ETH zürich

Data Sharing Strategies & Benefits in ICGC, TCGA and BRCA Challenge

Gunnar Rätsch

Biomedical Informatics Group

@gxr

#PrecisionMedicine #Cancer #Genomics #ClinicalData #SPHN

BIOMEDICAL INFORMATICS

DINFK

UniversityHospital Zurich

Memorial Sloan-Kettering Cancer Center

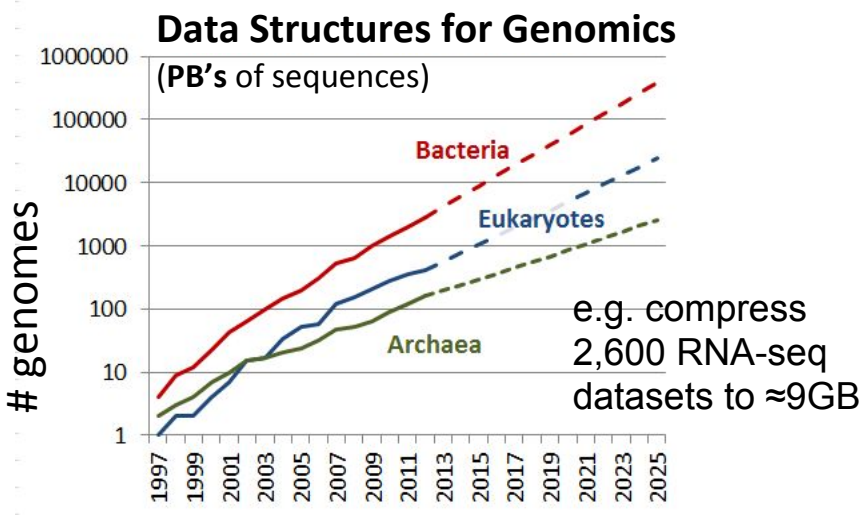
20 YEARS SIB

Biomedical Informatics Lab: Data Science for Biomedical Applications

A

Data Structures for Genomics

(**PB's** of sequences)



e.g. compress 2,600 RNA-seq datasets to ≈9GB

B

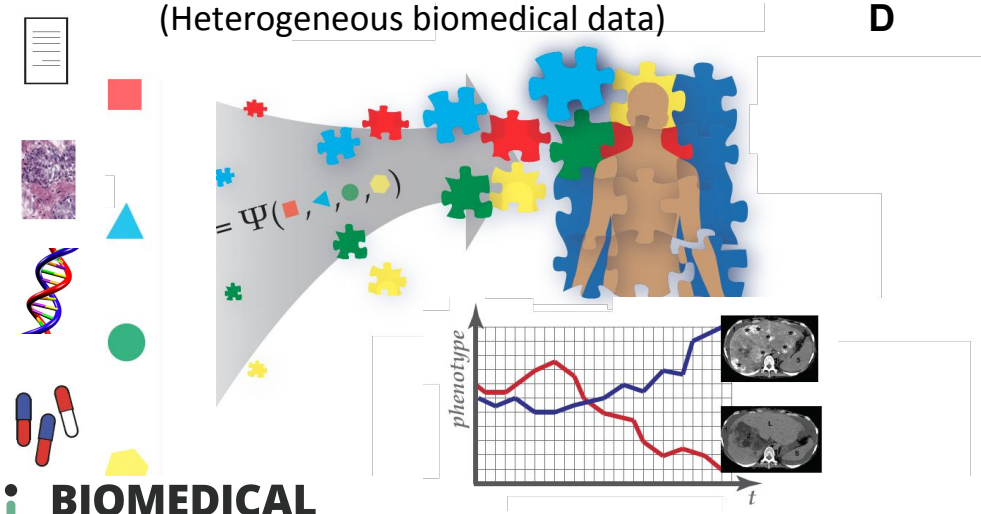
Data-rich Intensive Care Unit

(collaboration with Intensive Care Unit with ≈60.000 patients over last 10 years, **≈500 GB**)

C

Patient/Disease Modeling

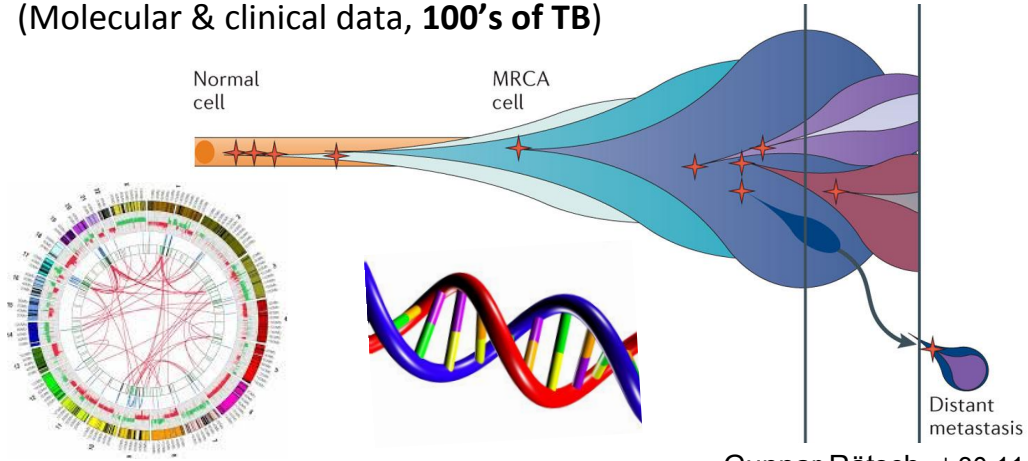
(Heterogeneous biomedical data)



