Cancer Genome (Data) Analysis

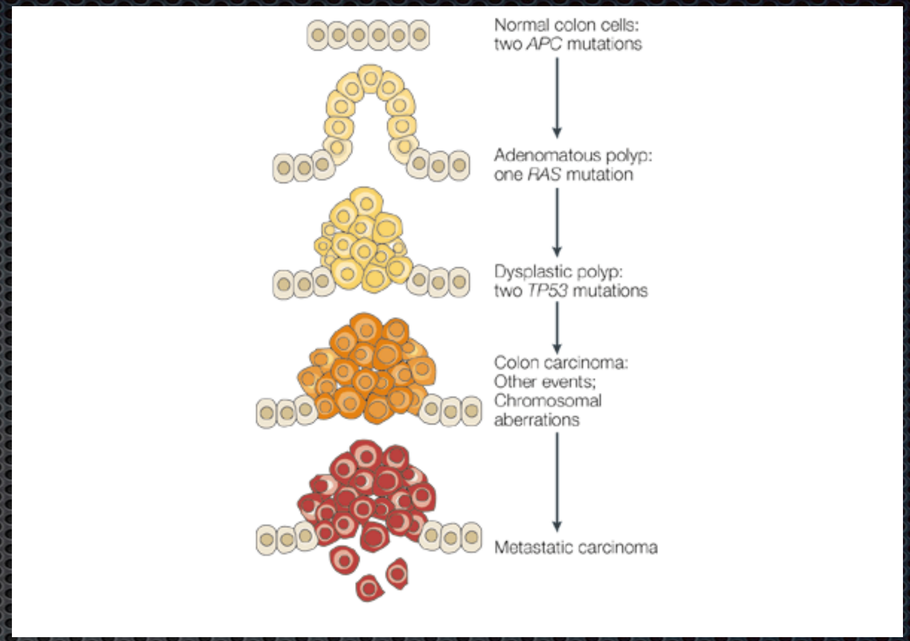
Why, Which, When, Whereto?

Michael Baudis
Computational Oncogenomics

University of Zurich[UZH]

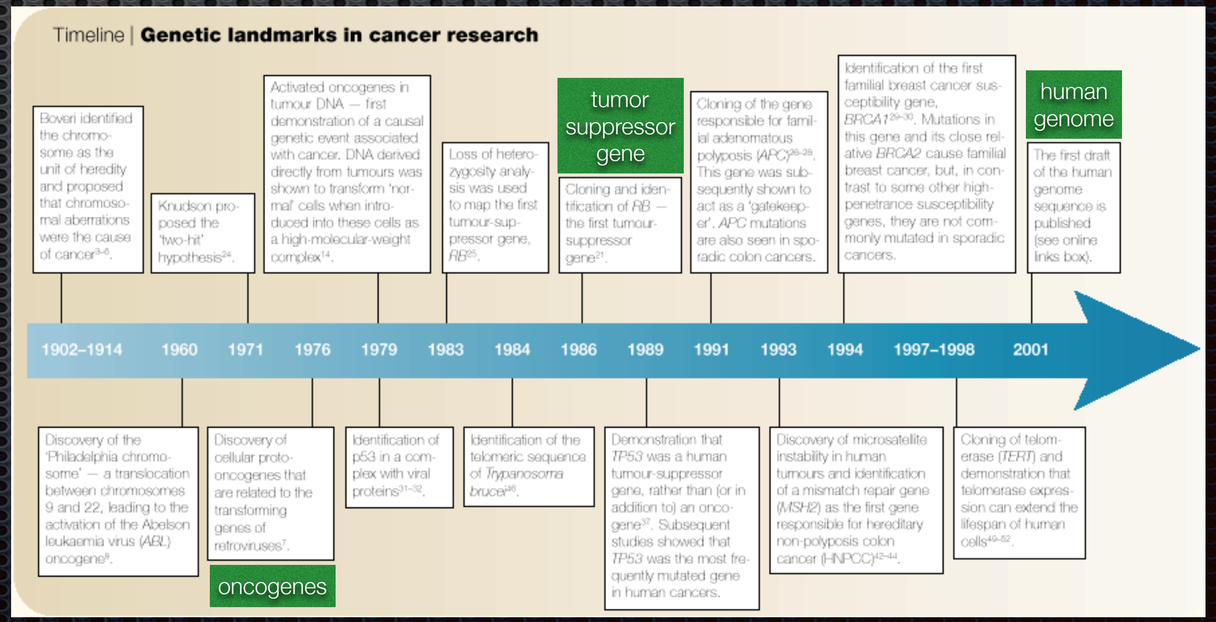
Cancers are diseases of **genomes**

- every cancer's genome is different
- every individual's genome is different



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

Cancers arise from the clonal accumulation of **somatic genome mutations**, with varying but **limited** contribution of **inherited risk**



Allan Balmain. *NatRev Cancer* (2001); 1: 77-82

Theodor Boveri (1914)

(based on observations in sea urchin eggs)

- **Cell-cycle checkpoints** (“Hemmungseinrichtung”)
- **Tumour-suppressor genes** (“Teilungshemmende Chromosomen”); can be eliminated during tumour progression
- amplified **Oncogenes** (“Teilungsfoerdernde Chromosomen” … “im permanenten Übergewicht”)
- sequential **Progression** (benign to malignant)
- Cancer **predisposition** through inheritance of less able suppressor “chromosomes”
- high-penetrance cancer syndromes - (e.g. xeroderma pigmentosum) through **homozygosity**
- Clonal origin & Genetic mosaicism; wounding and inflammation in tumour promotion; loss of cell adhesion in metastasis; sensitivity of malignant cells to radiation therapy (based on Hertwig *et al.*)

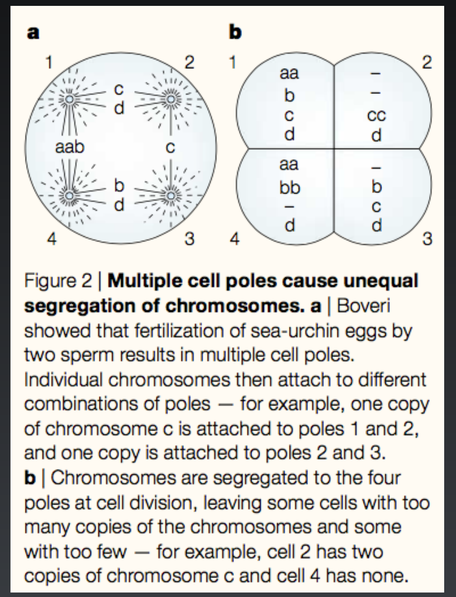
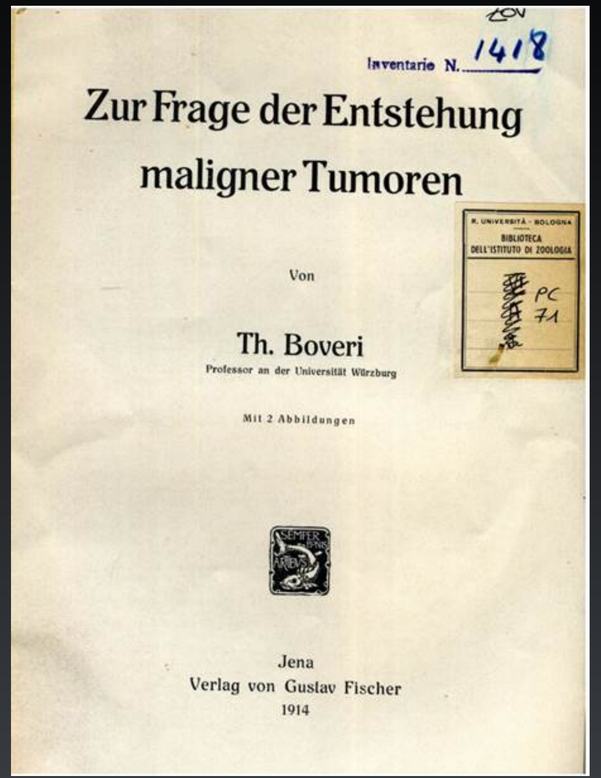
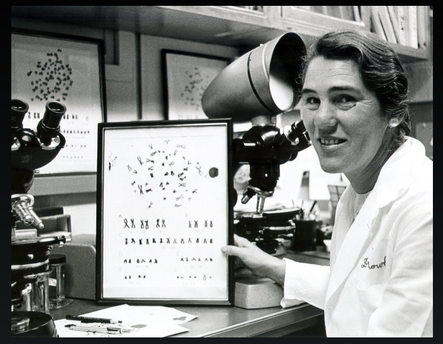


Figure 2 | **Multiple cell poles cause unequal segregation of chromosomes. a** | Boveri showed that fertilization of sea-urchin eggs by two sperm results in multiple cell poles. Individual chromosomes then attach to different combinations of poles — for example, one copy of chromosome c is attached to poles 1 and 2, and one copy is attached to poles 2 and 3. **b** | Chromosomes are segregated to the four poles at cell division, leaving some cells with too many copies of the chromosomes and some with too few — for example, cell 2 has two copies of chromosome c and cell 4 has none.

Allan Balmain
Cancer genetics: from Boveri and Mendel to microarrays.
NatRev Cancer (2001); 1: 77-82



Anna Di Lonardo , Sergio Nasi , Simonetta Pulciani
Cancer: We Should Not Forget The Past
Journal of Cancer (2015), Vol. 6: 29-39
(for book cover & summary)



Janet Rowley (1972/73)

Chromosomal translocations in cancer

- Recurrent chromosomal translocations in leukemias and lymphomas
- "Philadelphia chromosome" in CML (Nowell & Hungerford, 1960) represents a reciprocal translocation between chromosomes 9 and 22
- 1972: t(8;21) ALL manuscript rejected by NEJM
- 1973: t(9;22) manuscript rejected by *Nature* "*with some reasonable comments and some truly wrong*"
- Clinical implications: **Tyrosine Kinase inhibitors** as standard first-line therapy in CML
  - first trials in 1998 (STI-571; Imatinib/Gleevec)
  - cf. Druker BJ, Lydon NB (2000). Lessons learned from the development of an Abl tyrosine kinase inhibitor… J Clin Invest 2000;105:3-7)

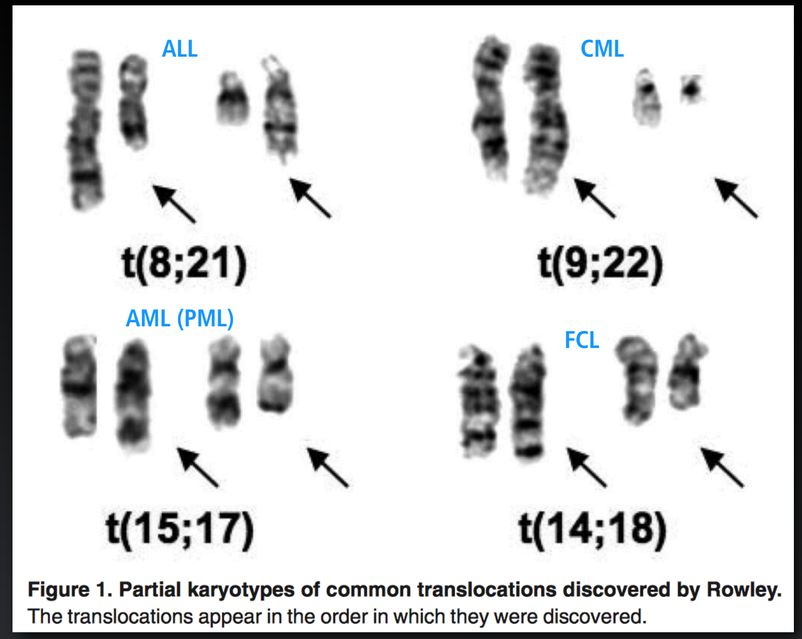
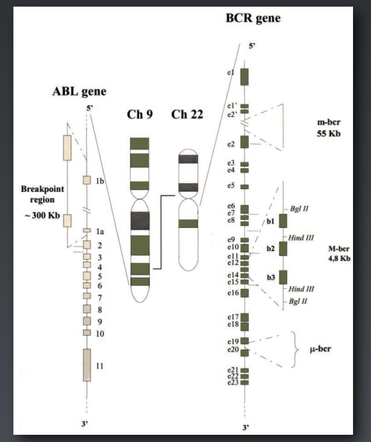
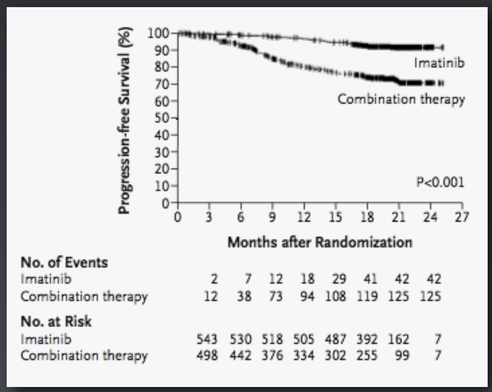


Figure 1. Partial karyotypes of common translocations discovered by Rowley. The translocations appear in the order in which they were discovered.

