) of the 91584 genomic array, 36747 chromosomal CGH and 4704 whole genome/exome based cancer genome datasets. The numbers are derived from the 2870 publications registered in the Progenetix database.

SHIFT TO SEQUENCING BASED TECHNIQUES LEADS TO SEVERELY LIMITED DATA ACCESSIBILITY

- Whole-genome screening is an important tool to understand the “mutational landscapes” in human malignancies
- “molecular” technologies (excluding traditional cytogenetics) have been, successively, chromosomal CGH, genomic arrays, Whole Exome/Genome Sequencing



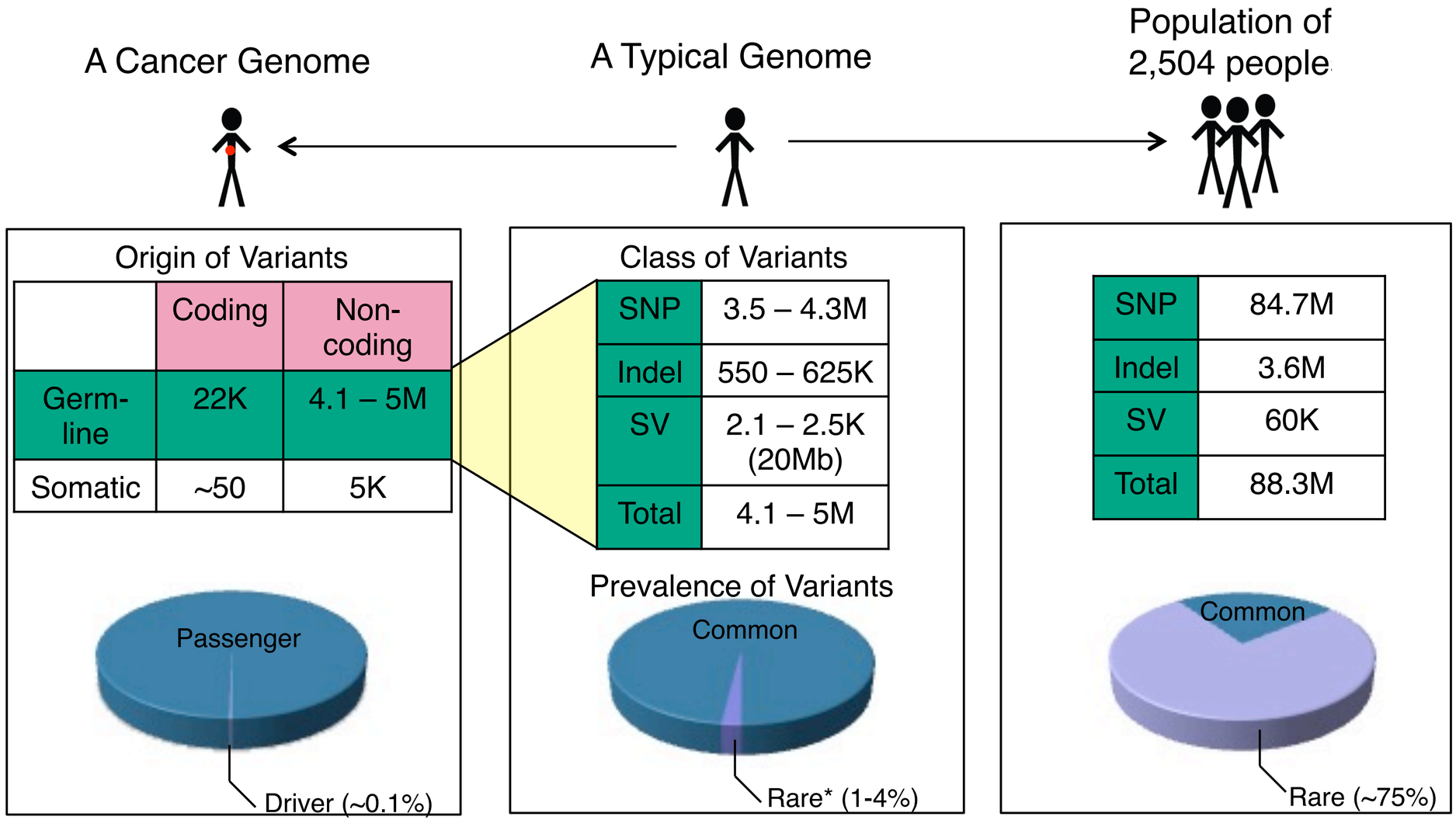
progenetix

GA4GH - Global Alliance for Genomics and Health

History, Structure, Projects

Global Alliance
for Genomics & Health

Human Genetic Variation



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

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

Genome data held in silos, unshared, not standardized for exchange

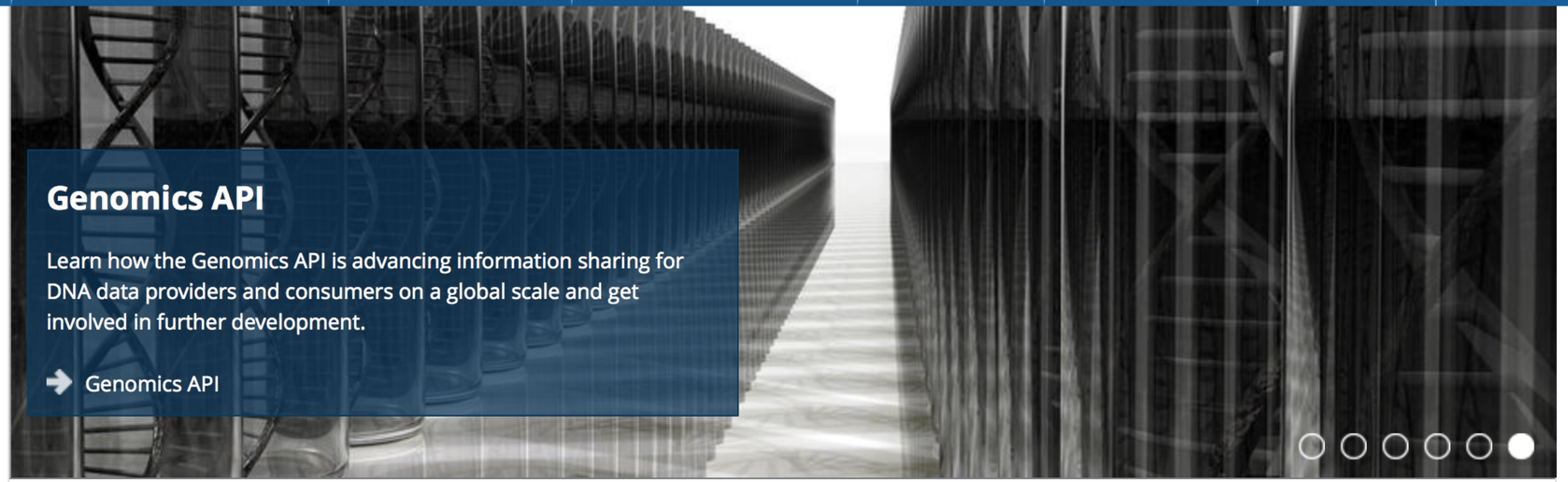
Global Alliance
for Genomics & Health

David Haussler

Global Alliance for Genomics & Health

Become a Member

ABOUT GLOBAL ALLIANCE | WORKING GROUPS | PRODUCTS & PROJECTS | MEMBERSHIP | NEWS & EVENTS | CONTACT US



Our Work

The diverse members of the Global Alliance are working together to create interoperable approaches to catalyze projects that will help unlock the great potential of genomic data. Our four Working Groups advance Initiatives that develop key Work Products.