D

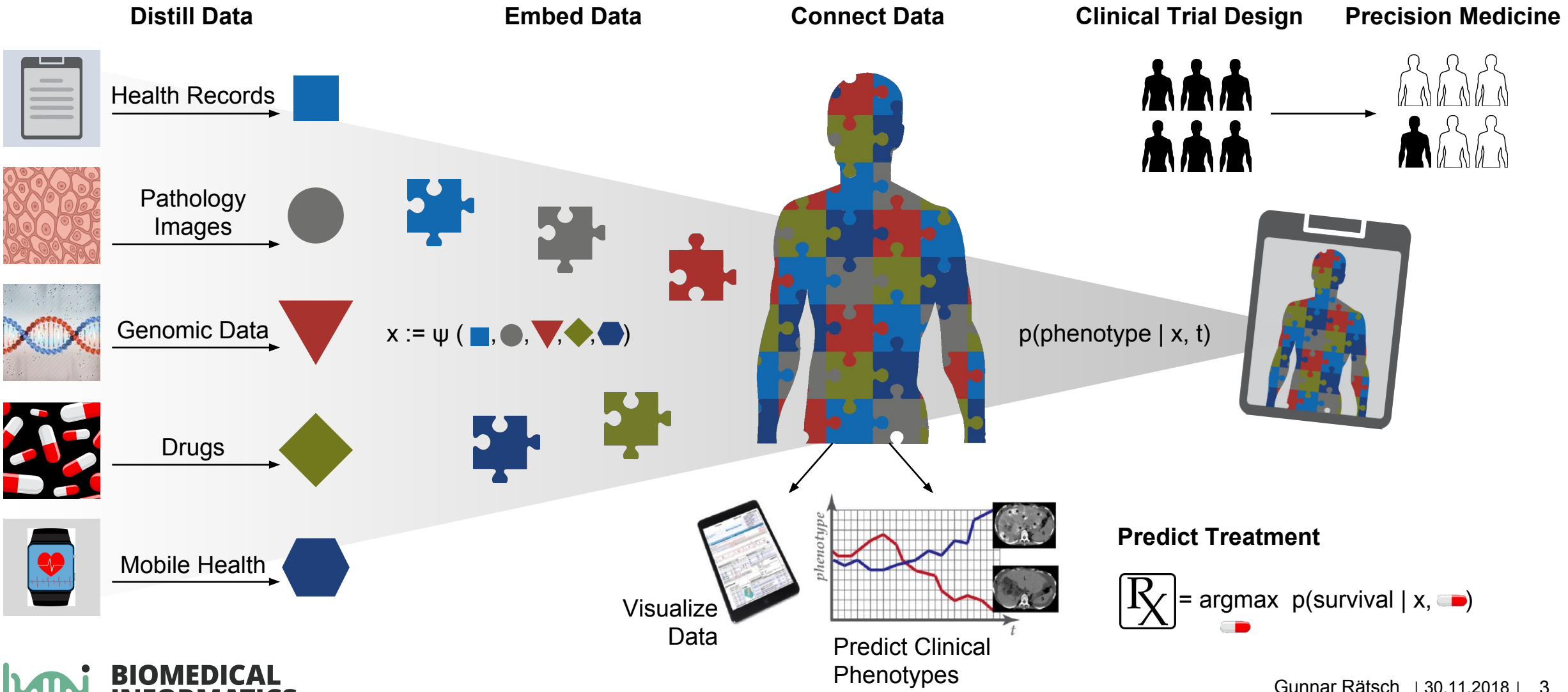
Cancer Genomics/Biology

(Molecular & clinical data, **100's of TB**)