Janet D Rowley. Chromosomal translocations: revisited yet again *Blood* (2008), 112(6)



Pane et al. BCR/ABL genes …. *Oncogene* (2002), 21 (56)

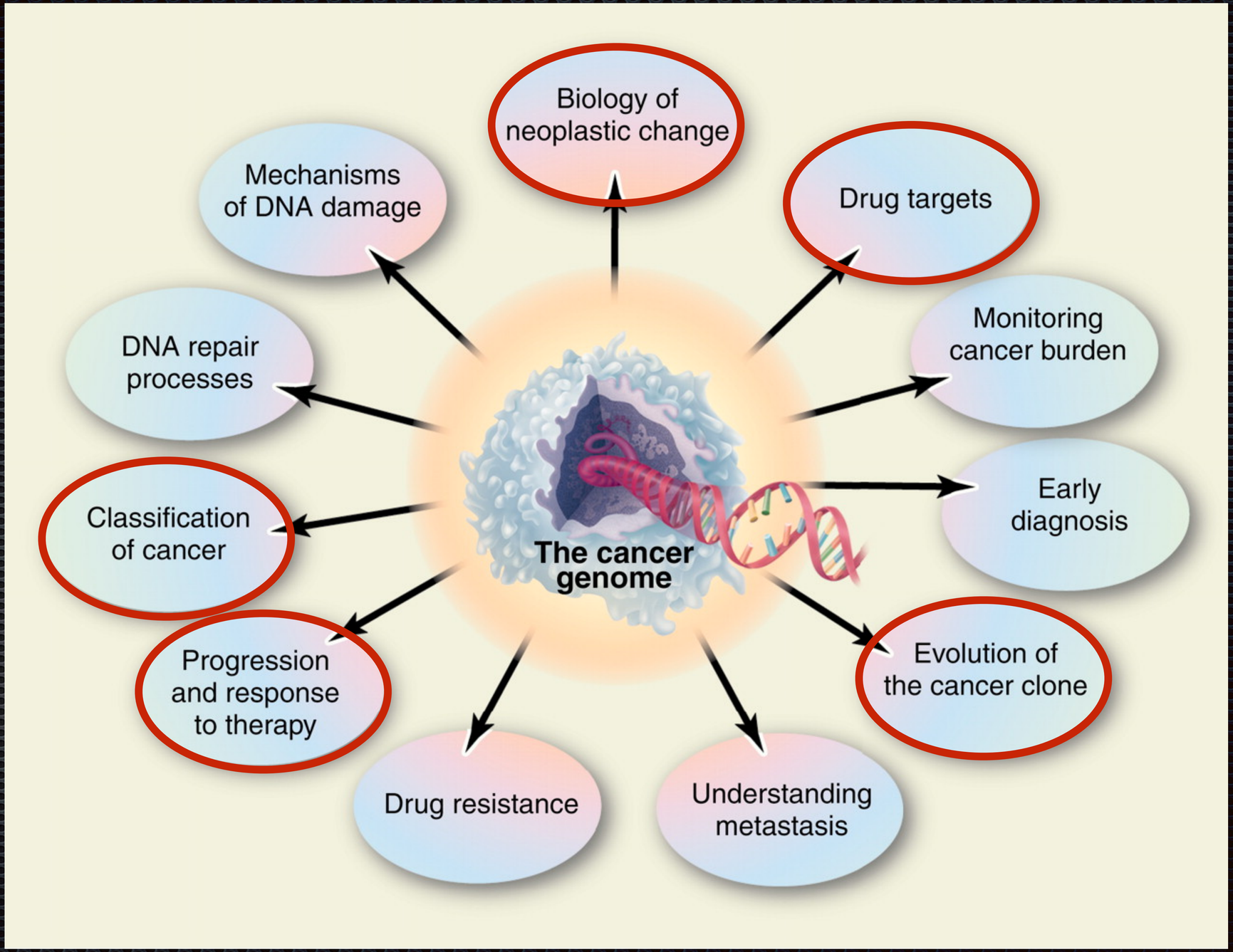


Event free Survival in first large Imatinib Trials

O'Brien et al. Imatinib compared with interferon and low-dose cytarabine… NEJM (2003) vol. 348 (11)

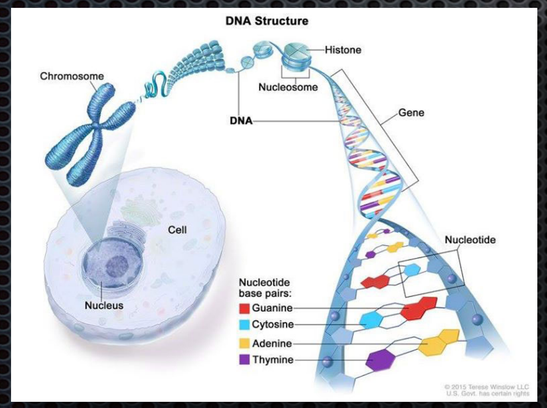
Why?

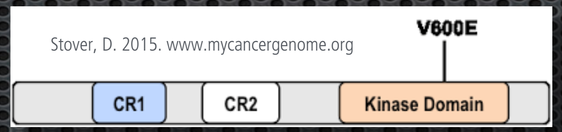
Michael R. Stratton. Exploring the Genomes of Cancer Cells: Progress and Promise. Science (2011)

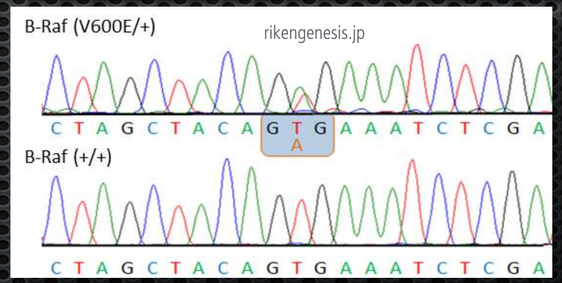


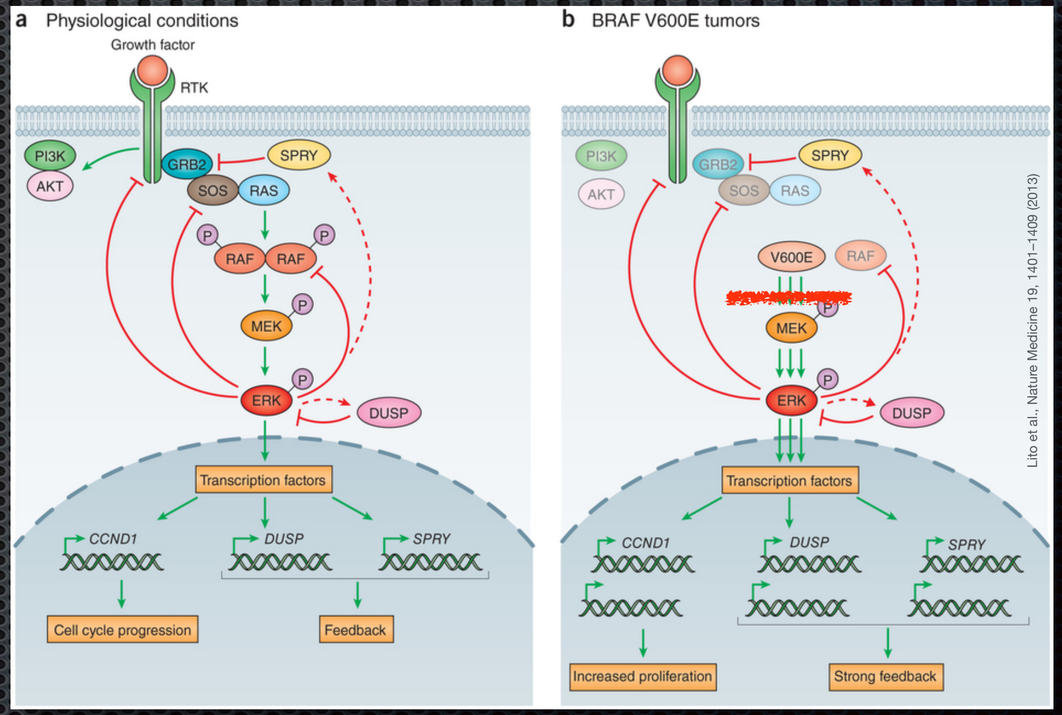
BRAF V600E (c.1799T>A) Mutation
Oncogene Activation by Single Nucleotide Alteration

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





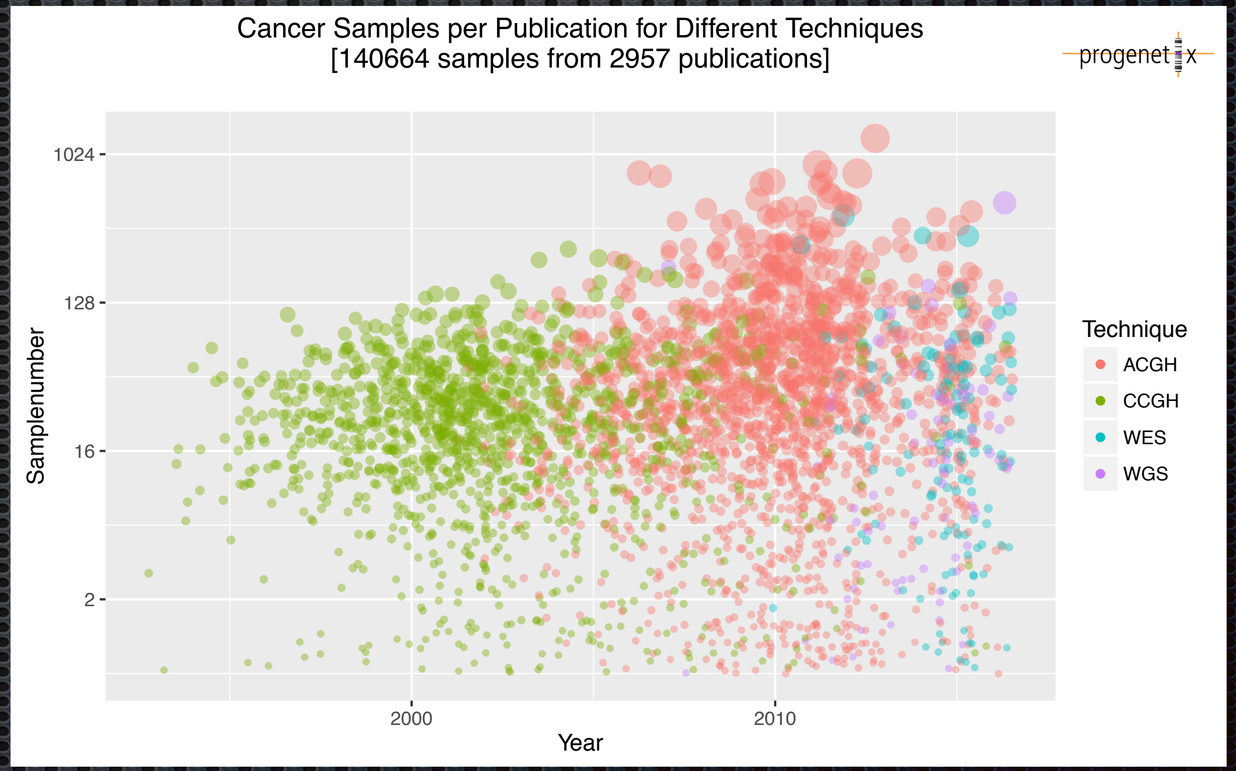




The BRAF V600E mutation leads to continuos phosphorylation of MEK, without the need for receptor based activation of the upstream pathway and loss of inhibitory feedback control.

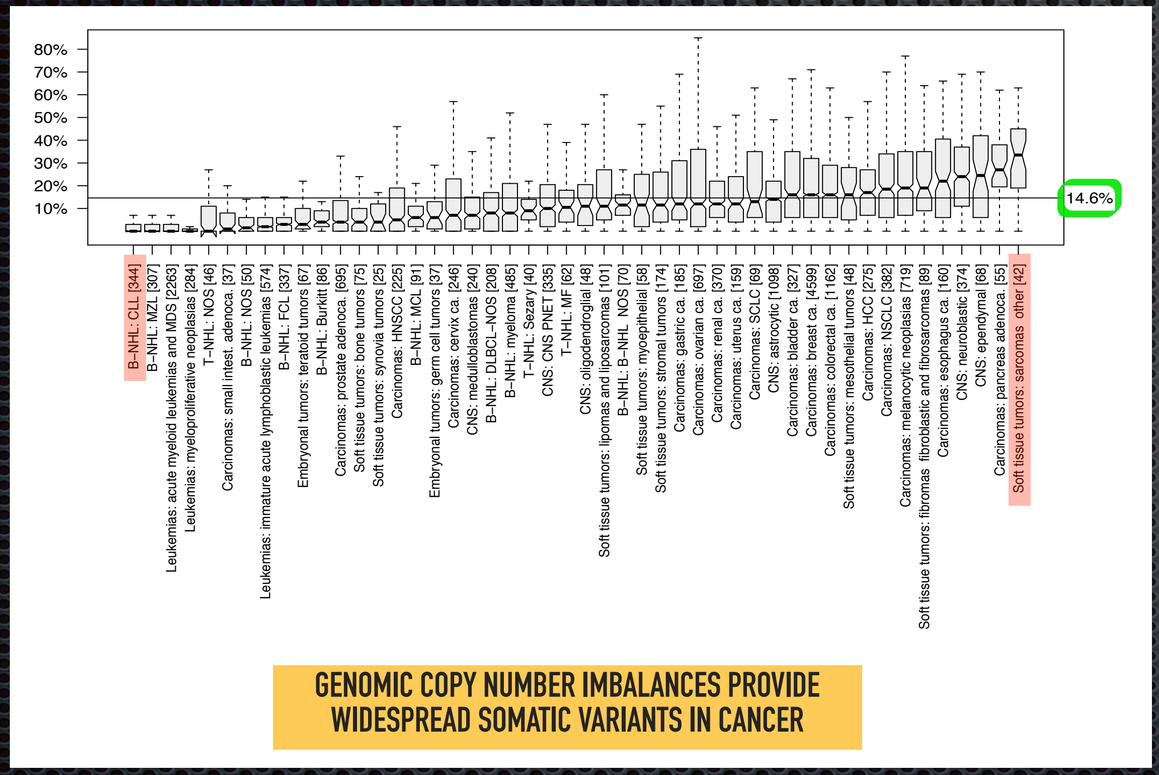
Molecular Cytogenetic & Sequencing Studies for **Whole Genome Profiling** in Cancer

- 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)