Clinical »

Aims to enable compatible, readily accessible, and scalable approaches for sharing clinical data and linking it with genomic data.

Data »

Concentrates on data representation, storage, and analysis of genomic data to develop approaches that facilitate interoperability.

Regulatory and Ethics »

Focuses on ethics and the legal and social implications of the Global Alliance, including harmonizing policies and standards.

Security »

Leads the thinking on the technology aspects of data security, user access control, audit functions, and developing or adopting data security standards.

GA4GH HISTORY & MILESTONES

Global Alliance for Genomics & Health

- 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


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

GA4GH MEMBERSHIP

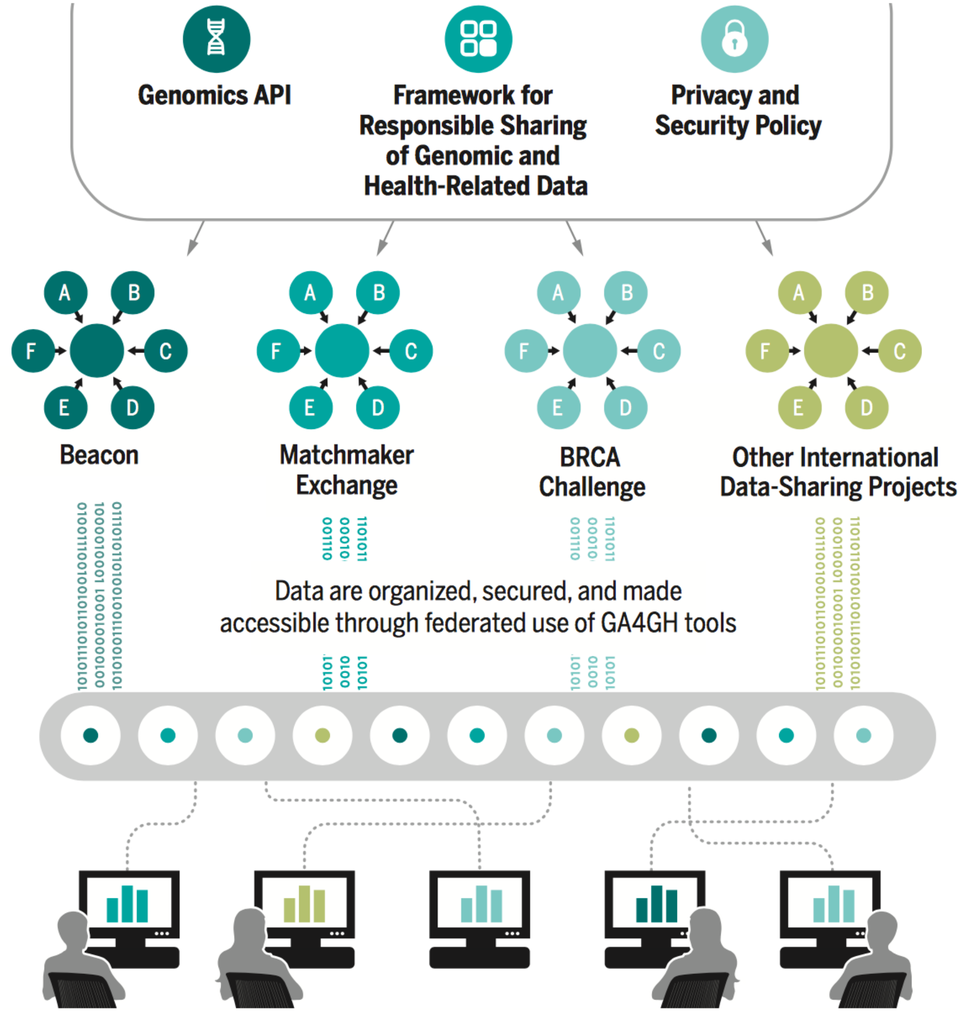
- host institutions
  - OICR, Broad Institute, Sanger Institute
  - Peter Goodhand, OICR, exec. director
- funding
  - CanSHARE, NIH, Wellcome Trust
- founding members
  - 229 partner organizations based in 30 countries as of October 3, 2014
  - SIB and UZH represented Switzerland
- membership status
  - 433 organizational members as of September 2016
  - open to individual registrations & participation
  - no financial commitment necessary; however, calls for support for events