BIOMEDICAL INFORMATICS

Towards Comprehensive Patient Models for Precision Medicine



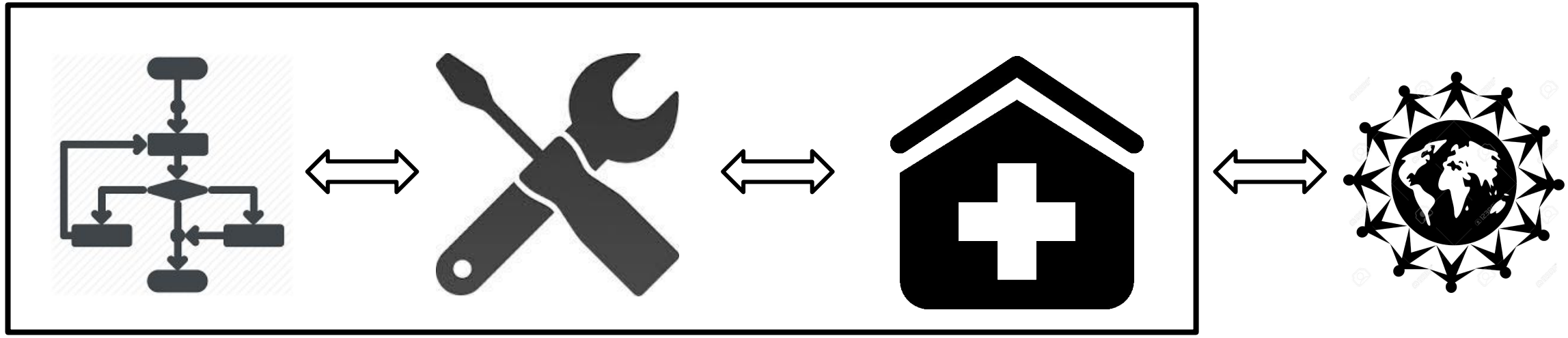
Data Science Research Challenges

Challenge 1: Develop novel data science approaches for medical data

Challenge 2: Provide analysis tools for the community

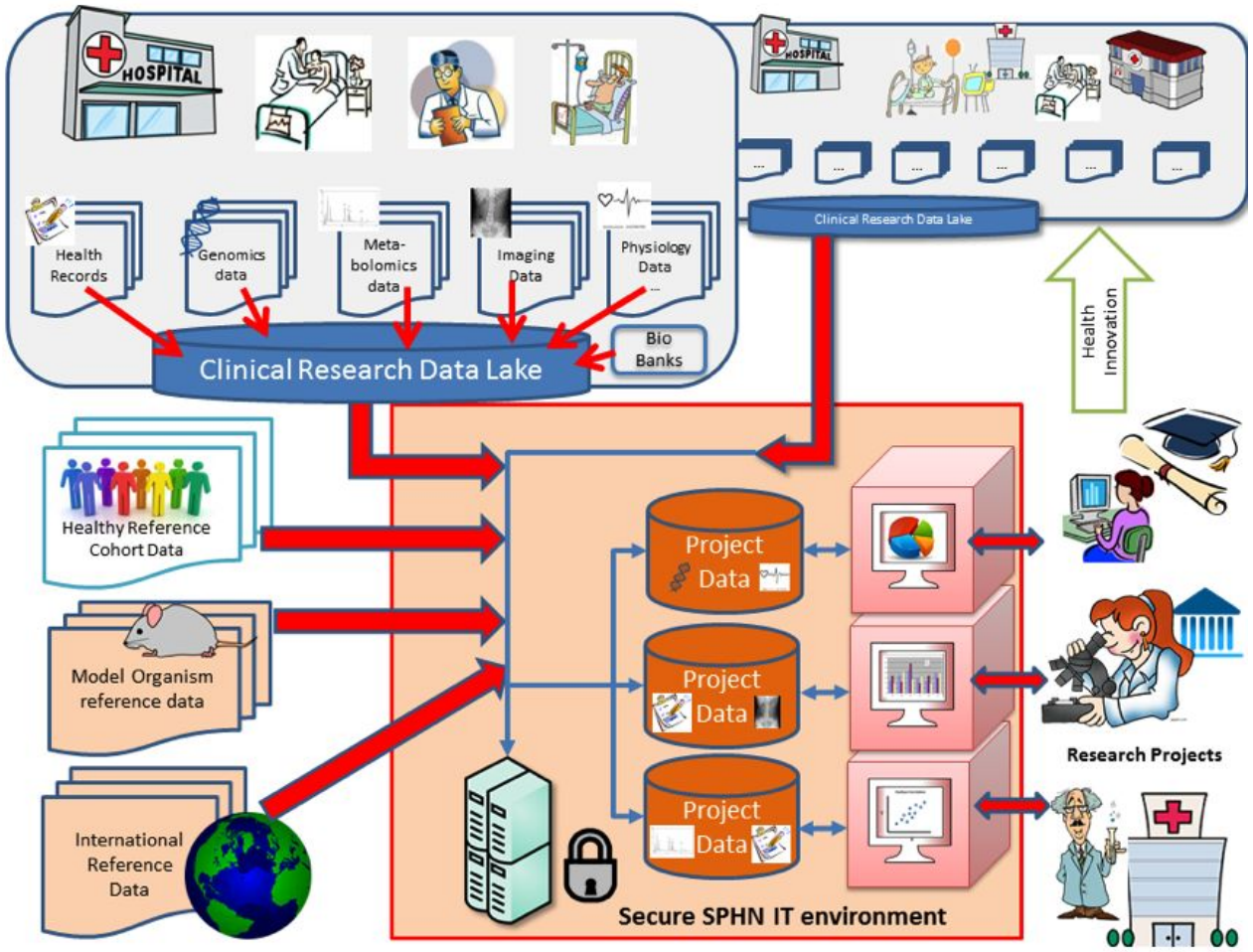
Challenge 3: Solve important biomedical problems through collaborations