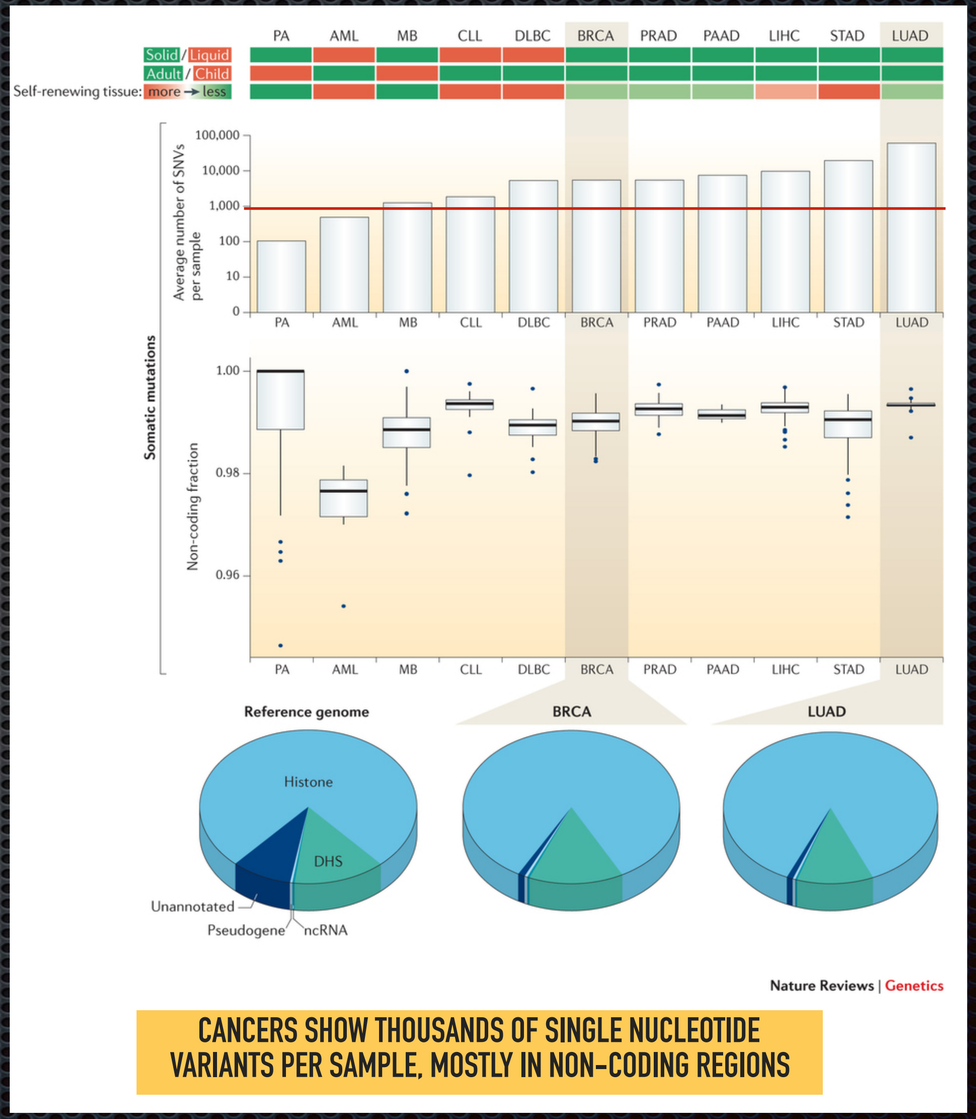
Overview of publications reporting whole-genome screening analysis of cancer samples, by molecular-cytogenic or genome sequencing methods. The data represents articles assessed for the progenetix.org cancer genome data resource (M. Baudis, 2001-2016)

Quantifying Somatic Mutations In Cancer



GENOMIC COPY NUMBER IMBALANCES PROVIDE WIDESPREAD SOMATIC VARIANTS IN CANCER

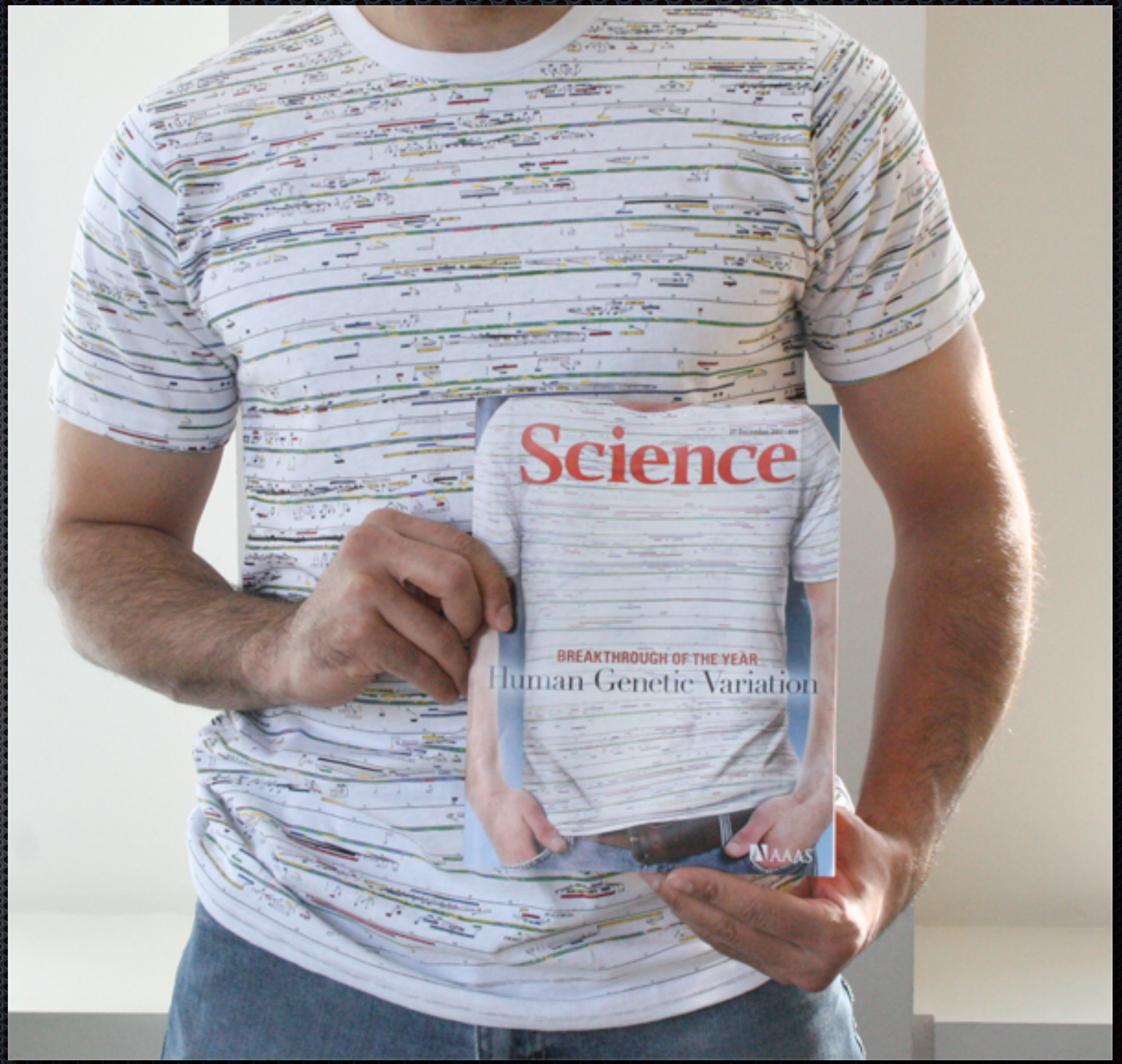
On average ~15% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on >30'000 cancer genomes from arraymap.org



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

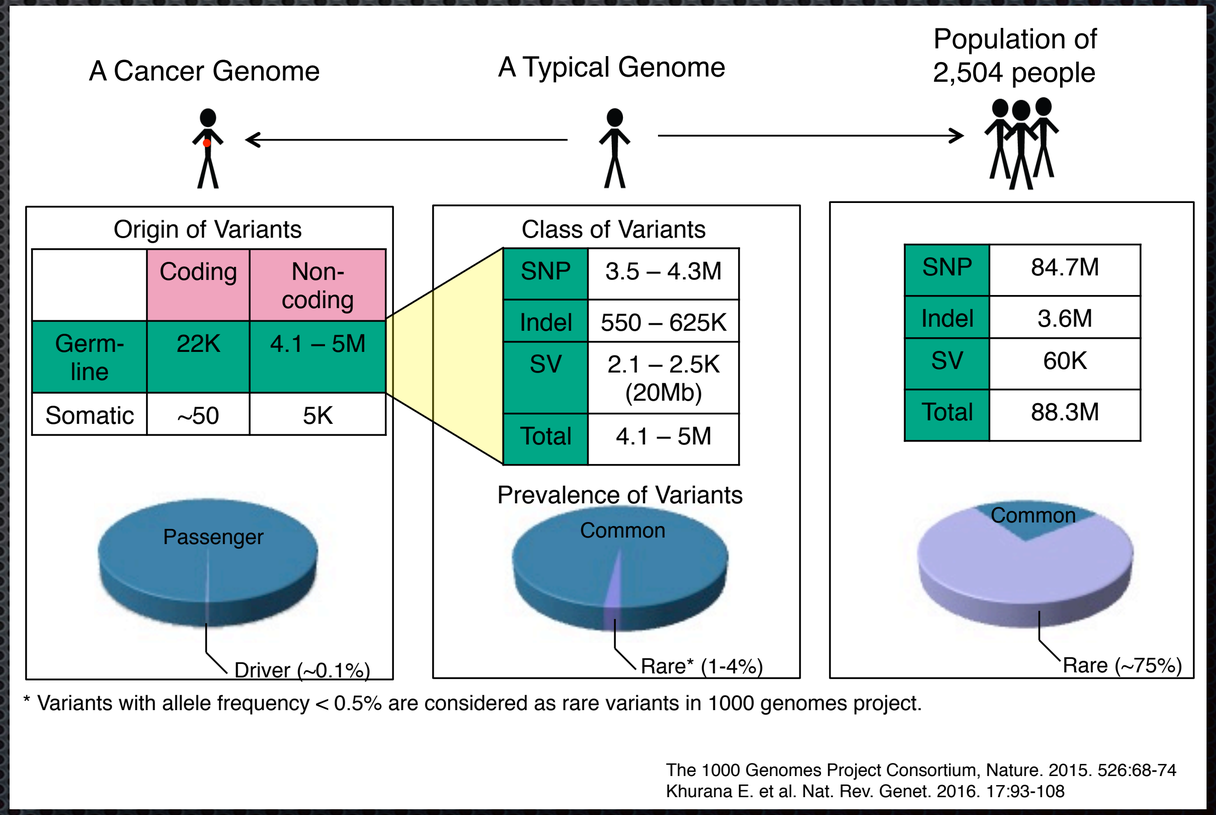
Pan-Cancer Analysis of Whole Genomes (PCAWG) data show widespread mutations in non-coding regions of cancer genomes (Khurana et al., Nat. Rev. Genet. (2016)

Personalized health & The trouble with human genetic variation



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

- a typical humen 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

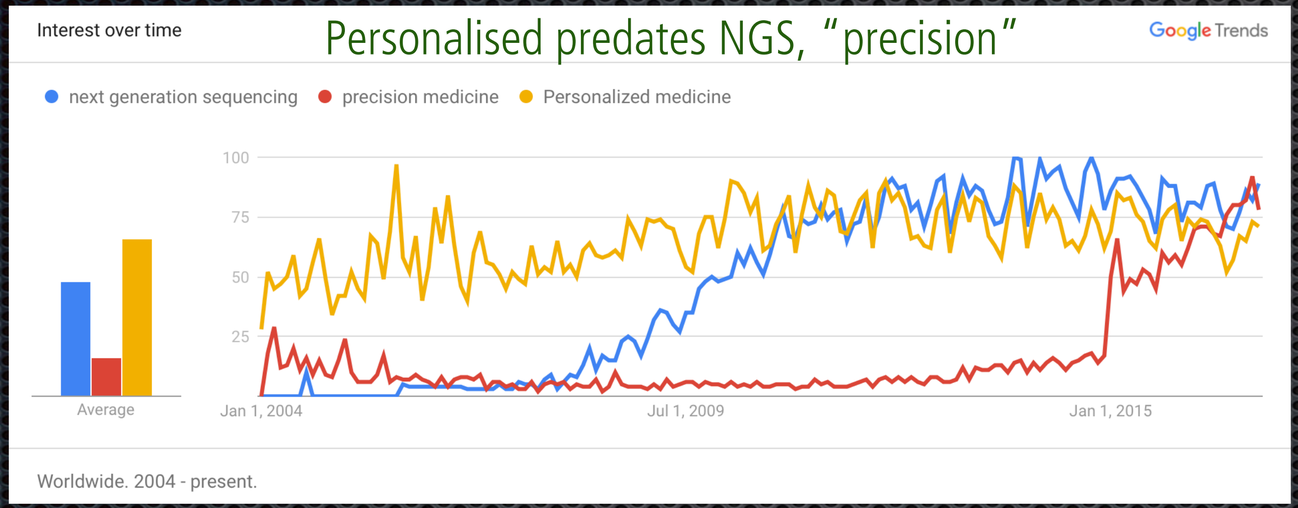
Graphic adapted from Mark Gerstein (GersteinLab.org; @markgerstein)

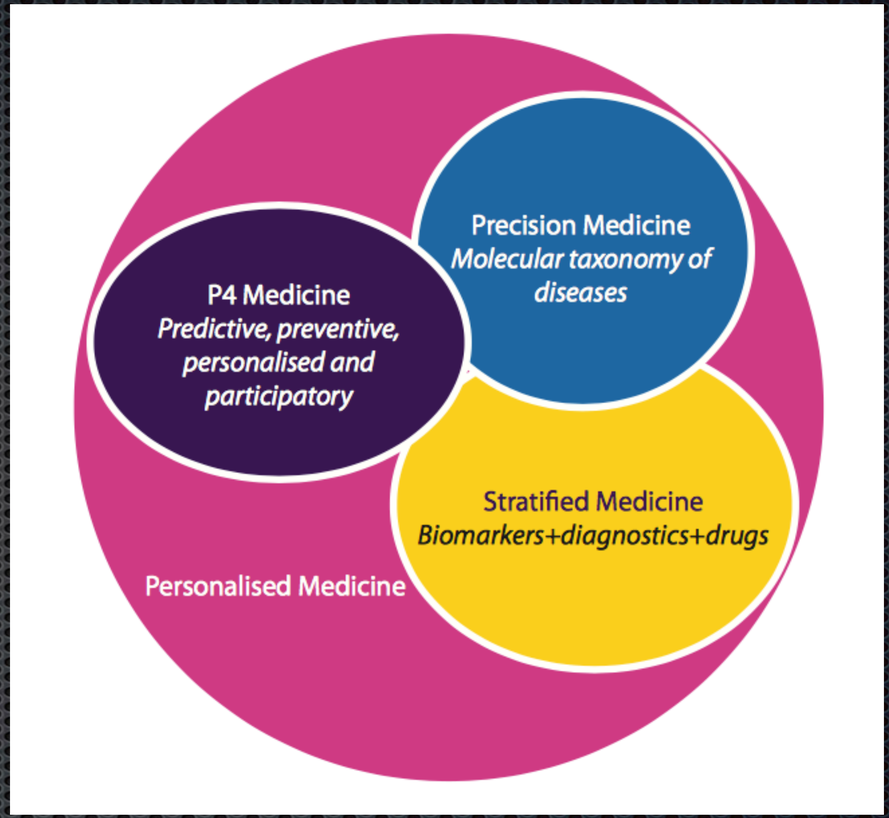
Genomic Background + Disease Parameters

Personalised Medicine

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 and individually targeted therapies.