Global Alliance for Genomics & Health

GA4GH API promotes sharing

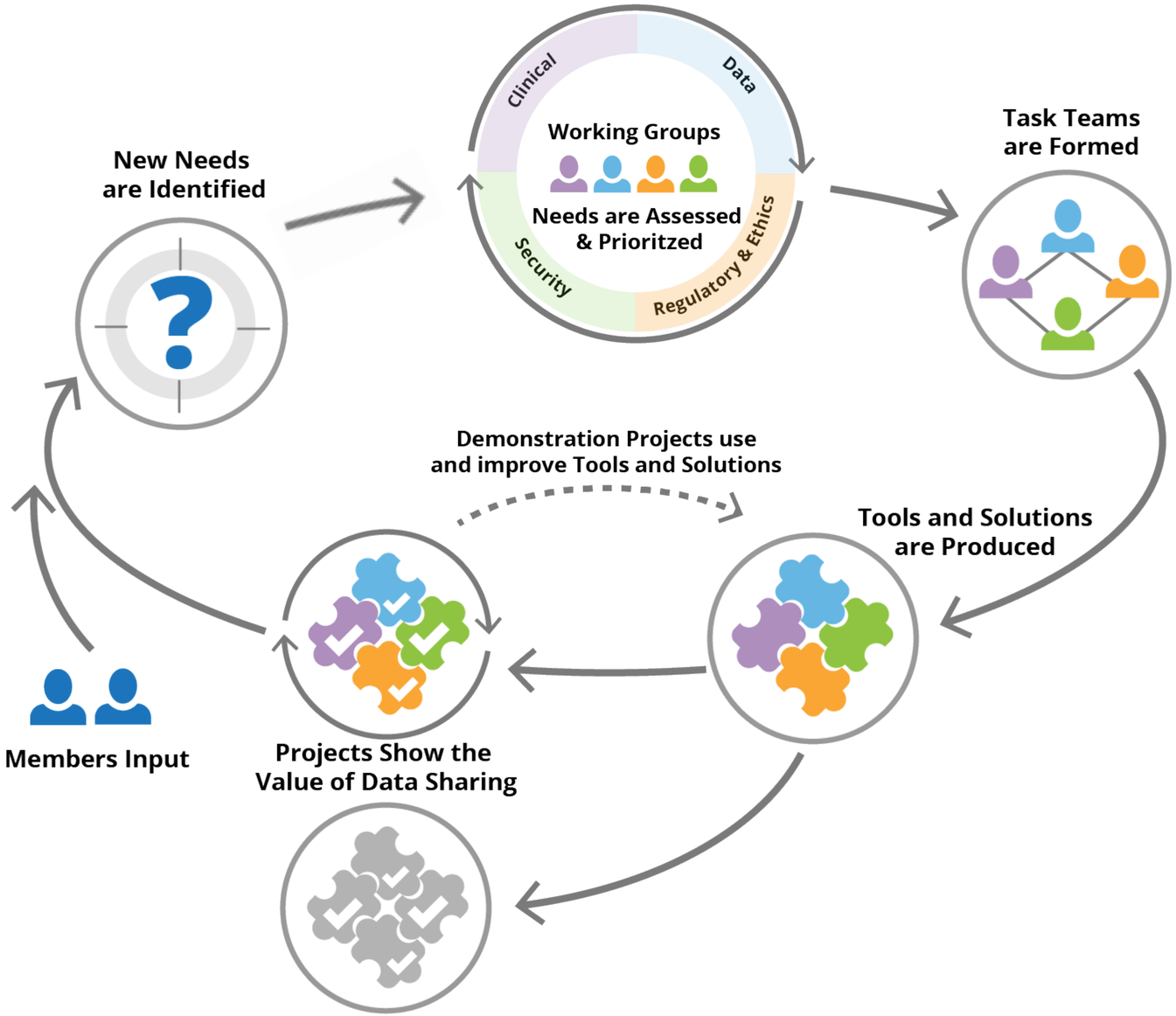
The mission of the Global Alliance for Genomics and Health is to accelerate progress in human health by helping to establish a **common framework** of harmonized approaches to enable **effective and responsible** sharing of **genomic and clinical data**, and by catalyzing data sharing projects that drive and demonstrate the value of data sharing.

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

How We Work


Clinical
Data
Working Groups
Needs are Assessed & Prioritzed
Security
Regulatory & Ethics
New Needs are Identified
Task Teams are Formed
Demonstration Projects use and improve Tools and Solutions
Tools and Solutions are Produced
Members Input
Projects Show the Value of Data Sharing

WORK PRODUCT: FRAMEWORK FOR RESPONSIBLE SHARING OF GENOMIC AND HEALTH-RELATED DATA

Foundational Principles for Responsible Sharing of Genomics and Health-Related Data

- Respect Individuals, Families and Communities
- Advance Research and Scientific Knowledge
- Promote Health, Wellbeing and the Fair Distribution of Benefits
- Foster Trust, Integrity and Reciprocity

Foundational Principles for Responsible Sharing of Genomic and Health-Related Data

- Respect Individuals, Families and Communities
- Advance Research and Scientific Knowledge
- Promote Health, Wellbeing and the Fair Distribution of Benefits
- Foster Trust, Integrity and Reciprocity

Global Alliance for Genomics and Health (GA4GH): Proposed Policy Template

Global Alliance for Genomics & Health

Published: September 10, 2014

WORK PRODUCT: CONSENT CODES

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]

1 Centre of Genomics and Policy, Faculty of Medicine, McGill University, Montreal, Quebec, Canada, **2** Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, **3** Centre for Genomic Regulation (CRG), Barcelona, Spain, **4** Universitat Pompeu Fabra (UPF), Barcelona, Spain, **5** Office of Science Policy, Office of the Director, National Institutes of Health, Bethesda, Maryland, United States of America, **6** Department of Genetics, University of Leicester, Leicester, United Kingdom, **7** European Molecular Biology Laboratory—European Bioinformatics Institute (EMBL—EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **8** Human Genetics Department, Leiden University Medical Center, Leiden, The Netherlands, **9** Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, Ontario, Canada, **10** Centre for Computational Medicine, Hospital for Sick Children, Toronto, Ontario, Canada, **11** Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, **12** Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom, **13** Department of Pathology, Brigham & Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States of America, **14** Bioreference Laboratories, Inc., Elmwood Park, New Jersey, United States of America, **15** DNAstack, Toronto, Ontario, Canada, **16** National Centre for Biotechnology Information, US National Library of Medicine, Bethesda, Maryland, United States of America