Challenge 4: Create an environment which allows us to work on the above



Source: Center: Google icons search

a) Large-scale Initiatives



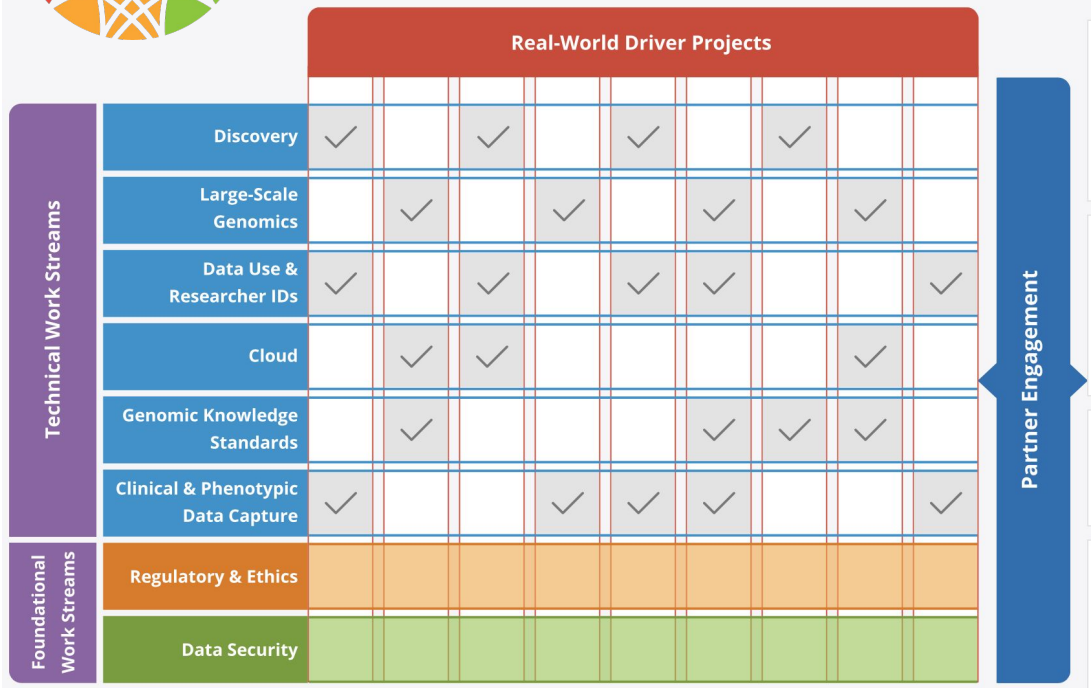
Source: Courtesy of Torsten Schwede

BIOMEDICAL INFORMATICS

Swiss Personalized Health Initiative
Aim: Data interoperability between hospitals and researchers

b) Data Sharing Standards

Global Alliance for Genomics & Health



GA4GH 6th Plenary Meeting

Source: GA4GH

3 – 5 October 2018, Basel, Switzerland

http://ga4gh.org

Large-scale Cancer Genome Projects (TCGA, ICGC, ICGC-ARGO, ...)

- Two **major cancer genome projects** started >10 years ago
  - Aim: collect and profile tumor & normal tissue samples
    - Whole Genome Sequencing (WGS), Exome (WXS), RNA-seq, Micro RNAs, some Mass Spec,
      pathology slides
    - All publicly/controlled accessible
  - TCGA: organized by US National Cancer Institute, total ≈12'000 donors, ≈1PB
  - ICGC: international effort, total ≈25'000 donors, ≈2'500 samples with WGS, ≈1PB
    - ICGC-ARGO: 10x ICGC + detailed medical records
- Hundreds of groups generating and analysing the data, thousands of papers written about it
- Anything left to be done?

Example: International Cancer Genome Consortium

≈800 scientists

International Cancer Genome Consortium

≈200 scientists

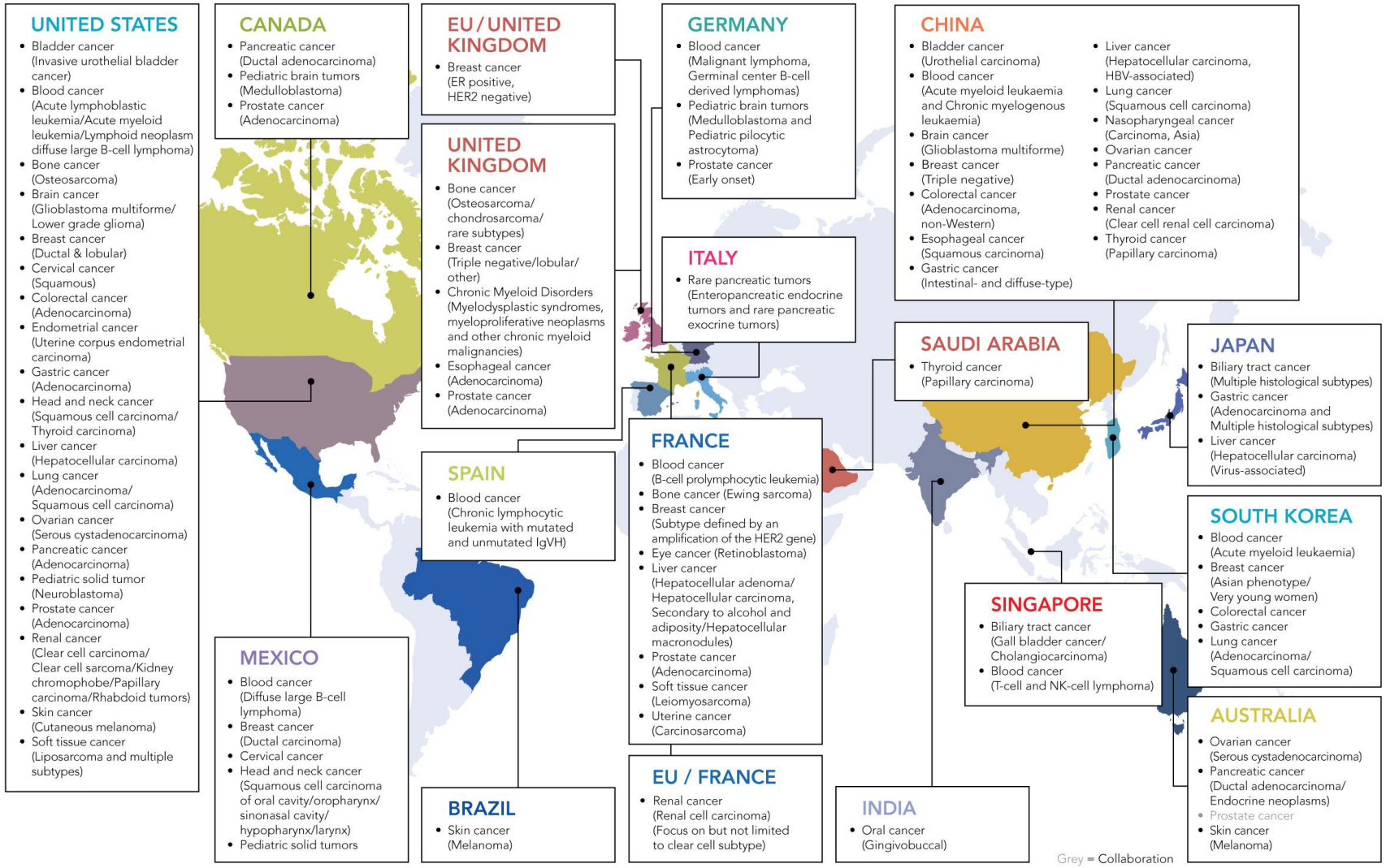
PCAWG PanCancer Analysis OF WHOLE GENOMES