Genome analyses at the core of Personalised Medicine™

Susceptibility, Pharmacogenomics, Classification, 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

RESEARCH ARTICLE Open Access

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

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

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]*
[1]*National Institute for Physiological Sciences,* [2]*Restart Postdoctoral Fellowship of Japan Society for the Promotion of Science, Tokyo, and* [3]*Department of Psychiatry, Nagoya University Graduate School of Medicine, Nagoya, Japan*

Common gene variants, mortality and extreme longevity in humans

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

NEURODEVELOPMENT

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

Mustafa Sahin* and Mriganka Sur*

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]

RESEARCH ARTICLE Open Access

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

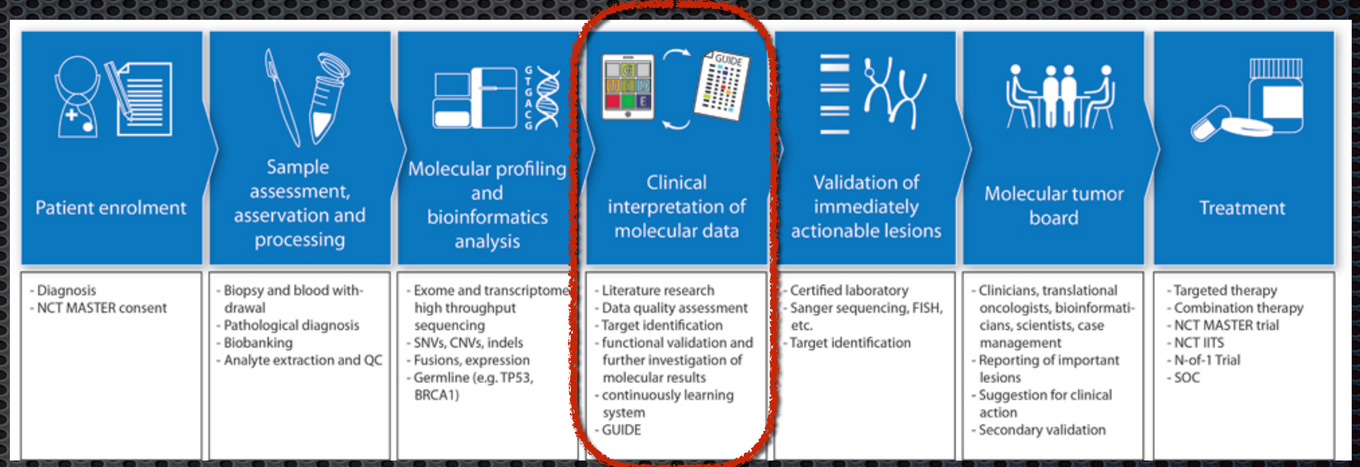
Haoyang Cai[1,2], Nitin Kumar[1,2], Homayoun C Bagheri[3], Christian von Mering[1,2], Mark D Robinson[1,2*] and Michael Baudis[1,2*]

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

Personalised Medicine in Cancer - A Genome Based Approach

- personalized cancer therapy uses information about the **individual genetic background** and **tumor sequence analysis** for the identification of somatic variants



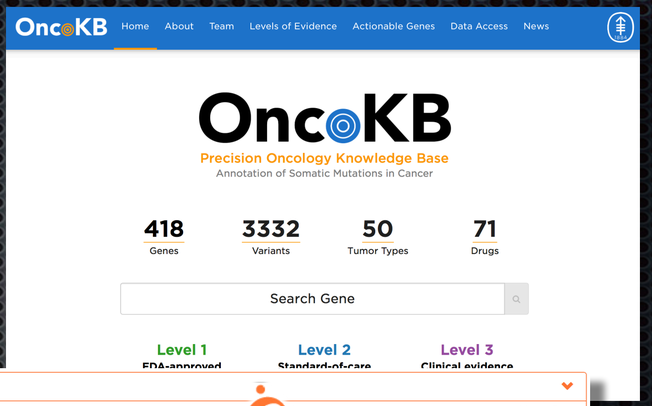
Workflow of a cancer treatment protocol based on "personalized" assessment of actionable genomic lesions (source: NCT Heidelberg).

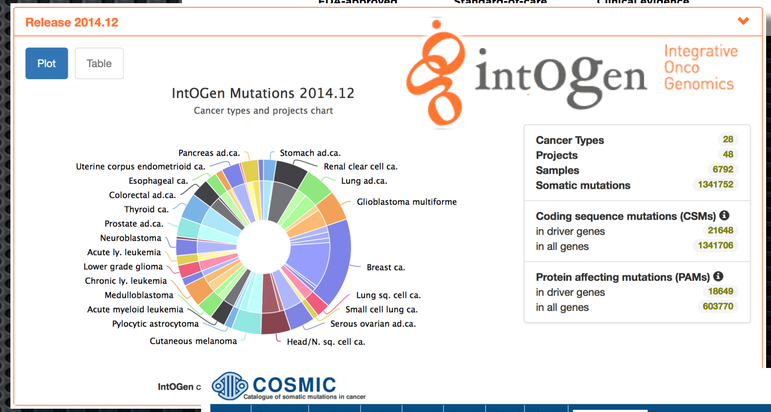
- currently mostly use of **targeted / panel sequencing** for identification of tens - hundreds of most common "actionable" mutations
- knowledge resources and literature search for interpretation of non-standard variants

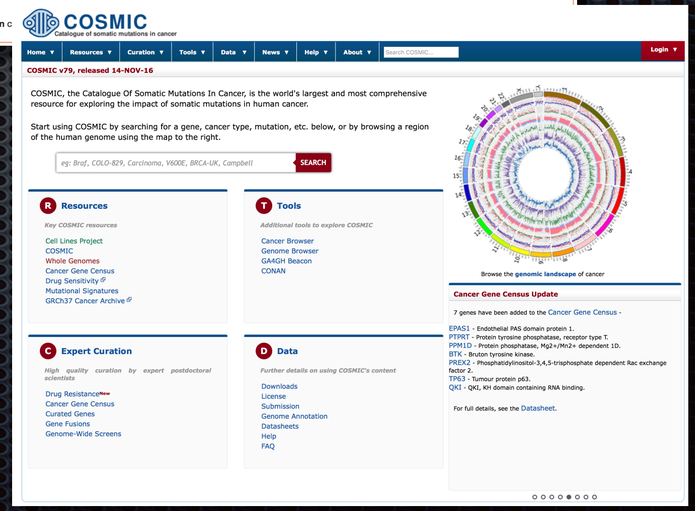
Curated Variant Data Resources as Backbone of Personalised Cancer Therapy

- cancer variant interpretation resources apply manual **data curation** and **bioinformatics** methods to provide information about putative targets and possible interventions

Database	Institute	Organized by
TARGET	BROAD	Gene
PCT	MD Anderson	Gene
cBioPortal / OncoKB	MSK	TCGA diseases
COSMIC	Sanger	Gene
IntOGen	University Pompeu Fabra	Gene
My Cancer Genome	Vanderbilt	Disease
CIViC	Washington University	Variant
DGIdb	Washington University	Drug/gene interaction



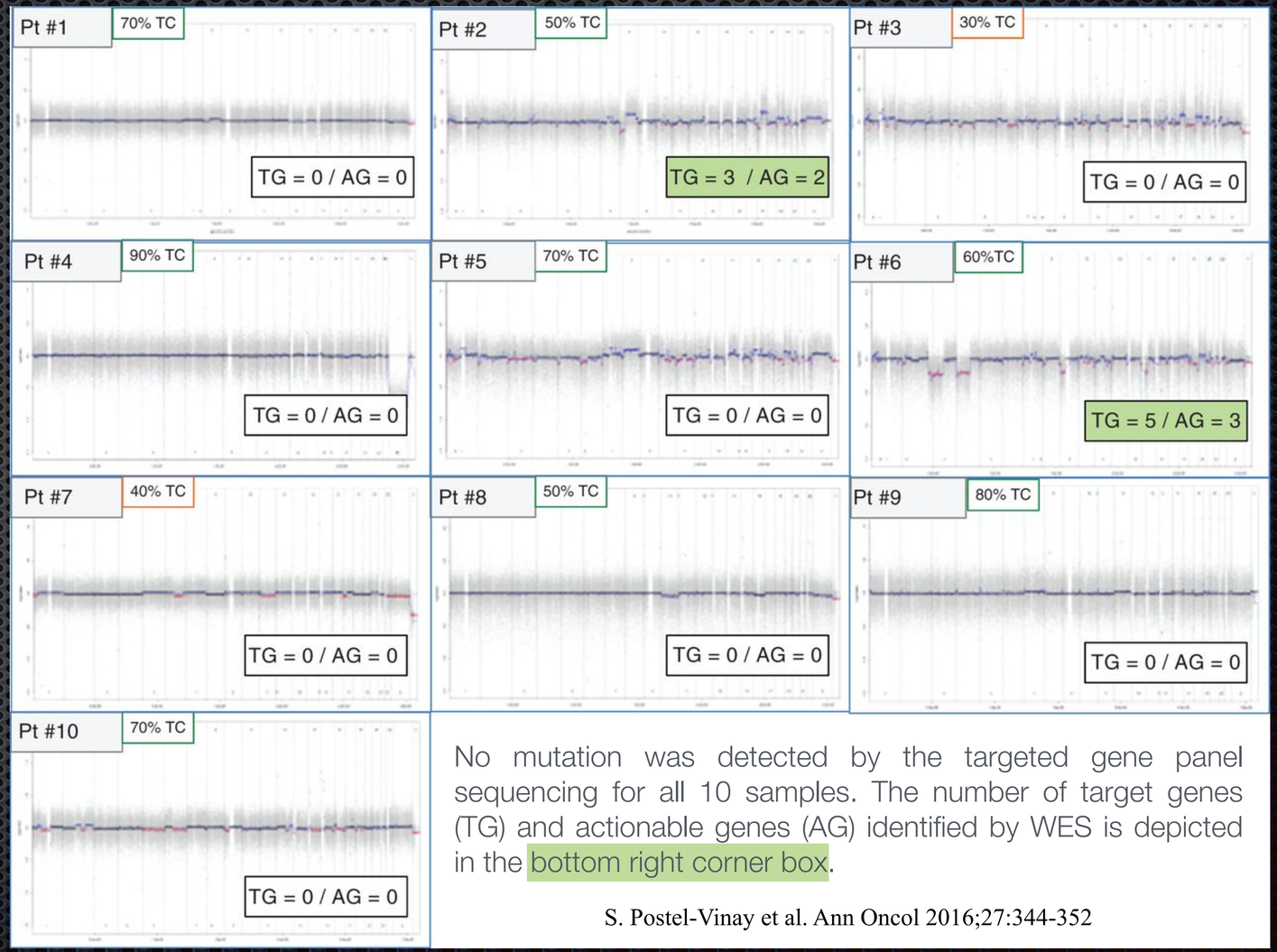




Drivers in tumours with apparent normal molecular profile on CGH and **targeted gene panel** sequencing: what is the added value of **WES**?

French "MOSCATO-01" trial (2011-2013)

- samples **without** mutation in hot spot regions of a **panel** of 46 critical cancer-related genes were submitted to WES
- actionable mutations were identified in **2/10** randomly selected patients

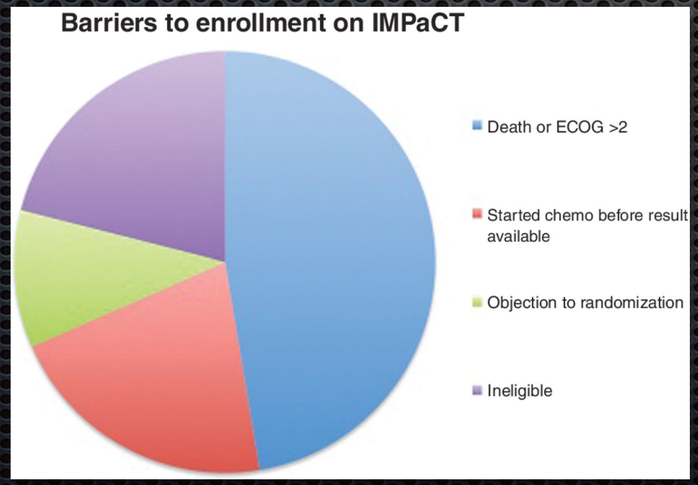


No mutation was detected by the targeted gene panel sequencing for all 10 samples. The number of target genes (TG) and actionable genes (AG) identified by WES is depicted in the bottom right corner box.

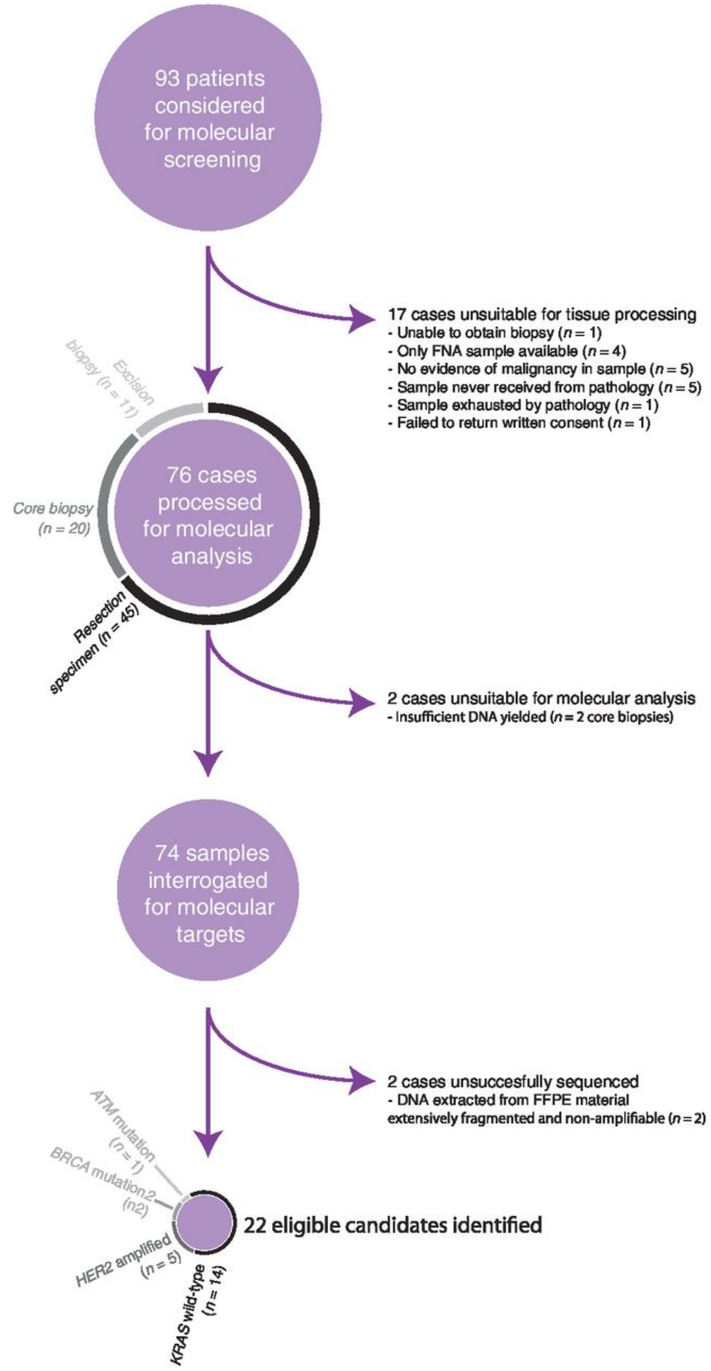
S. Postel-Vinay et al. Ann Oncol 2016;27:344-352