PLOS Genetics | DOI:10.1371/journal.pgen.1005772 January 21, 2016

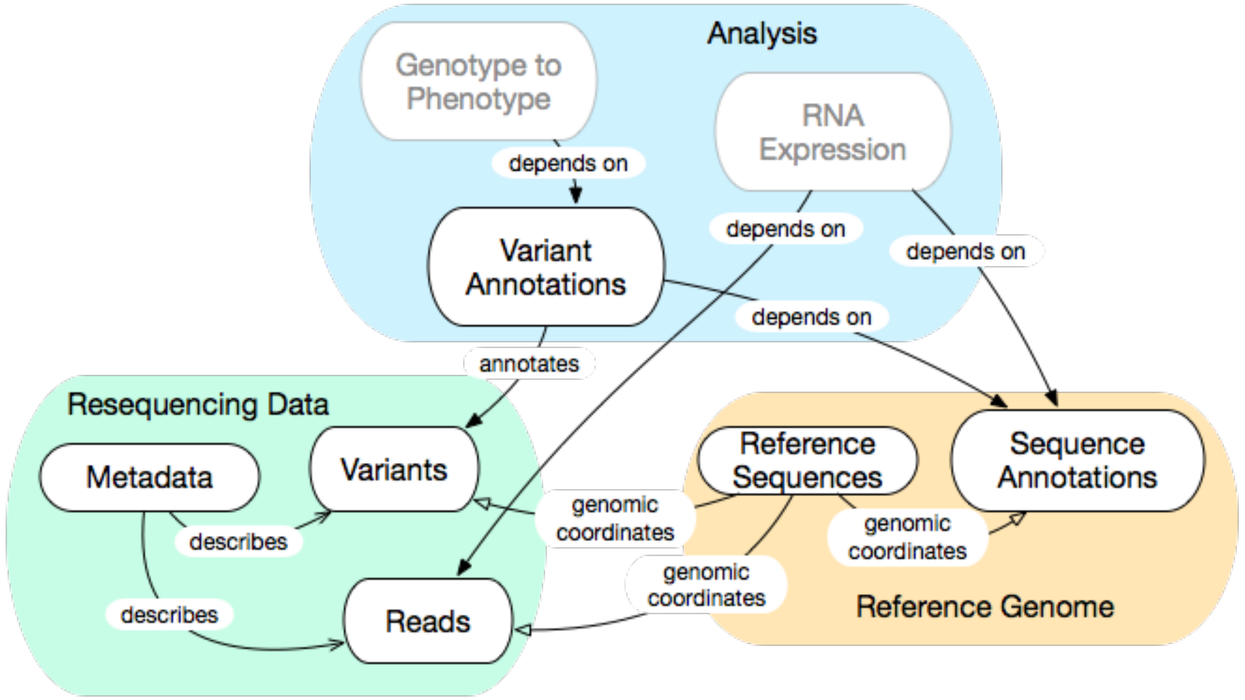
Global Alliance for Genomics & Health

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*

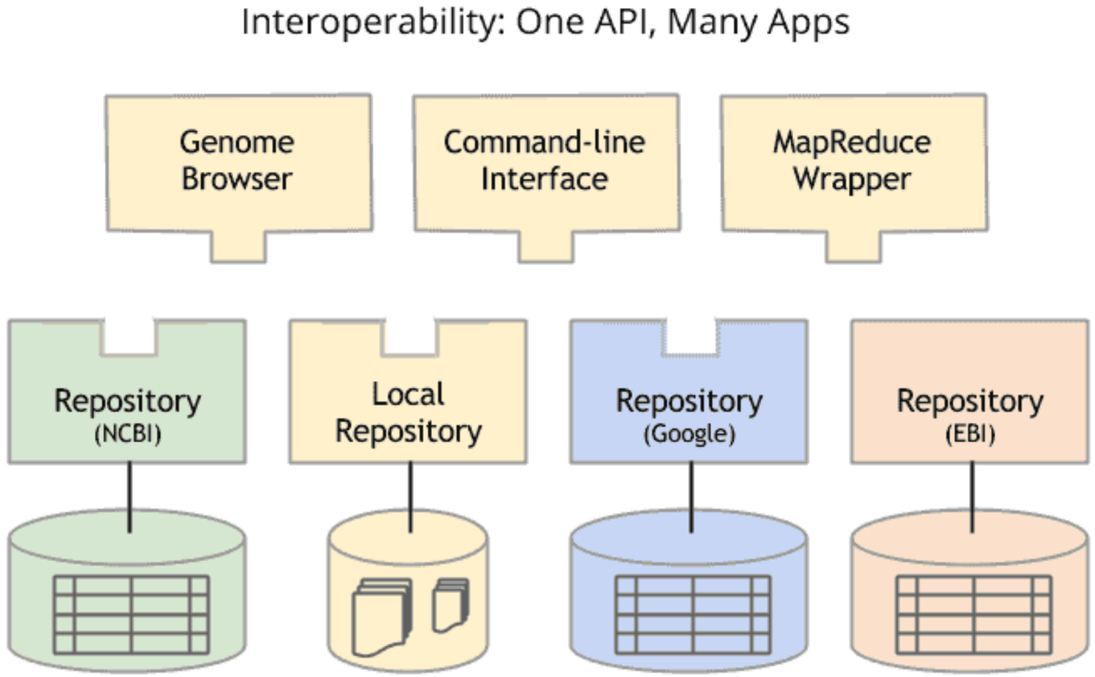
Contact: Dr. Stephanie Dyke (stephanie.dyke@mcgill.ca)

WORK PRODUCT: GENOMICS API



Global Alliance for Genomics and Health

Data Working Group



Genomics API

The Global Alliance for Genomics and Health (GA4GH) Genomics API will allow the interoperable exchange of genomic information across multiple organizations and on multiple platforms. This is a freely available open standard for interoperability, that uses common web protocols to support serving and sharing of data on DNA sequences and genomic variation. The API is implemented as a webservice to create a data source which may be integrated into visualization software, web-based genomics portals or processed as part of genomic analysis pipelines. It overcomes the barriers of incompatible infrastructure between organizations and institutions to enable DNA data providers and consumers to better share genomic data and work together on a global scale, advancing genome research and clinical application.

GA4GH: SOME OPINIONS

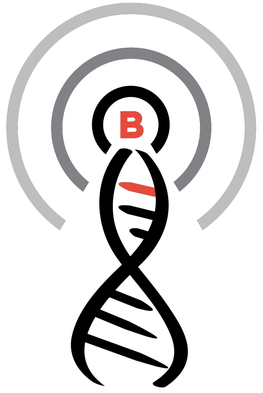
- GA4GH is a very large, dynamic and heterogeneous organization
- activity & impact is particularly high in U.S. | CA | U.K., with some emphasis on locally relevant projects (e.g. EHR integration via FHIR)
- data working group projects and products (file formats, API) should be considered (emerging) standards
- driver projects (esp. Beacon, BRCA challenge) offer test beds & references for data sharing implementations
- concrete implementations (e.g. Beacon instances, API, tiered access) are now driven forward by organizations/initiatives loosely affiliated w/ GA4GH (example: EGA/ELIXIR support)

DRIVING GA4GH METADATA SCHEMA

- arrayMap for GA4GH
  - metadata schema development through implementation of arrayMap resource data
  - OntologyTerm objects for covering biodata
  - implementation using EMBL-EBI ontology services