≈15 working groups

PCAWG Workgroup 3 Transcriptome/Genome Analysis Group
WGL: Brooks, Brazma, Rätsch

≈50 scientists

≈40 PCAWG Papers have been or will be submitted to the Nature family in 2018



Source: ICGC

BIOMEDICAL INFORMATICS

Collaborative Project Principles/Goals

"It is amazing what you can accomplish if you do not care who gets the credit."
— Harry S. Truman

- Be part of something great
- Support junior faculty and students
- Change clinical and research practices
- Have fun working together

Important: Clear rules of collaboration/publication in consortium

Courtesy of Mark Rubin

Project 1: Integrate Diverse Transcriptomic Alterations to Identify Cancer-relevant Genes and Signatures

PanCancer Analysis of Whole Genomes and Transcriptomes Working Group (PCAWG-3), PCAWG Consortium, Kjong-Van Lehmann[1,2], André Kahles[1,2], Alvis Brazma[3], Angela N. Brooks[4], Claudia Calabrese[3], Nuno A. Fonseca[3], Jonathan Göke[5], Roland F Schwarz[3,6], Gunnar Rätsch[1,2], Zemin Zhang[7,8]

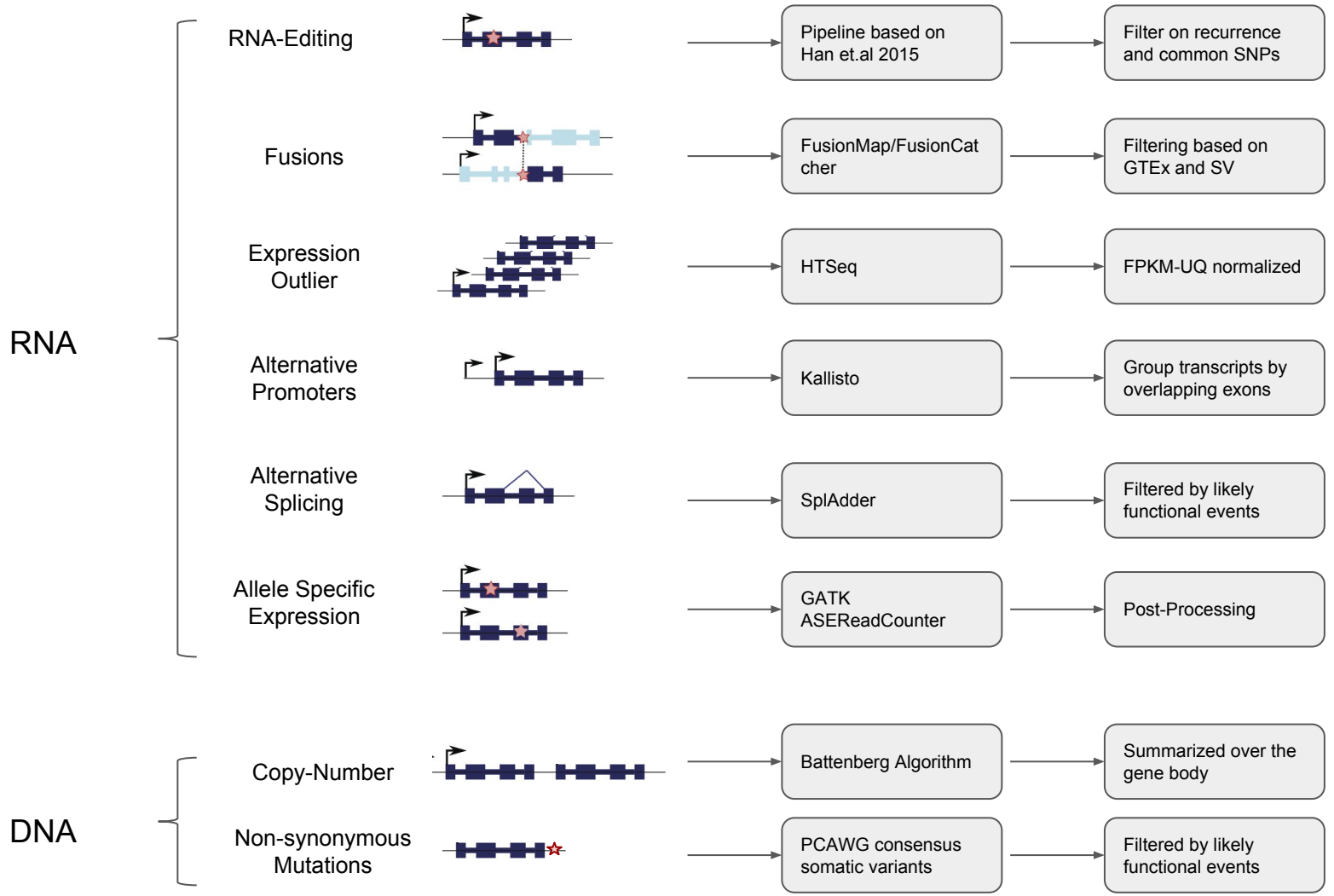
[1]ETH Zürich, Computer Science Dept, Universitätsstrasse 6, 8092 Zürich, Switzerland;
[2]Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York 10065, USA;
[3]European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK;
[4]University of California, Santa Cruz, CA 95060; [4]BaylorCollege of Medicine, Houston, TX, USA; [5]Duke-NUS Graduate Medical School, 8 College Road, Singapore 169857, Singapore; [6]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany; [7]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, 100871, China; [8]Genome Institute of Singapore, 60 Biopolis Street, Genome #02-01, Singapore 138672, Singapore;