Personalised Cancer Therapy Needs More Targets

- with the current knowledge, targeted molecular analysis will not lead to the identification of actionable interventions in a majority of cancer cases



Precision Medicine for Advanced Pancreas Cancer: The Individualized Molecular Pancreatic Cancer Therapy (IMPaCT) Trial. Chantrill *et al.*, Clin Cancer Res. 2015 May 1;21(9):2029-37

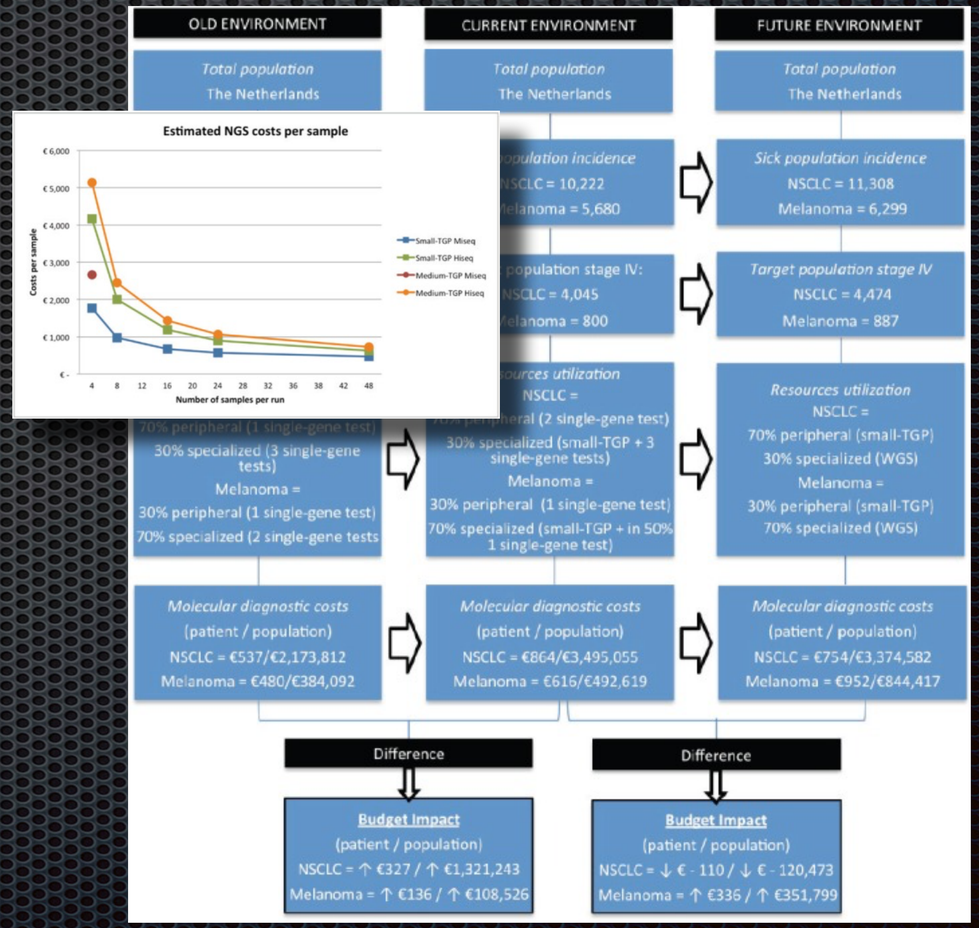


Next-generation sequencing strategies in malignancies

	Targeted panels	Whole exome	Whole genome
Pro	▶ High depth of coverage ▶ Readily standardisable ▶ Rapid interpretation for clinical use ▶ Low costs ▶ Easy clinical implementation	▶ Detection of unknown variants ▶ Detection of CNVs ▶ Research applications ▶ Feasible in clinical routine ▶ Low price/performance ratio	▶ Comprehensive assessment of cancer genomes ▶ Highest resolution of genomic alterations ▶ SNVs in enhancer/promoter and ncRNA regions ▶ Decreasing costs ▶ Subject to future studies
Contra	▶ Limited, 'peephole' observations ▶ Limited value for research ▶ Limited assessment of complex aberrations	▶ Not fully comprehensive ▶ Lower CNV resolution ▶ Amplification or exon capture necessary ▶ High bioinformatic effort ▶ Demanding clinical interpretation ▶ Time-consuming workflow	▶ Uncertain value for clinical interpretation ▶ Most expensive

CNV, copy number variant; ncRNA, non-coding RNA; SNV, single nucleotide variant.

Horak P, et al. ESMO Open 2016;1:e000094.



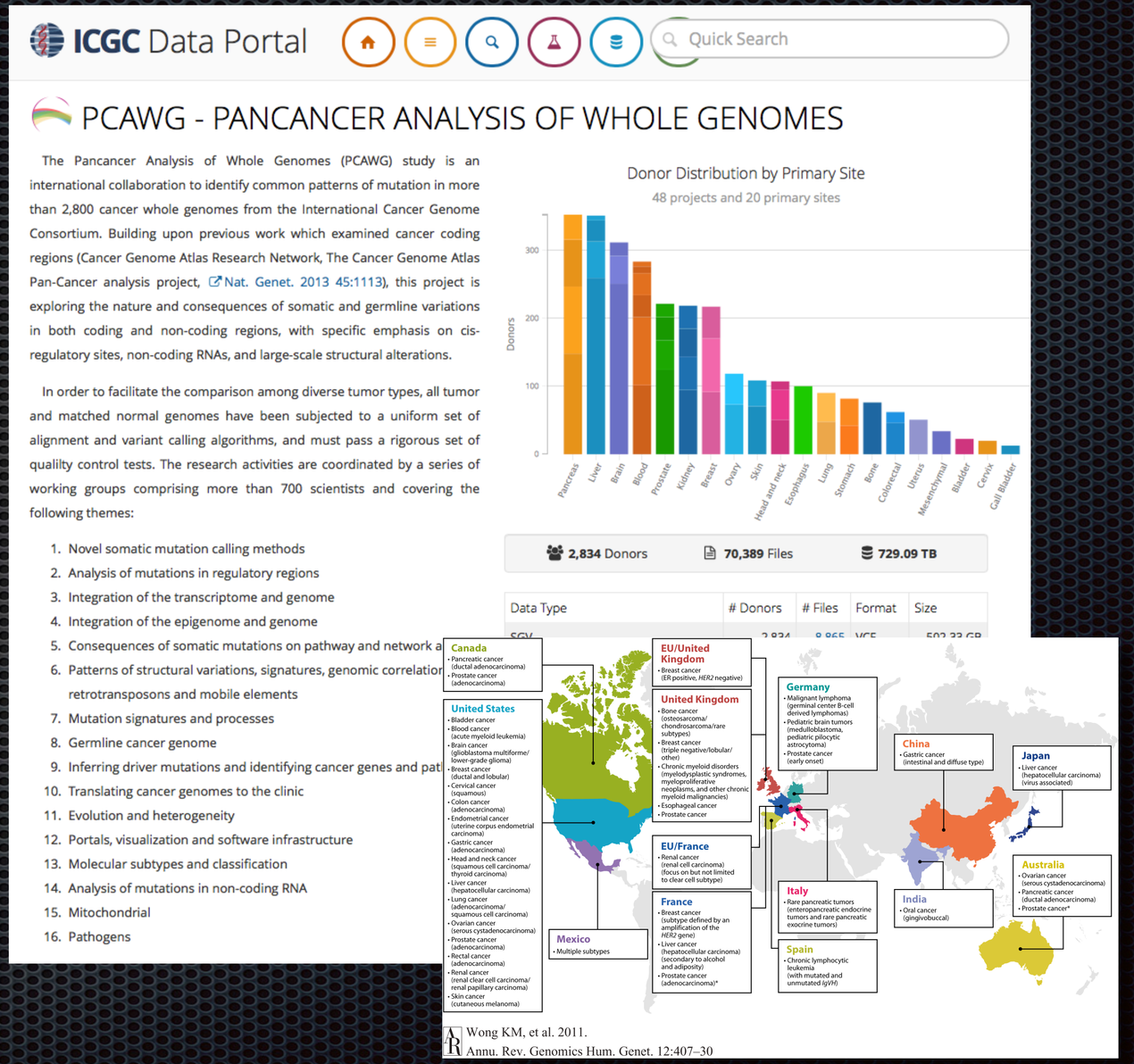
Estimated costs per sample for small- and medium TGP and WGS; van Amerongen et al., Ecancermedicalscience. 2016

WGS use is expected to reduce the cost of additional tests

Genome Data Access in Cancer
Mining for New Knowledge

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

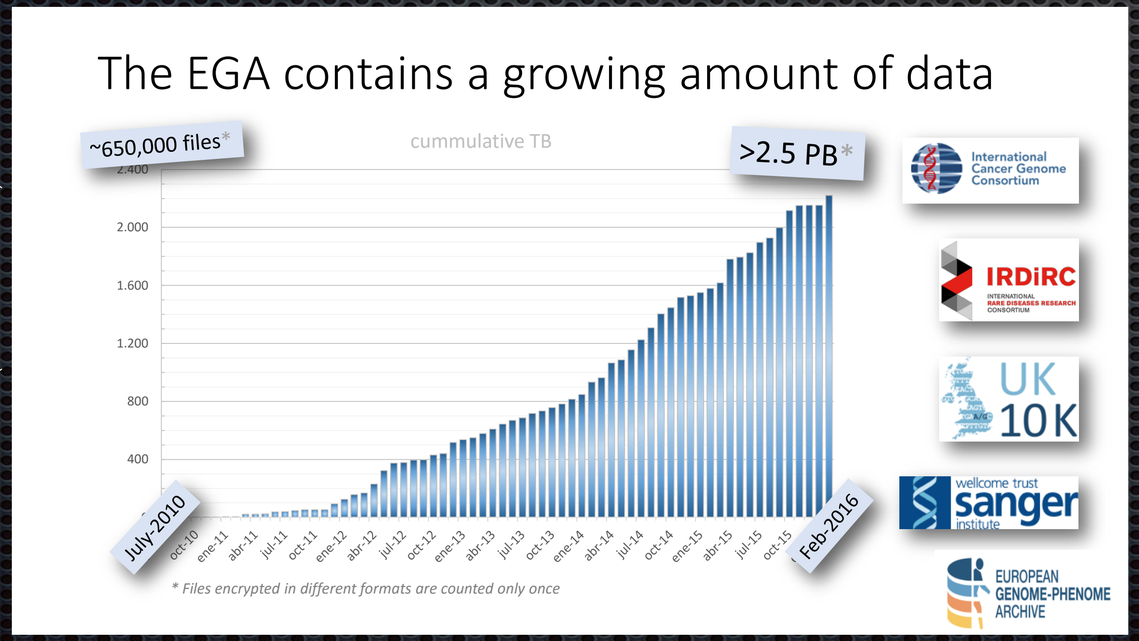
- the International Cancer Genome Consortium (ICGC) as leading example of deep analysis of multiple cancer entities
- international collbaoration of leadinc research centers for each of ~20 tumor types
- limitations:
  - focus on prominant cancer types w/ limited representation of rare entities
  - data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



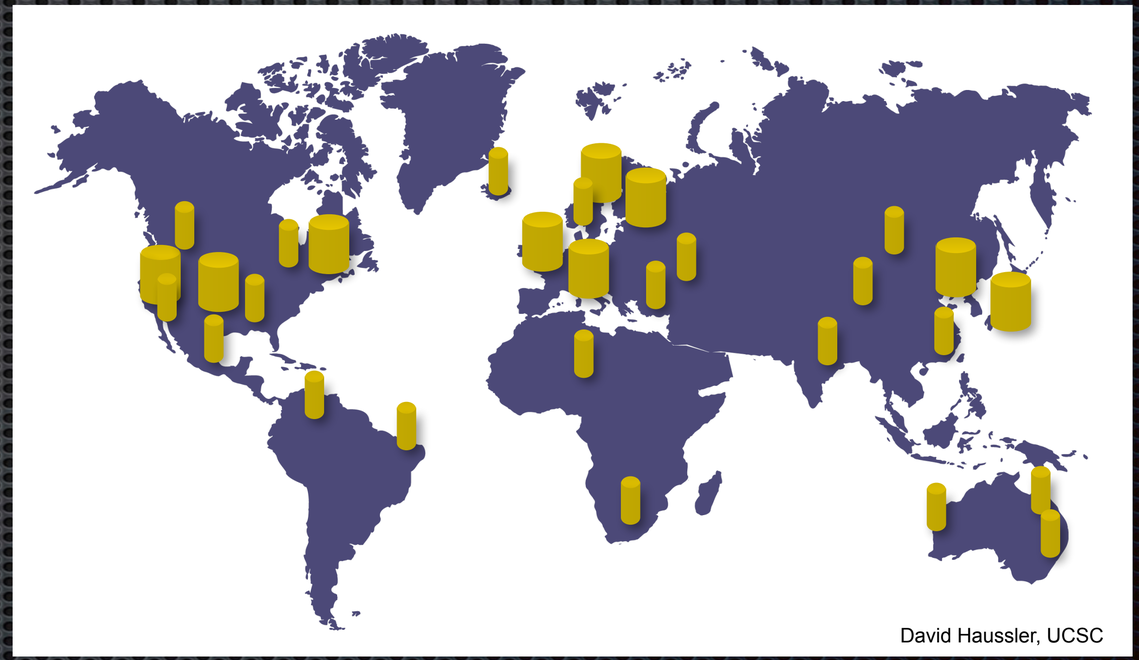
Wong KM, et al. 2011.
Annu. Rev. Genomics Hum. Genet. 12:407–30