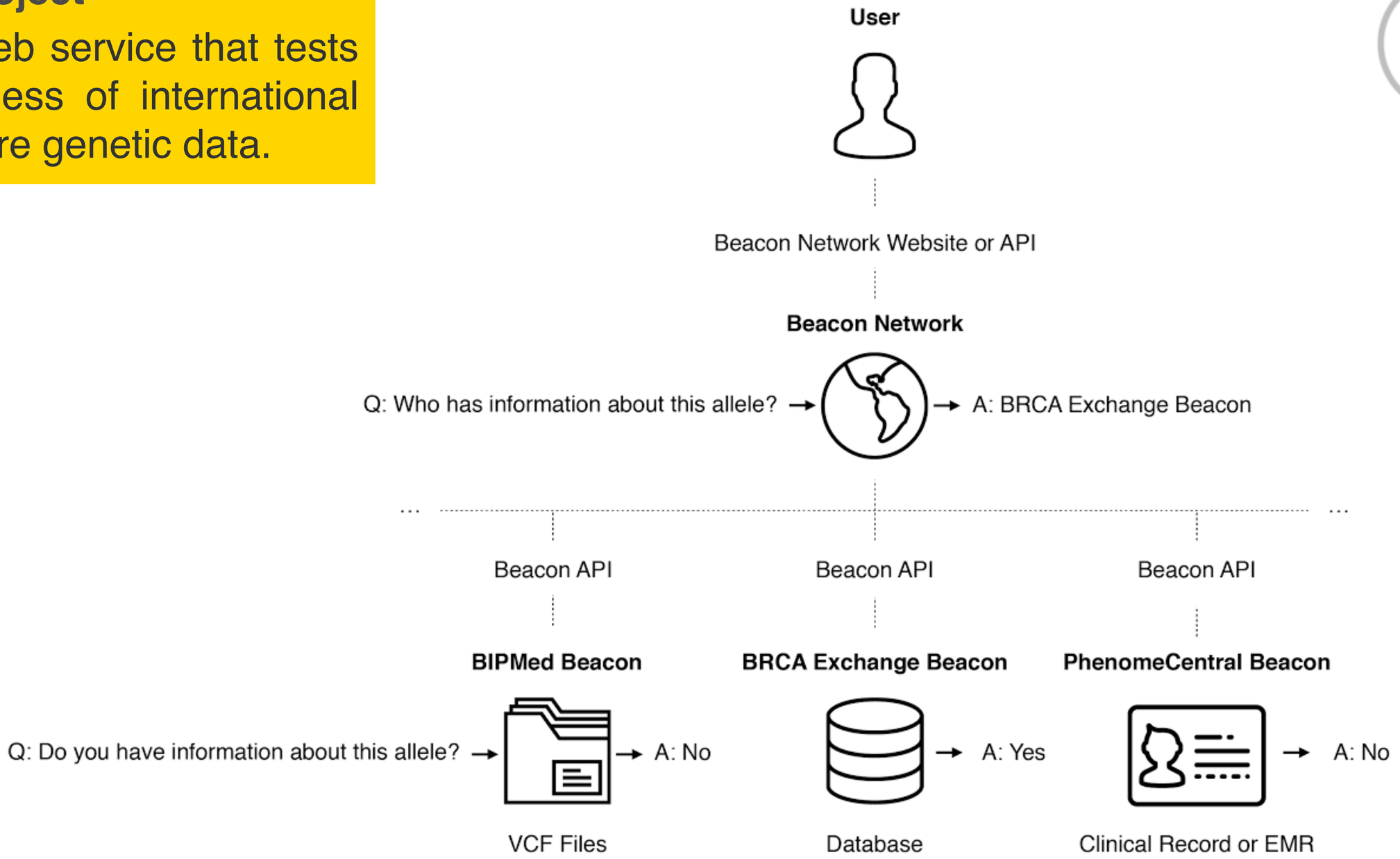
DRIVING BEACON DEVELOPMENT

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



Beacon Project

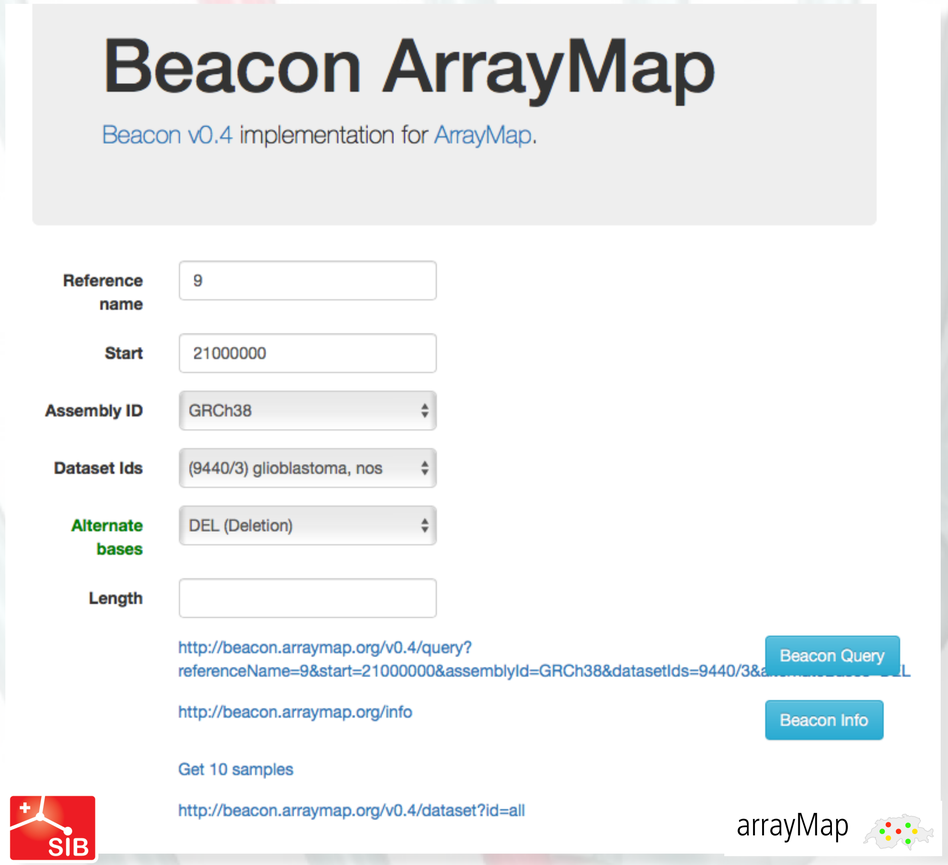
An open web service that tests the willingness of international sites to share genetic data.