PCAWG-3 Marker Manuscript
https://doi.org/10.1101/183889

BIOMEDICAL INFORMATICS

Collaborative Analysis of Multiple Alteration Types



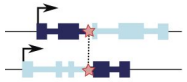
Needed collaboration of 7 top research groups

Source: PCAWG-3 Working group

Summarize Alterations for a Gene

RNA-Editing

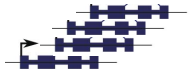
Fusions



Non-synonymous Mutations

Copy-Number

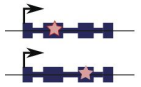
Expression Outlier



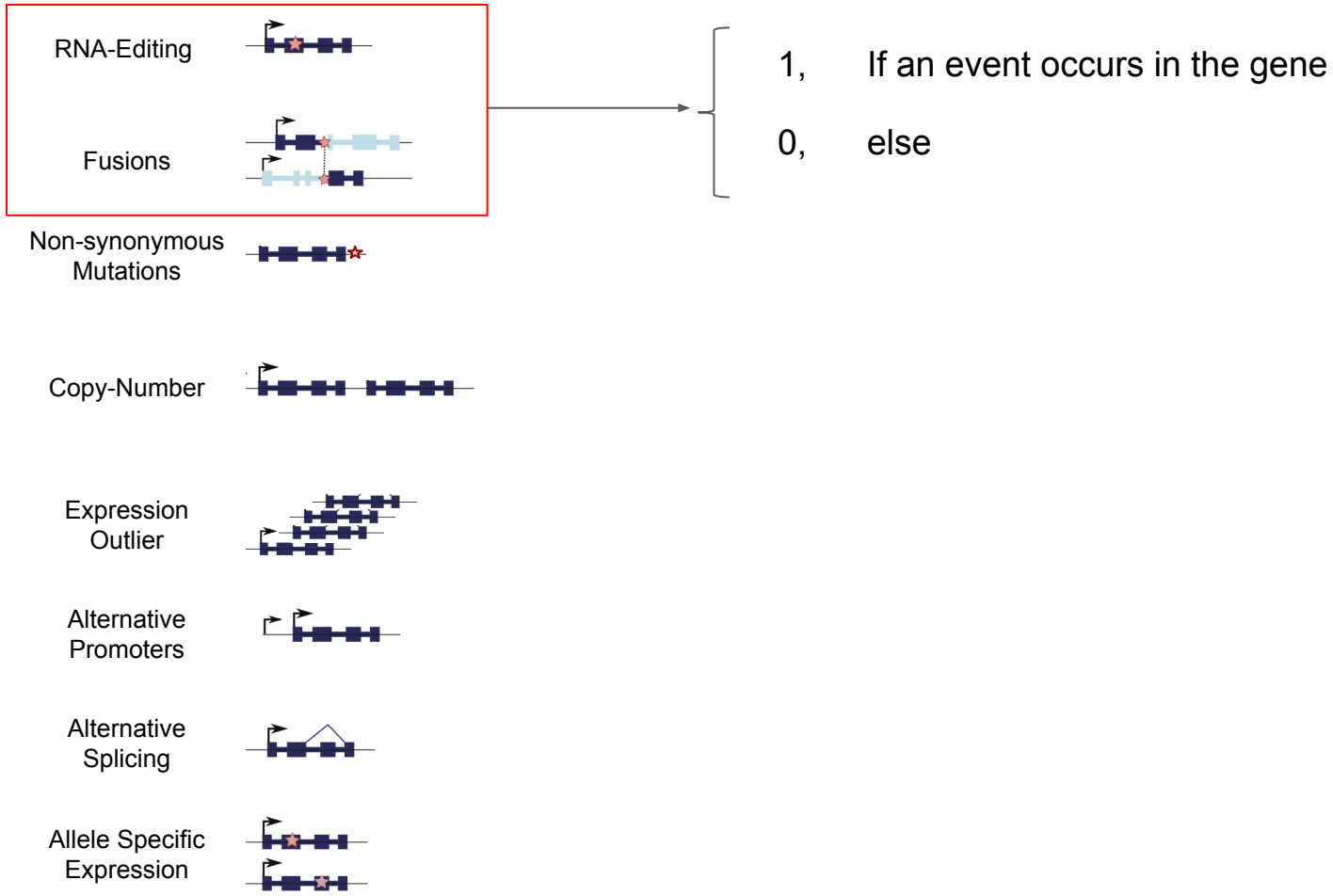
Alternative Promoters

Alternative Splicing

Allele Specific Expression



Summarize Alterations for a Gene

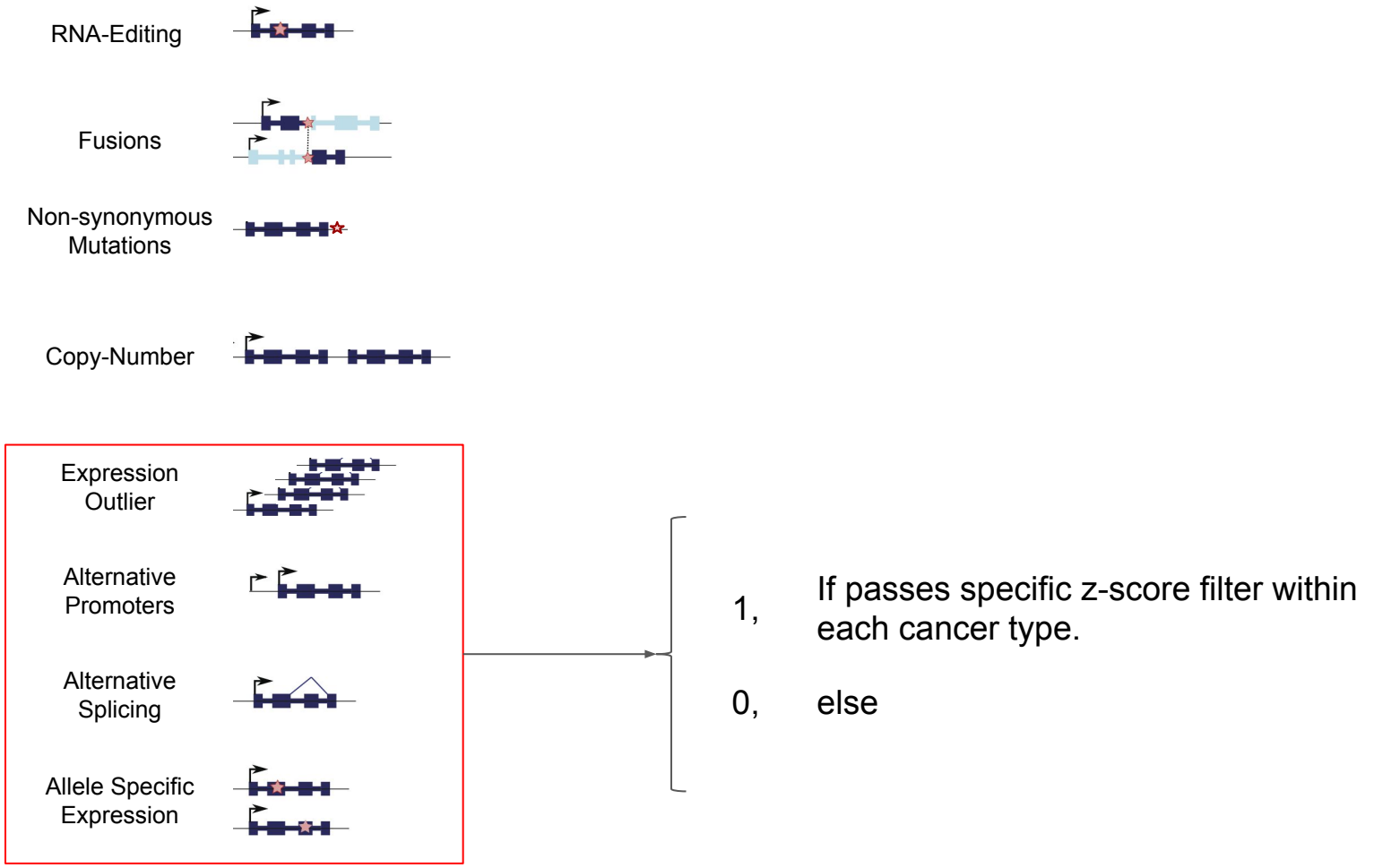


Courtesy of Natalie Davidson

Summarize Alterations for a Gene

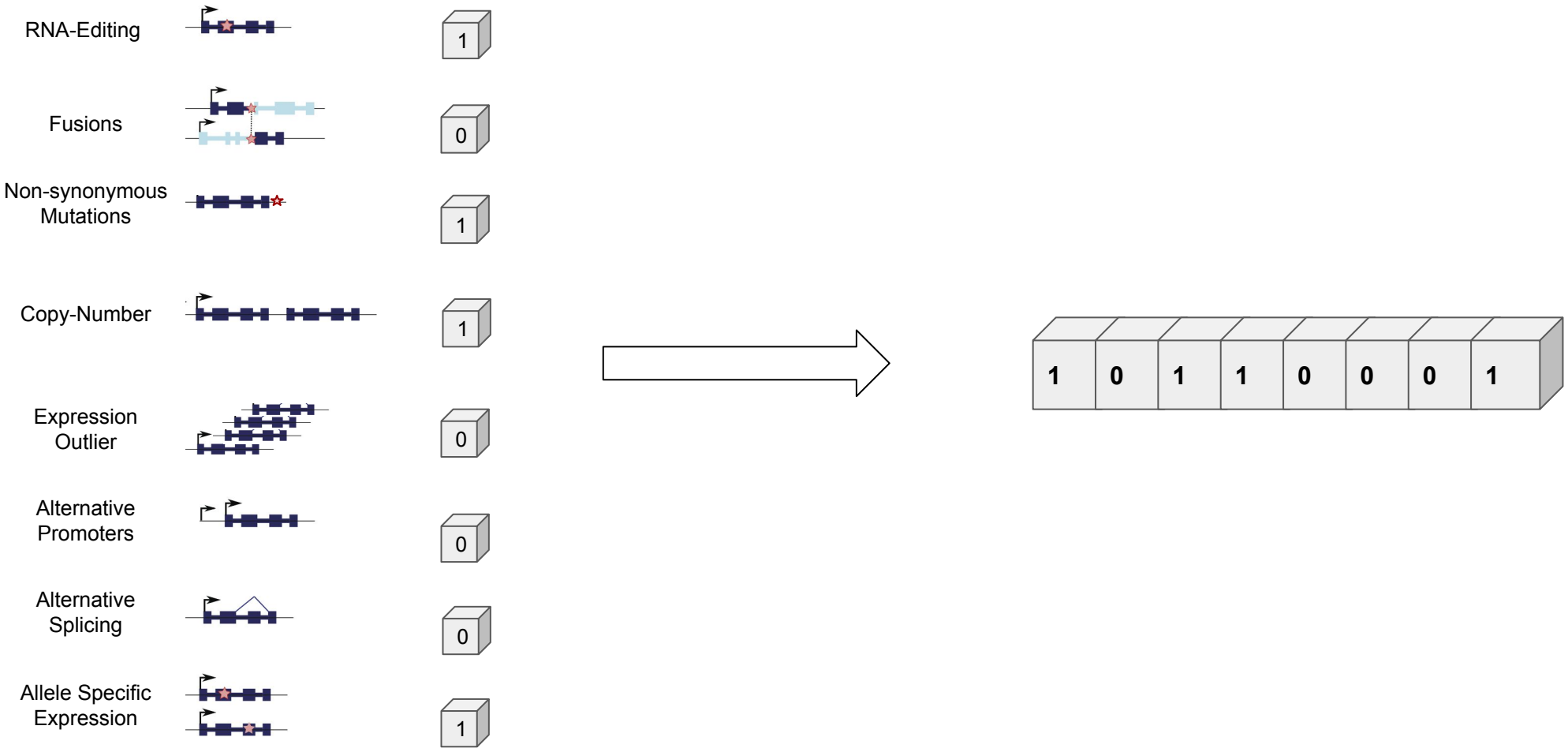