Genome Datasets: Rapid Growth, Limited Access

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



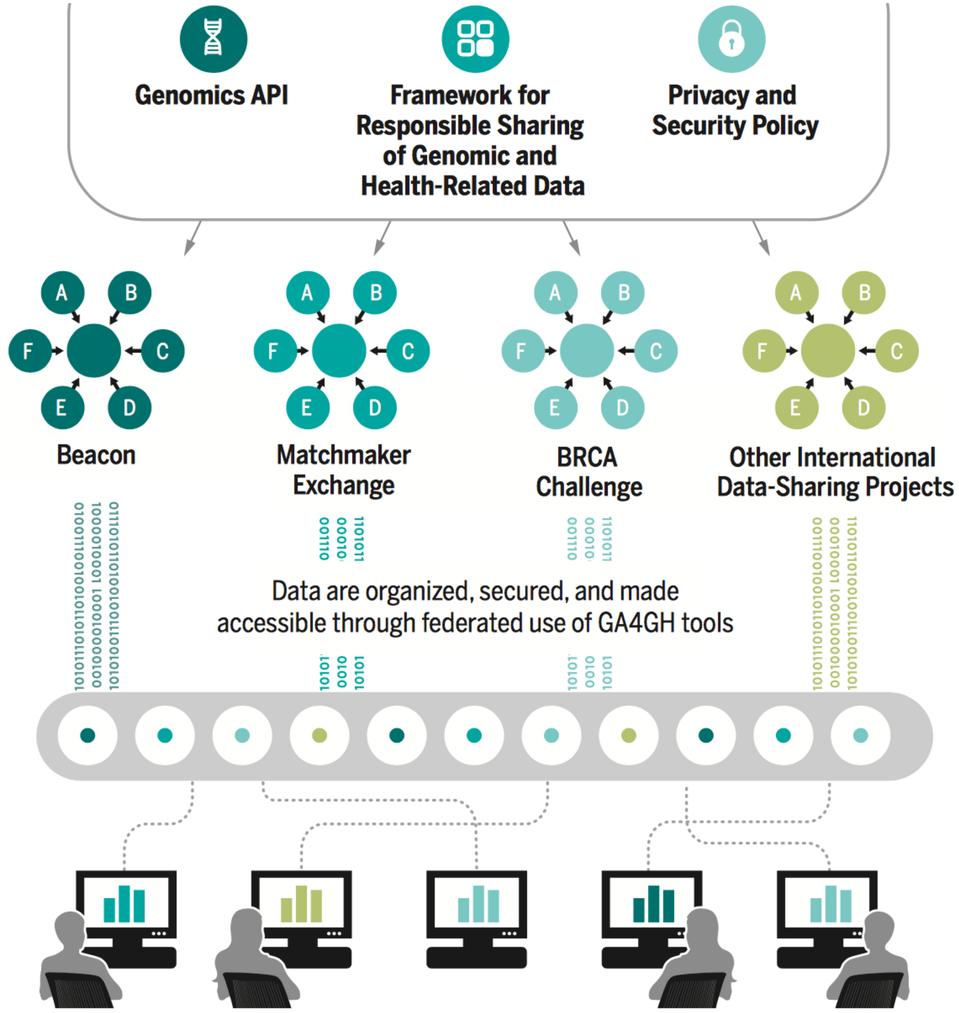
Source: Jordi Rambla (CRG Barcelona) 2016



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

GA4GH API promotes sharing

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



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.

Source: genomicsandhealth.org

Global Alliance for Genomics & Health

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

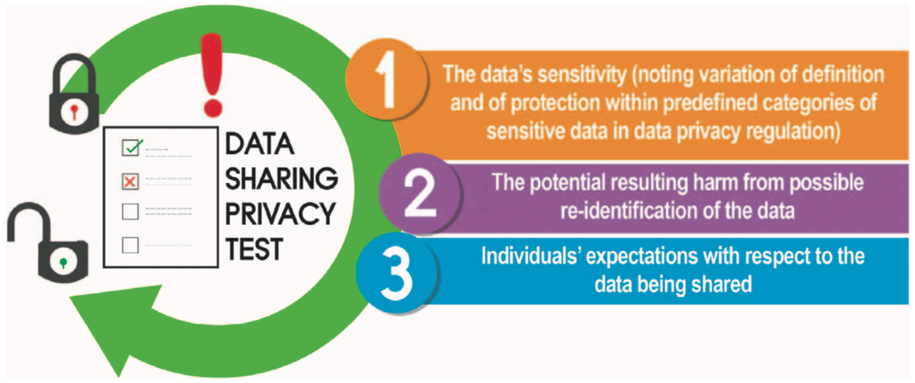


Figure 1. The three steps of a Data Sharing Privacy Test to distinguish degrees of data sensitivity within categories of data recognised as 'sensitive'.

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. (Source: SAMW)

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)

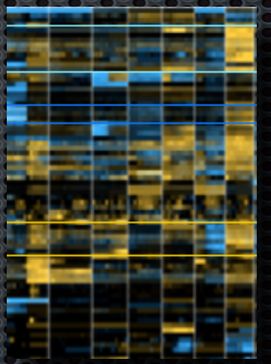
Cancer Genome Data & Beyond

Contributing our share

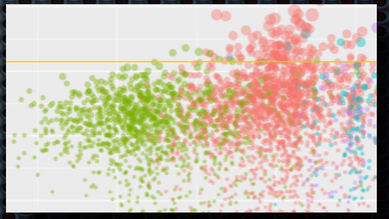
- Cancer genome data resources
- Software tools for data analysis & visualisation
- Parsing the cancer genome landscape: Patterns & targets
- GA4GH: Standards & Beacons
- Quantifying cancer genomic research
- The Swiss Personalized Health Network initiative
- Collaborations!

progenetix

arrayMap



Global Alliance for Genomics & Health



SAMW ASSM

akademien der wissenschaften schweiz

SIB Swiss Institute of Bioinformatics

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

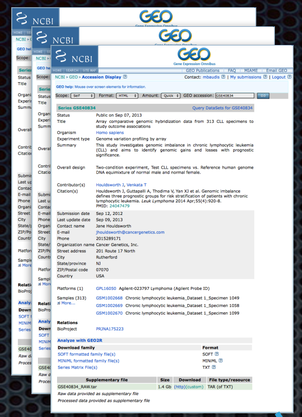
	progenetix	arrayMap
techniques	cCGH, aCGH, WES, WGS	aCGH (+?)
scope	sample (e.g. combination of several experiments)	experiment
content	**>31000** samples	**>60000** arrays
raw data presentation	no (link to sources if available)	yes (raw, log2, segmentation if available)
per sample re-analysis	no; supervised 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 ...)

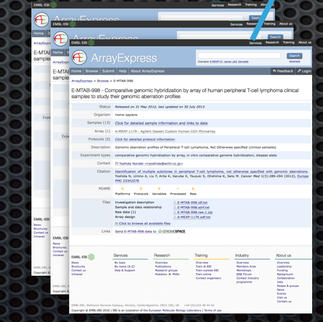
Michael Baudis | UZH | 2016-12-19

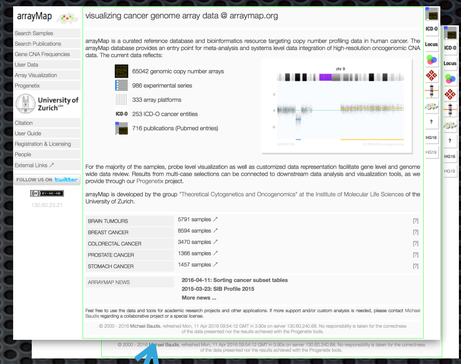
DATA PIPELINE

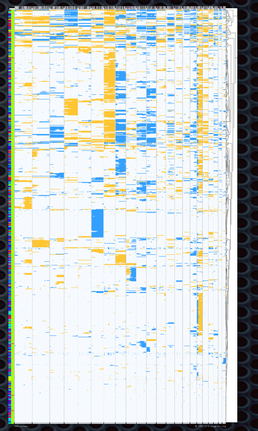
BIOCURATION

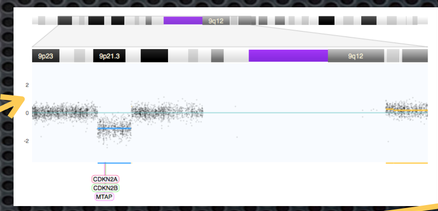
BIOINFORMATICS

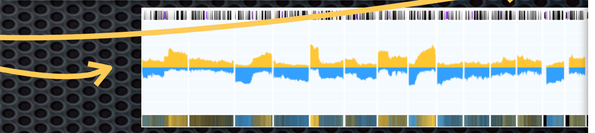


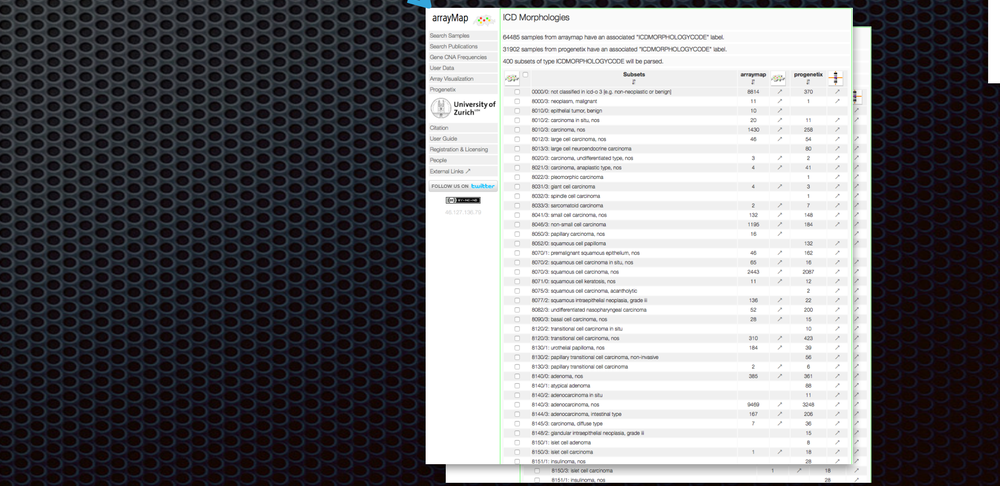


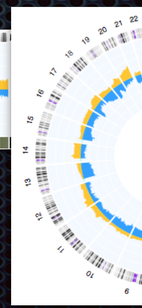












ARRAYMAP DATA "PIPELINE"

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

arrayMap

The largest resource for copy number variation data in cancer

arrayMap

- Search Samples
- Search Publications
- Gene CNA Frequencies
- User Data
- Array Visualization
- Progenetix

University of Zurich[UZH]

- Citation
- User Guide
- Registration & Licensing
- People
- External Links ↗

FOLLOW US ON twitter

(cc) BY-SA

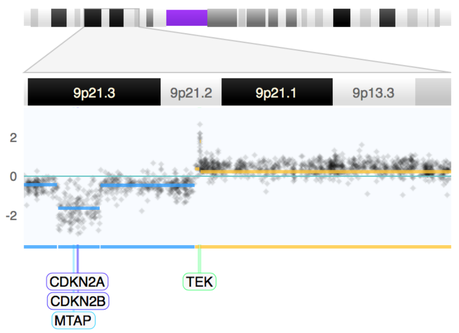
130.60.23.21

visualizing cancer genome array data @ arraymap.org

arrayMap is a curated reference database and bioinformatics resource targeting copy number profiling data in human cancer. The arrayMap database provides an entry point for meta-analysis and systems level data integration of high-resolution oncogenomic CNA data.

The current data reflects:

- 63060 genomic copy number arrays
- 763 experimental series
- 145 array platforms
- ICD-O 141 ICD-O cancer entities
- 554 publications (Pubmed entries)



Genomic copy number imbalances on chromosome 9 in a case of Glioblastoma (**GSM491153**), indcating, among others, a homozygous deletion involving CDKN2A/B.

ICD-O
Locus
HG18
HG19

For the majority of the samples, probe level visualization as well as customized data representation facilitate gene level and genome wide data review. Results from multi-case selections can be connected to downstream data analysis and visualization tools, as we provide through our Progenetix project.

arrayMap is developed by the group "Theoretical Cytogenetics and Oncogenomics" at the Institute of Molecular Life Sciences of the University of Zurich.

BRAIN TUMOURS	5653 samples ↗	[?]
BREAST CANCER	8329 samples ↗	[?]
COLORECTAL CANCER	3238 samples ↗	[?]
PROSTATE CANCER	991 samples ↗	[?]
STOMACH CANCER	1062 samples ↗	[?]

ARRAYMAP NEWS

2016-08-03: SVG graphics
2016-05-17: Transitioning to Europe PMC
More news ...

Feel free to use the data and tools for academic research projects and other applications. If more support and/or custom analysis is needed, please contact Michael Baudis regarding a collaborative project or a special license.

© 2000 - 2016 Michael Baudis, refreshed Mon, 19 Sep 2016 10:20:09 GMT in 6.87s on server 130.60.240.68. No responsibility is taken for the correctness of the data presented nor the results achieved with the Progenetix tools.

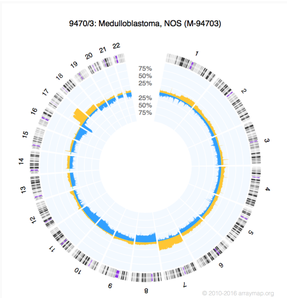
ICD Morphologies

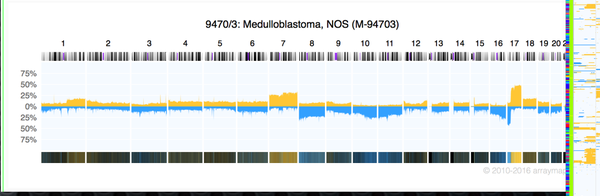
2021 samples from arraymap have an associated "ICDMORPHOLOGYCODE" label.

9470/3: Medulloblastoma, NOS (M-94703)

Synonyms

- Medulloblastoma, NOS
- Melanotic medulloblastoma

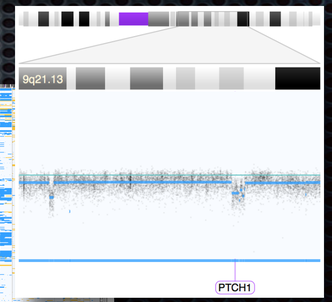




FIND CNAS BY GENE OR REGION: TP53 | [ERBB2] 17:35097862-35138441:1 [?]
REGION SIZE | MAX COVERAGE (KB): 0 kb * 5000 250000 kb [?]
CLINICAL DATA: no followup required
CITY: 20 km [?]
Query Database

1949 of 65042 cases matched the selection criteria.