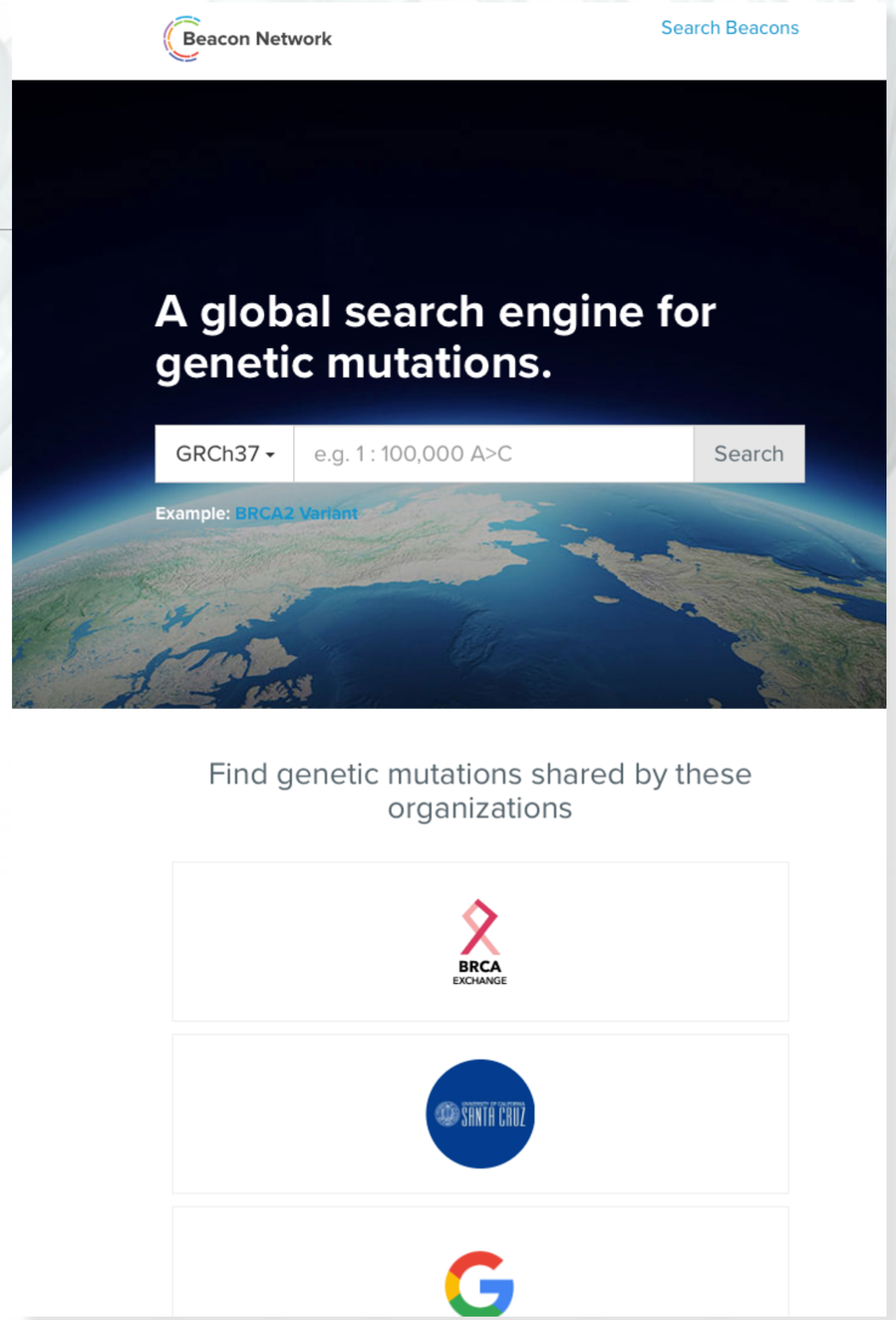


Global Alliance for Genomics & Health

Beacon+ Concept

- Implementation of cancer beacon prototype, backed by arrayMap data
- structural variation, quantitative queries, metadata





Heinz Stockinger, Séverine Duvaud & SIB Technology Group

GA4GH METADATA OBJECT IMPLEMENTATION

METADATA SCHEMA – "BIOMETADATA" OBJECTS

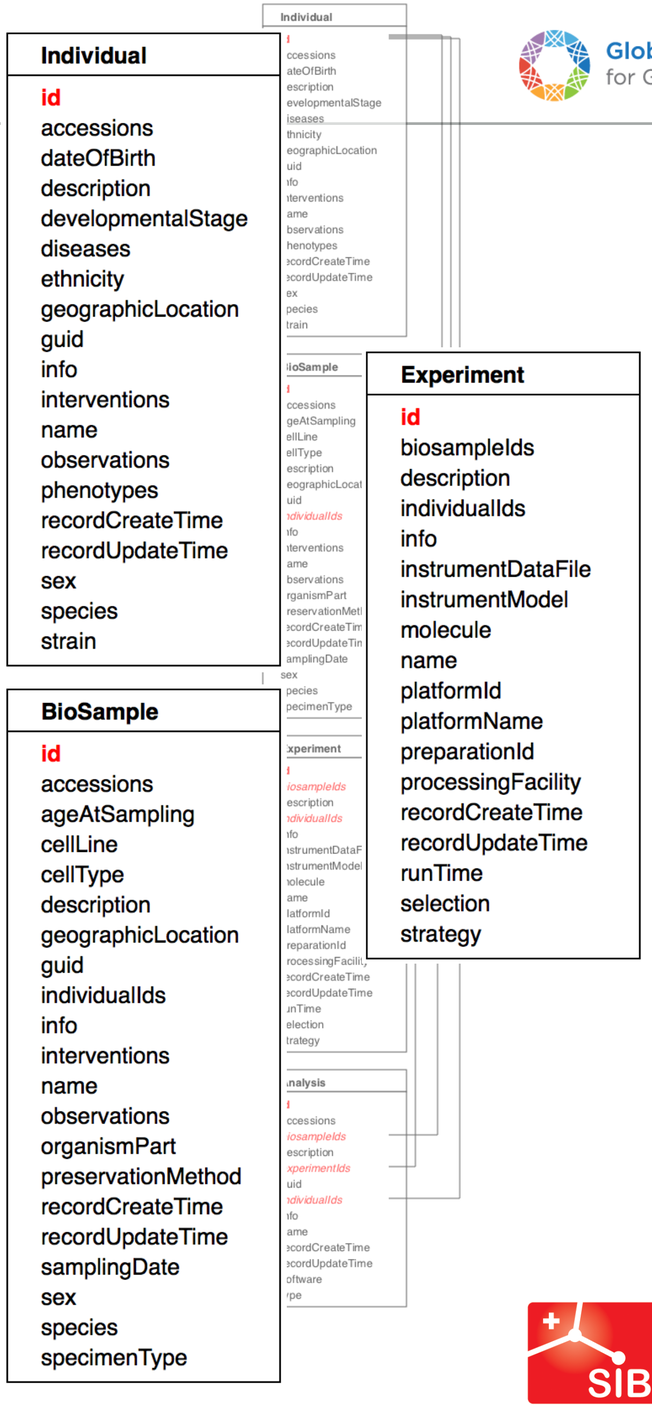
- ***Individual*** (i.e., basic biological entity)
- ***BioSample*** (e.g. micro dissected part of a tissue biopsy, environmental sample) are the basic "Bioobjects"

METADATA SCHEMA – "ASSAYMETADATA" OBJECTS

- ***Experiment*** describes the technical procedures used in the analysis of (an aliquot) of the *BioSample*
- ***Analysis*** for "interpretations" results of an experiment

ARRAYMAP FOR GA4GH SCHEMA DEVELOPMENT

- metadata schema development through implementation of arrayMap resource data
- testing of ***OntologyTerm*** object for biometadata
- implementation using EMBL-EBI ontology services



CORE PROBLEMS AND CONCEPTS TO BE ADDRESSED

- suitability of **ontologies as core of metadata** features for federated data mining
- mapping ontologies: **WHO** and **HOW**
- identification of essential non-OT attributes and stable definition using internationally accepted standards
- development of a strategy for implementation of **ontology based data annotations for reference data resources**, e.g. Elixir, EBI, SIB …

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 a superabundance of observations and classification items.
- We do not have (and never will?) enough ontology concepts for detailed disease descriptions (Where to stop?)
- Relationships may help - but how to do them uniformly?

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

ONTOLOGIES ARE RARELY USED IN CASE REPORTING

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

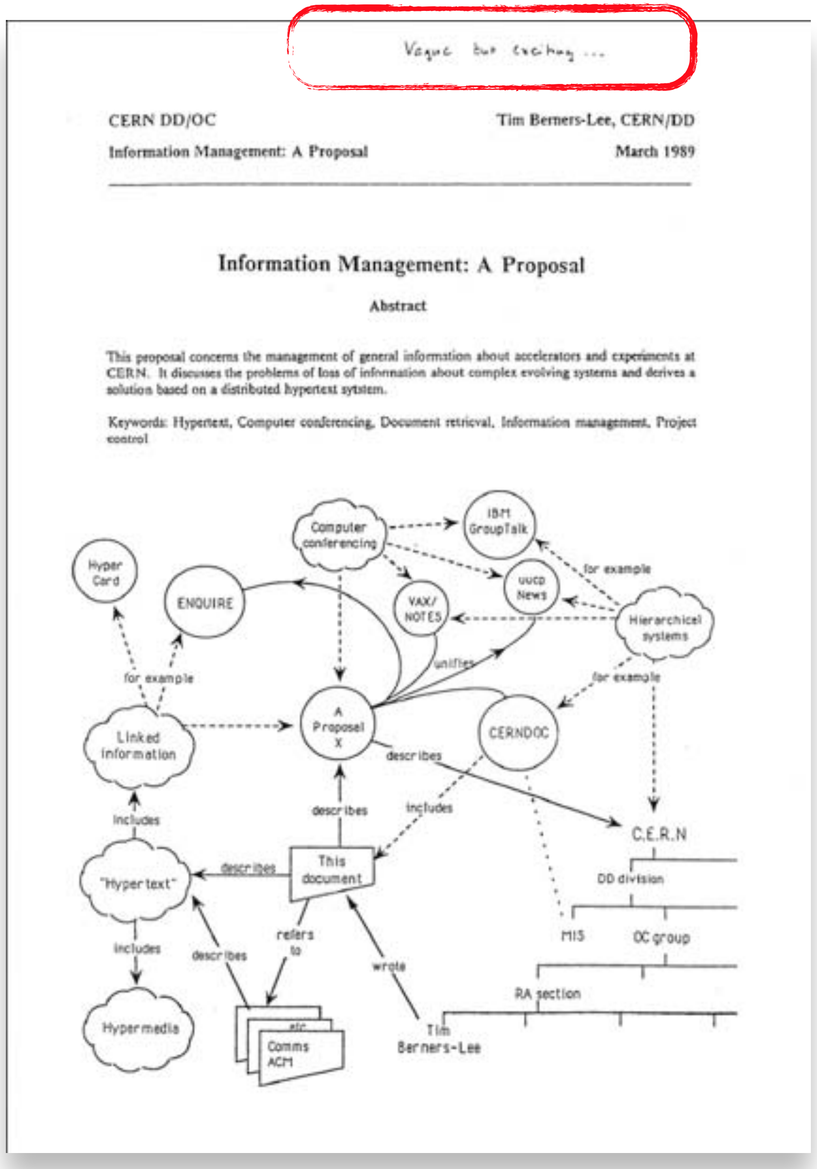
Extended ontologies

Ontology tools && services

Structured data && Data structures

Clinical data curation

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



Vague but exciting ...

CERN DD/OC — Tim Berners-Lee, CERN/DD

Information Management: A Proposal — March 1989

Information Management: A Proposal

Abstract

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext sytstem.

Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control

Tim Berners-Lee: Information Management: A Proposal (CERN 1989) & WWW: First Page (1990)

World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
: Pointers to the world's online information, subjects , W3 servers, etc.

Help
: on the browser you are using

Software Products
: A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)

Technical
: Details of protocols, formats, program internals etc

Bibliography
: Paper documentation on W3 and references.

People
: A list of some people involved in the project.

History
: A summary of the history of the project.

How can I help ?
: If you would like to support the web..

Getting code
: Getting the code by anonymous FTP , etc.

Genome Privacy

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.

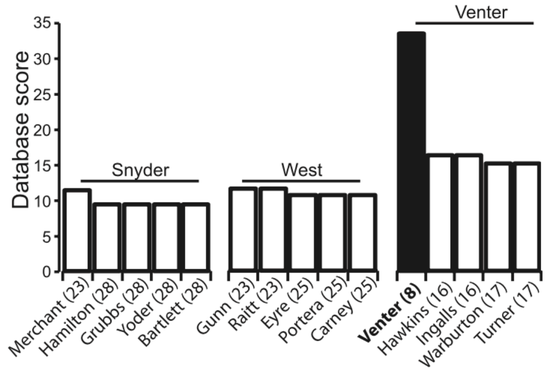


Fig. 2. The top five records retrieved after searching Ysearch with the Y-STR haplotypes of Michael Snyder, John West, and Craig Venter. The expected number of generations to the MRCA is given in parentheses for each record. Searching with Craig Venter returned a "Venter" record (closed bar) as the top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

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

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.