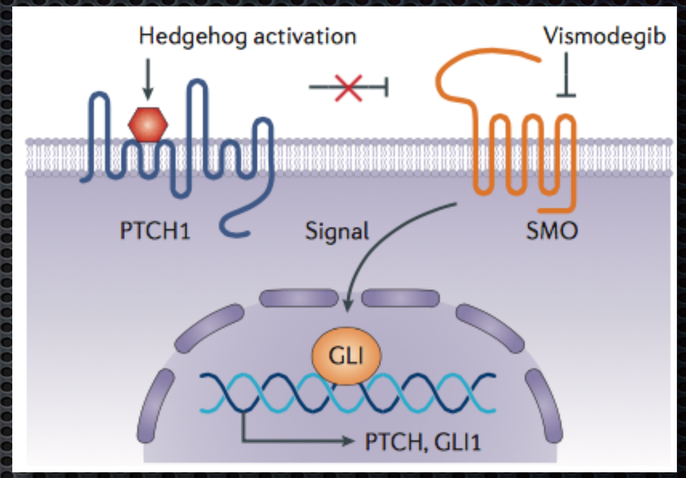
SUBSET	PERCENT IN SUBSET
8507/3: Invasive micropapillary carcinoma (13/39)	33.3
C692: retina (14/82)	17.1
8260/3: Papillary adenocarcinoma, NOS (11/65)	16.9
8500/3: Invasive carcinoma of no special type (1201/8188)	14.7
8560/3: Adenosquamous carcinoma (3/21)	14.3
Carcinomas: breast ca. (1254/8837)	14.2
C50: breast (1254/8929)	14.0
8500/2: Ductal carcinoma in situ, NOS (25/225)	11.1
C32: larynx (3/29)	10.3
8010/2: Carcinoma in situ, NOS (2/20)	10.0
C187: sigmoid incl. rectosigmoid junction (13/140)	9.3
8480/3: Mucinous adenocarcinoma (12/132)	9.1
8522/3: Infiltrating duct and lobular carcinoma (4/44)	9.1
8460/3: Micropapillary serous carcinoma [C56.9] (32/513)	6.2
8130/1: Urothelial papilloma, NOS (11/184)	6.0
C680: other urinary organs (11/184)	6.0
C54: corpus uteri (19/330)	5.8
8441/3: Serous adenocarcinoma, NOS (31/542)	5.7
Carcinomas: esophagus ca. (32/571)	5.6
Carcinomas: gastric ca. (80/1492)	5.4



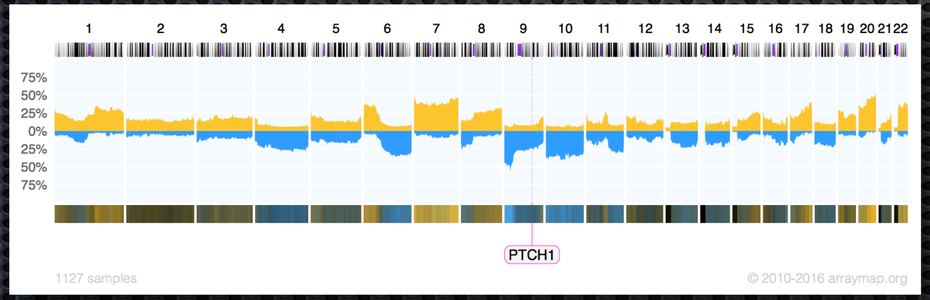
UID	SERIESID	PMID	ICDMORPHOLOGYCODE	ICDTOPOGRAPHYCODE
GSM1000061	GSE36942	23457519	8070/3	C10
GSM1000062	GSE36942	23457519	8070/3	C10
GSM1001316	GSE40777	23571474	8070/3	C53
GSM1001317	GSE40777	23571474	8010/3	C34
GSM1001318	GSE40777	23571474	8070/3	C09
GSM1001319	GSE40777	23571474	8010/3	C34
GSM1002668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

Rare Events & Hidden Therapeutic Options?
Example: PTCH1 deletions in malignant melanomas

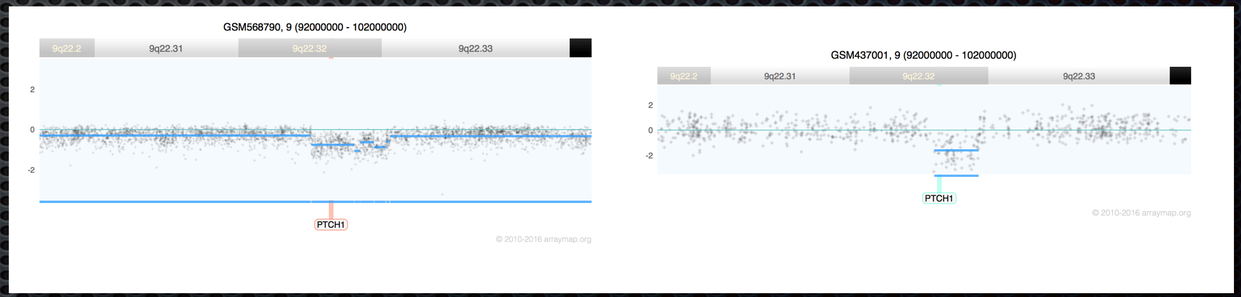
- PTCH1 is a actionable tumor suppressor gene, which has been demonstrated in e.g. basaliomas and medulloblastomas
- analysis of 1127 samples from 26 different publications could identify **focal** deletions in 4 samples
- a current project addresses the focal involvement of all mapped genes, in >50'000 cancer genome profiles



In its normal function, PTCH1 is a tumor suppressor gene in the sonic hedgehog pathway and inhibits SMO driven transcriptional activation. A loss of PTCH1 function (mutation, deletion) can be mitigated through drugs antagonistic to SMO activation.



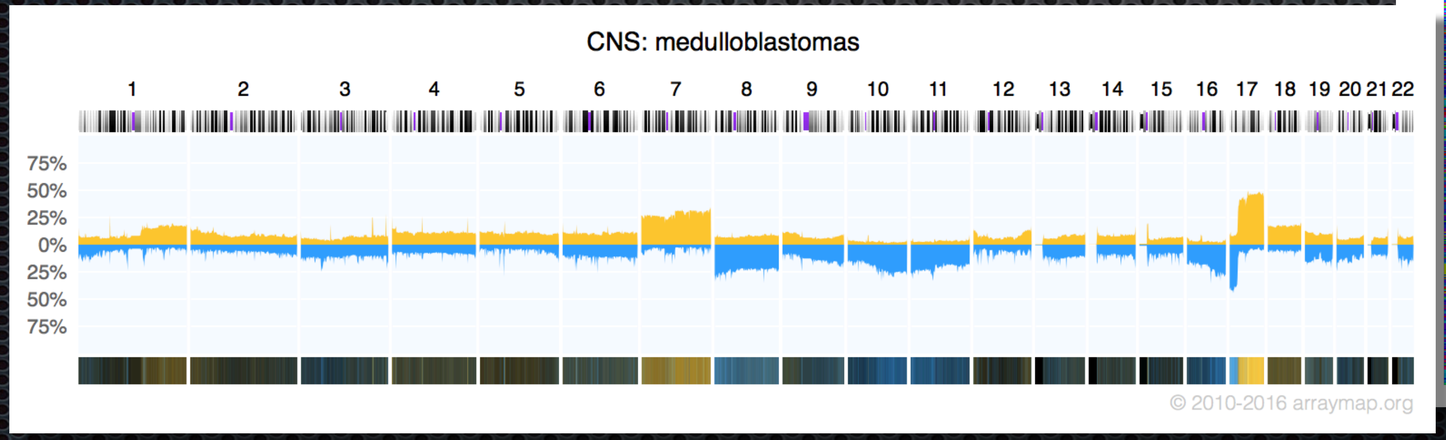
Summary of somatic copy number aberrations from the analysis of 1127 genome profiles of malignant melanomas, collected in our arraymap.org cancer genome resource. While PTCH1 does not represent a deletion hotspot, the genomic locus is part of larger deletions in ~25% of melanoma samples.

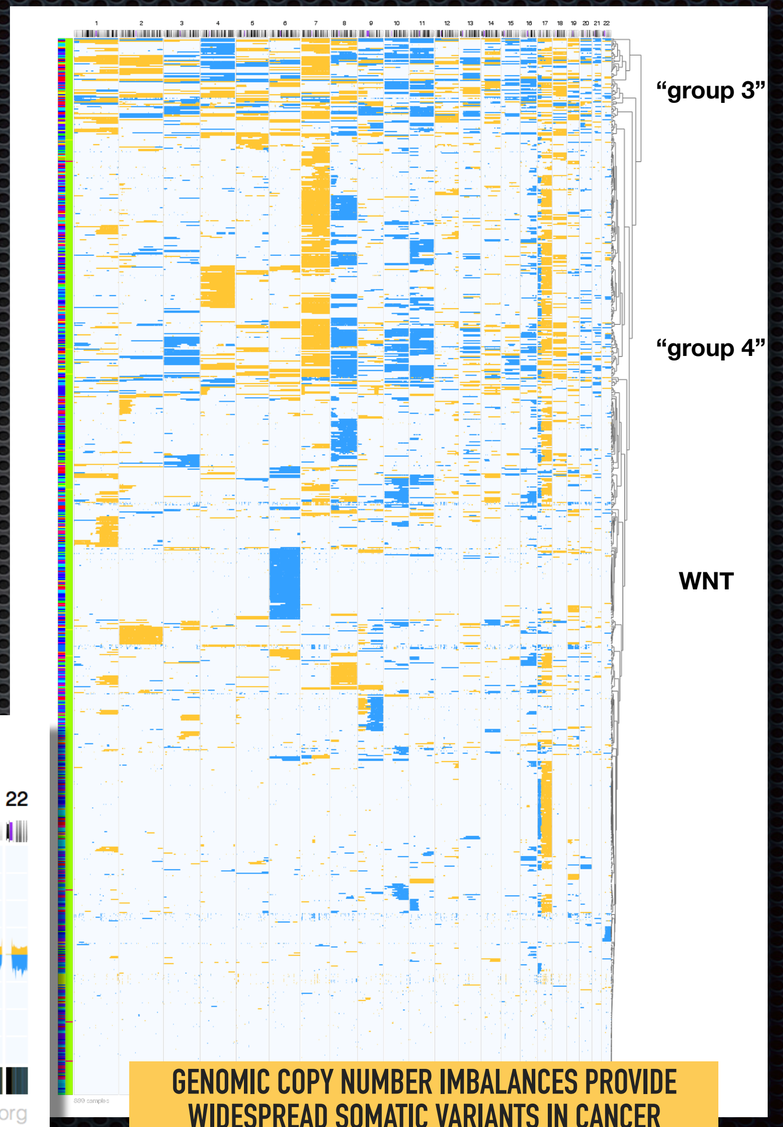


Examples of focal / homozygous PTCH1 deletions detected in the analysis of 1127 genomic array datasets. Focal somatic imbalance events are considered an indicator for oncogenic involvement of the affected target genes.

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?





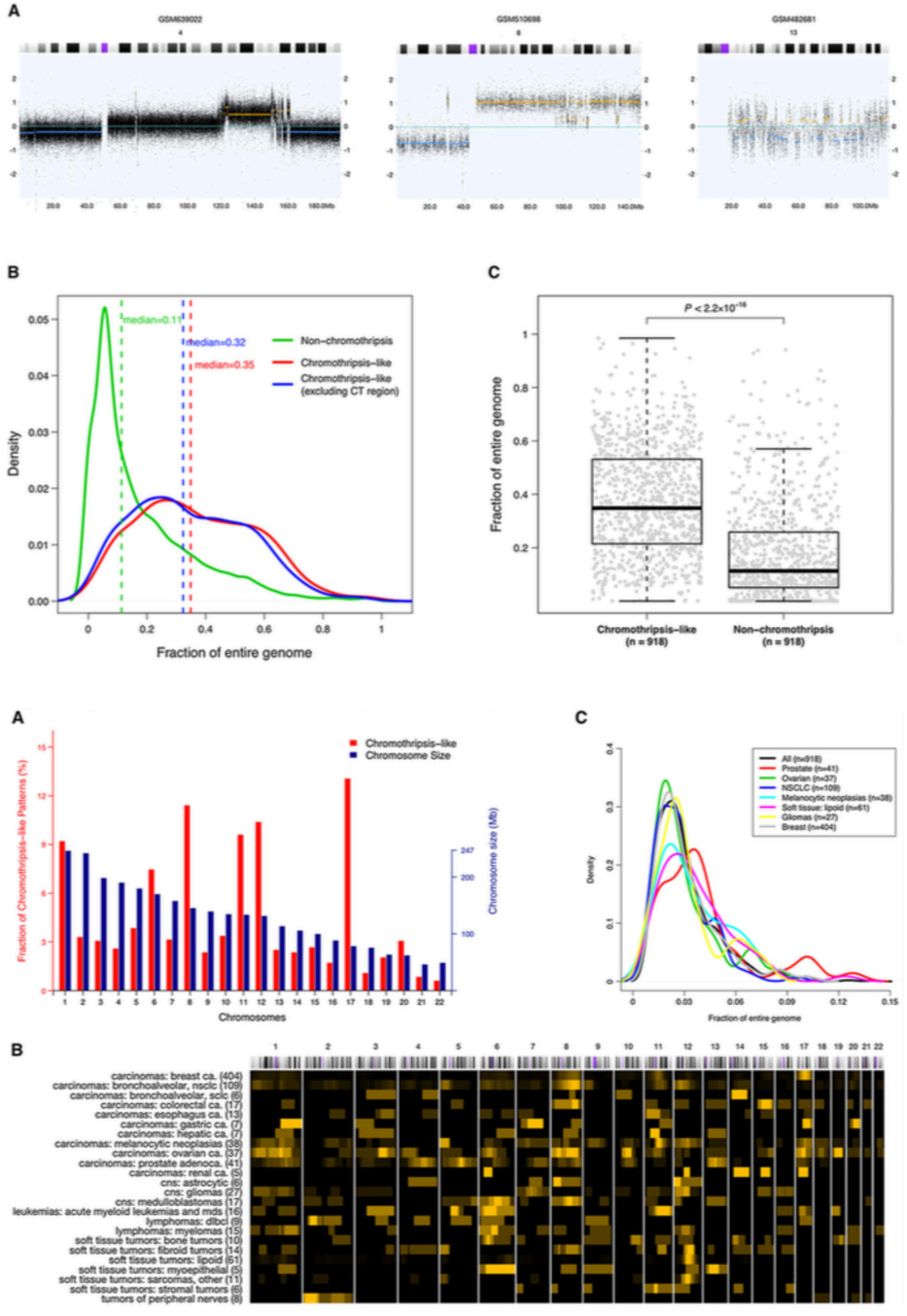
GENOMIC COPY NUMBER IMBALANCES PROVIDE WIDESPREAD SOMATIC VARIANTS IN CANCER

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 II

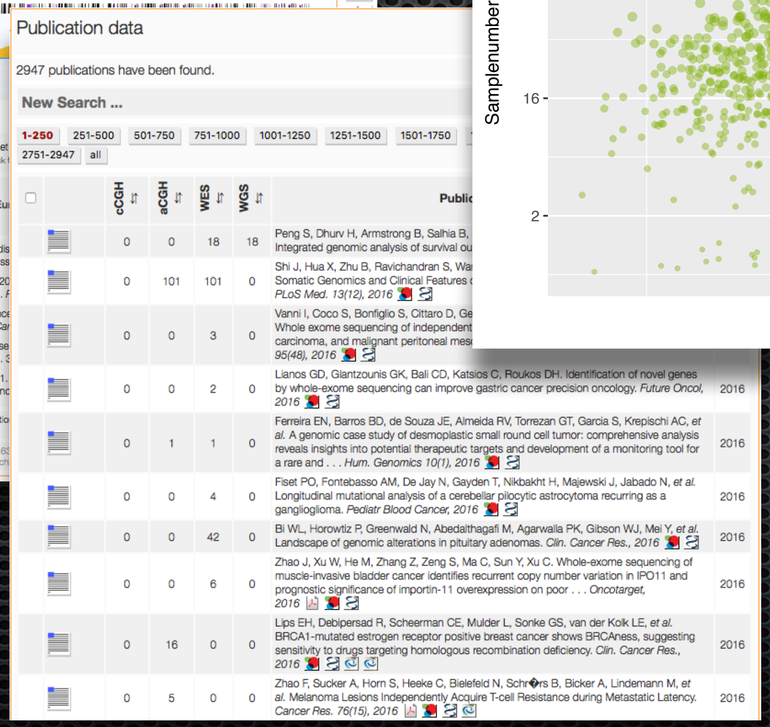
Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens