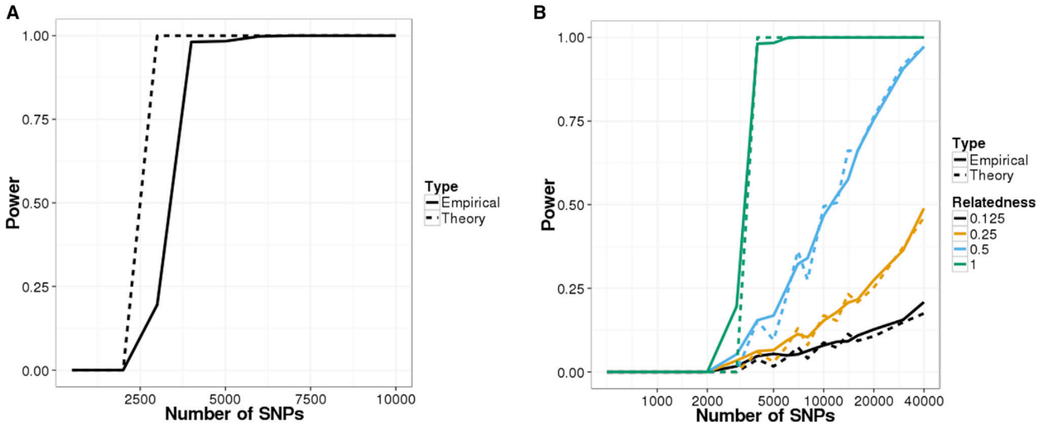
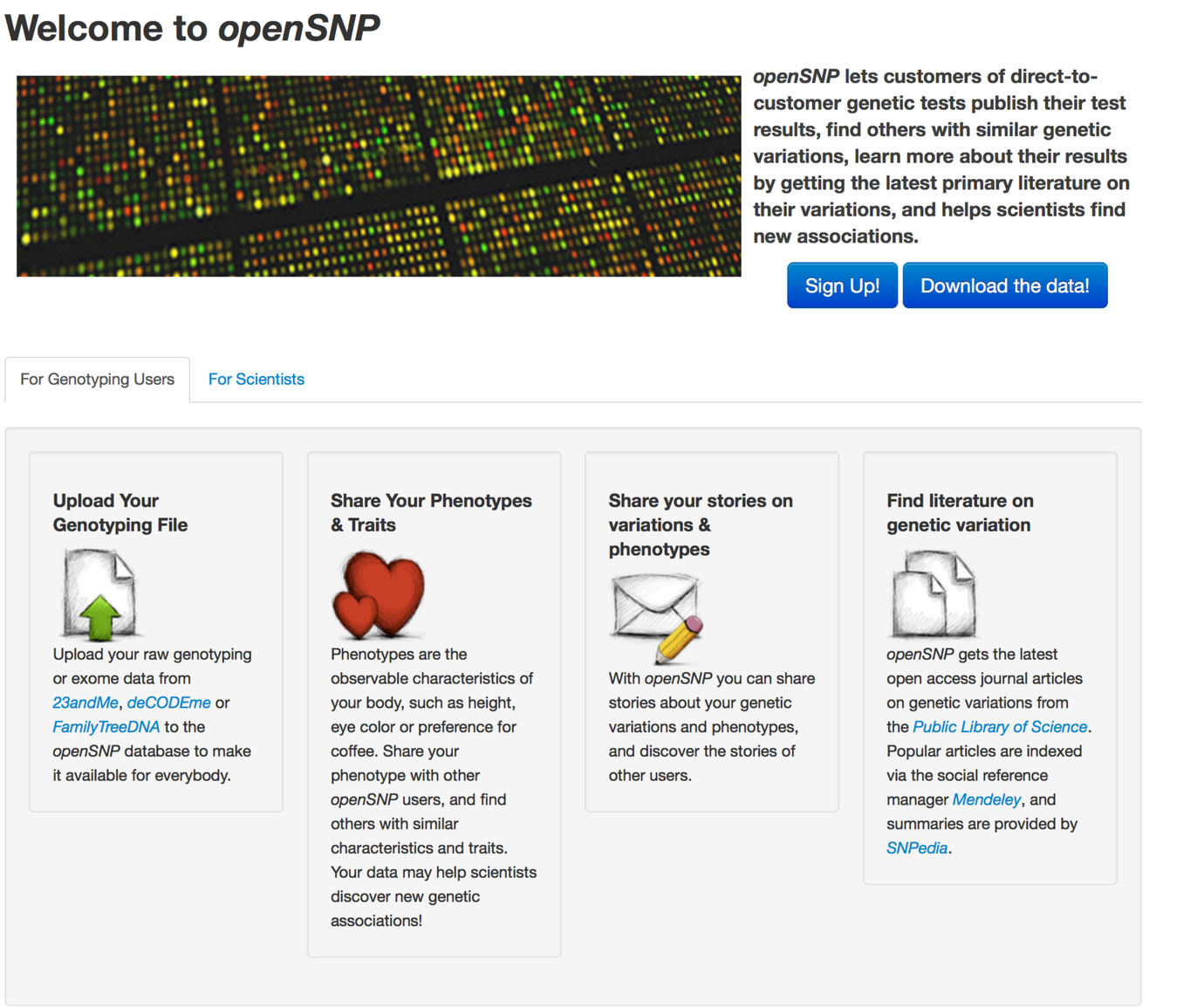


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

- rare allelic variants can be used to identify an individual (or her relatives) in a genome collection wothout having access to individual datasets
- however, such an approach requires previous knowledge about the individual's SNPs

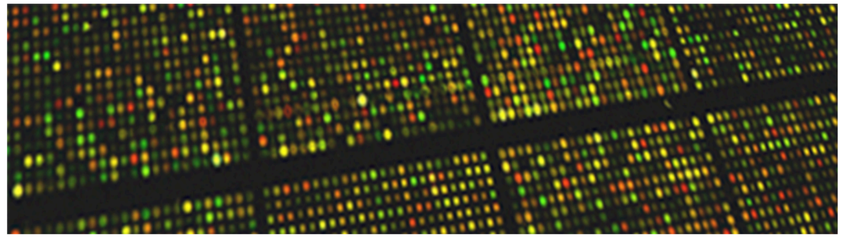
SHARE YOUR GENOME DATA?

- 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*.

BAUDISGROUP @ UZH

NI AI
MICHAEL BAUDIS
(HAOYANG CAI)
PAULA CARRIO CORDO
LINDA GROB
SAUMYA GUPTA
ROMAN HILLJE
(NITIN KUMAR)
ALESSIO MILANESE

University of Zurich[UZH]

progenetix

arrayMap

SIB

HEINZ STOCKINGER
SÉVERINE DUVAUD
VASSILIOS IOANNIDIS
DANIEL TEIXEIRA

SIB Swiss Institute of Bioinformatics

Global Alliance for Genomics & Health

EMBL-EBI

UNIVERSITY OF CALIFORNIA SANTA CRUZ

GA4GH DWG + CWG

JACQUI BECKMANN
ANTHONY BROOKES
MELANIE COURTOT
MARK DIEKHANS
MELISSA HAENDEL
DAVID HAUSSLER
SARAH HUNT
STEPHEN KEENAN
SUZY LEWIS
DAVID LLOYD
MICHAEL MILLER
HELEN PARKINSON
ELEANOR STANLEY
DAVID STEINBERG

European Genome-phenome Archive

elixir

ELIXIR & CRG

JORDI RAMBLA DE ARGILA
SABELA DE LA TORRE PERNAS
SUSANNA REPO