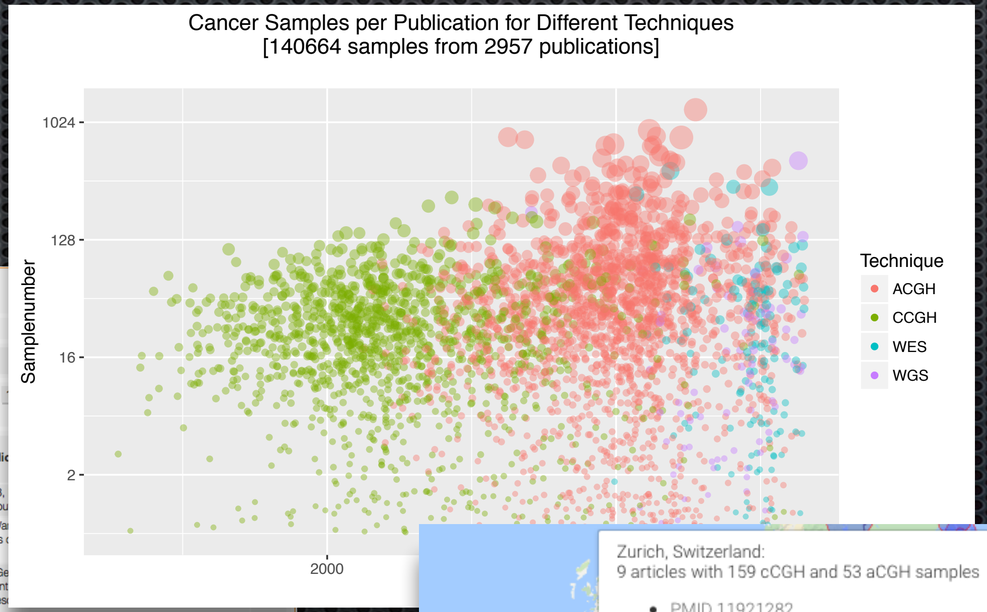
- “Chromothripsis” is a recently described mechanism in which hundreds of genomic fragments are re-assembled in a single rearrangement event, which may oblivate gradual accumulation of mutations
- many cancer genome profiles haven been attributed to chromothripsis
- our analysis of >22’000 cancer genome profiles pointed to heterogeneity and relation to predominant overall genome instability in cancers with such chromothripsis-like genome patterns

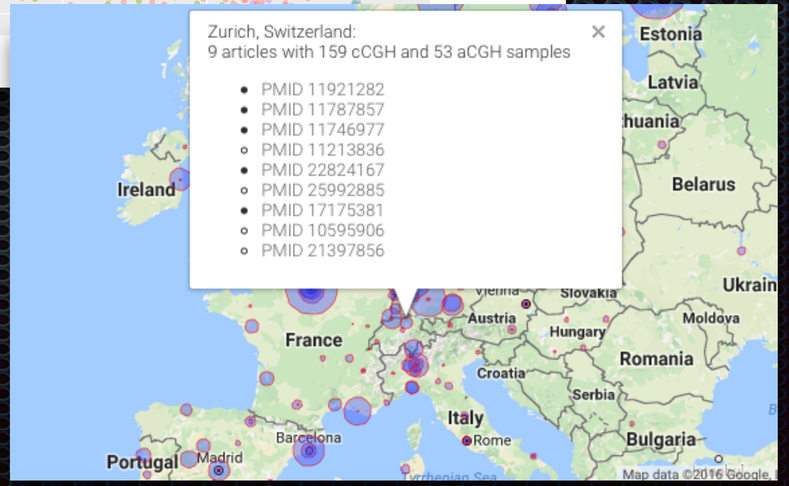


Cai et al. BMC Genomics 2014, 15:82

Progenetix: Cancer Genome Profiles, Article Metrics, Epistemology, **Resource Hub**

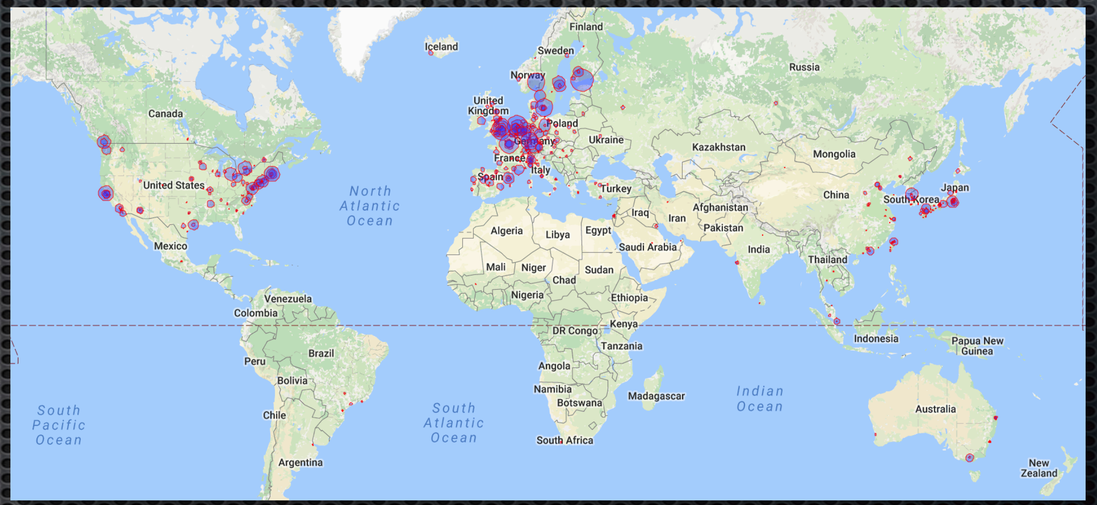


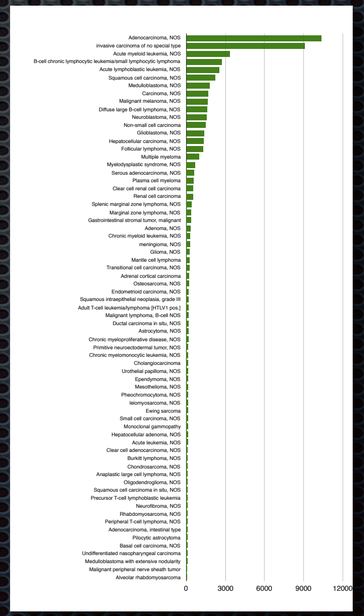




Bias in Ascertainment / Background / Environment in Cancer Genome Studies

- 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



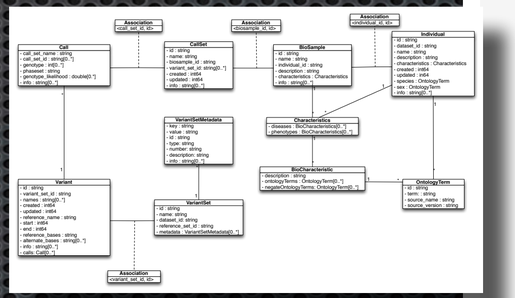


Geographic distribution of >140'000 cancer genome profiles reported in the literature. The numbers are derived from the 2947 publications registered in the Progenetix database.

progenetix

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

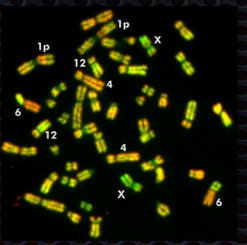
elixir

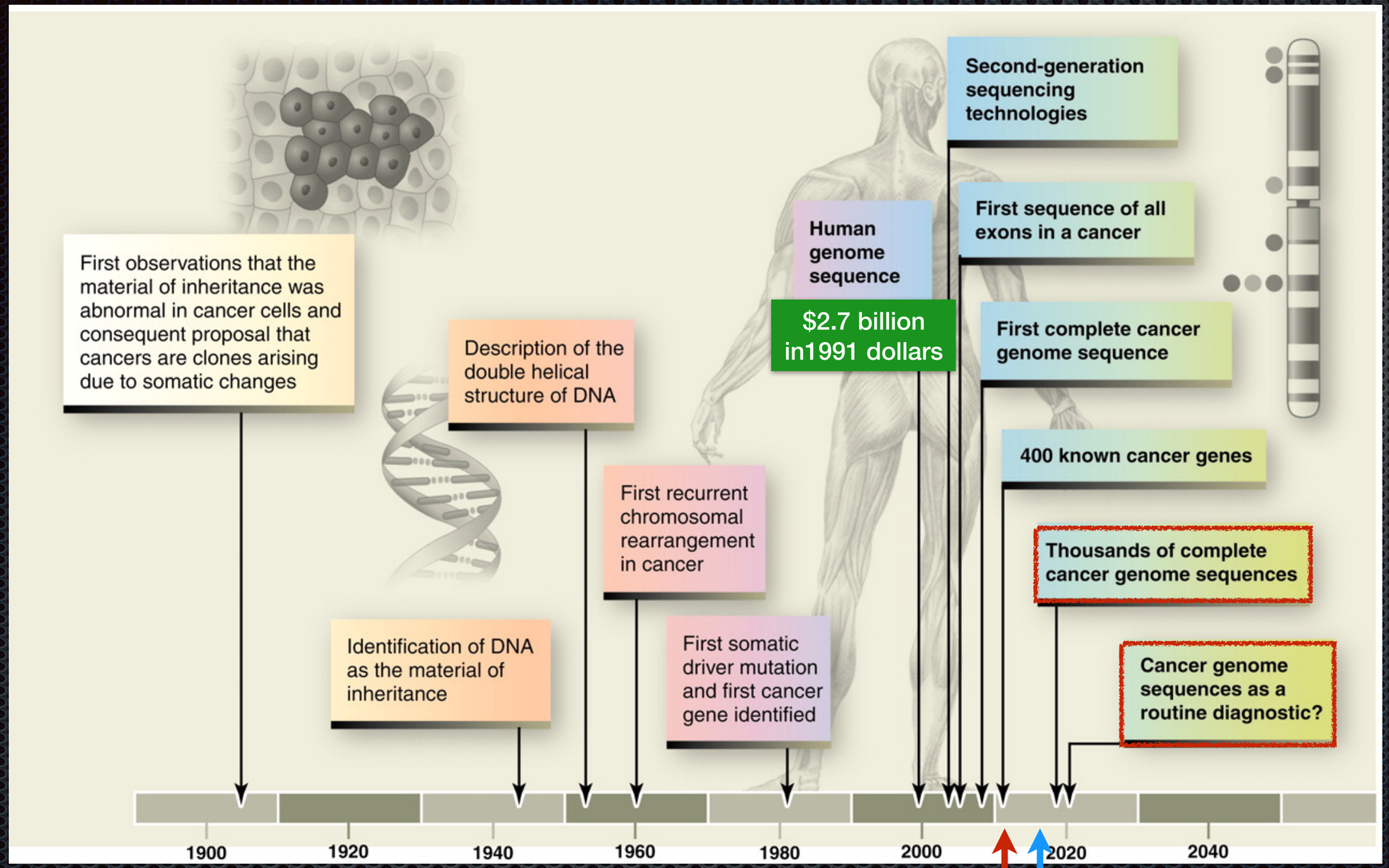
SIB Swiss Institute of Bioinformatics


```
{
        "_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"),
}
```

Cancer Genome Data: Where Do We Need To Go

- balancing clinical medicine (**panel** sequencing of **actionable** cancer targets) and the necessary knowledge generation (**complete genomes**, multi"omes", genetic background)
- "genetic **awareness**" and regulatory **security**
- vastly increasing **data curation efforts** to make best use of existing data
- data **sharing frameworks** & federated analysis
- everywhere in the world...





Not yet practical for full human genomes, but real product! (OxfordNanopore)

Michael R. Stratton. Exploring the Genomes of Cancer Cells: Progress and Promise. Science (2011)

Th. Boveri

Nowell/Hungerford
J. Rowley

2016

BAUDISGROUP @ UZH

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

SIB

HEINZ STOCKINGER
SÉVERINE DUVAUD
VASSILIOS IOANNIDIS
DANIEL TEIXEIRA

THOMAS EGGERMANN
ROSA NOGUERA
REINER SIEBERT
CAIUS SOLOVAN

... MY COLLEAGUES AT THE INSTITUTE OF MOLECULAR LIFE SCIENCES AND OTHER MEMBERS OF THE UZH

University of Zurich[UZH]

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
GUNNAR RÄTSCH
ELEANOR STANLEY
DAVID STEINBERG

ELIXIR & CRG

JORDI RAMBLA DE ARGILA
SABELA DE LA TORRE PERNAS
SUSANNA REPO
SERENA SCOLLEN