Courtesy of Natalie Davidson

Summarize Alterations for a Gene



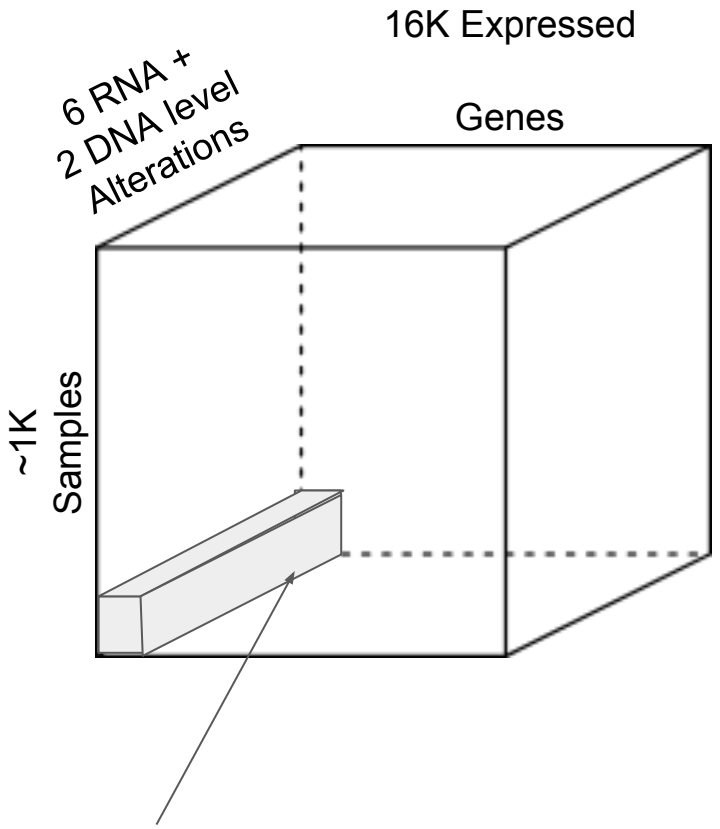
Courtesy of Natalie Davidson

Summarize Alterations for a Gene



Courtesy of Natalie Davidson

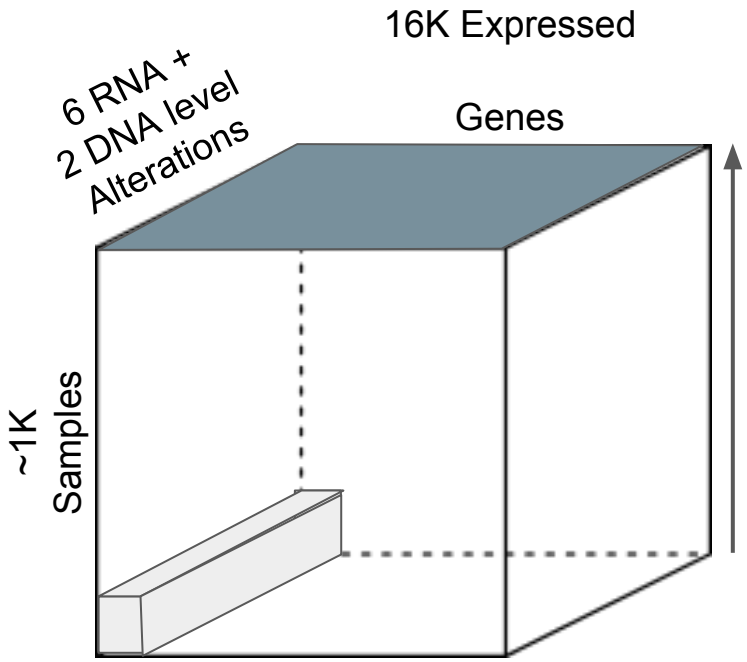
Summarize Alterations Across Genes or Samples



Binary value for each triplet (sample, gene, alteration).

Courtesy of Natalie Davidson

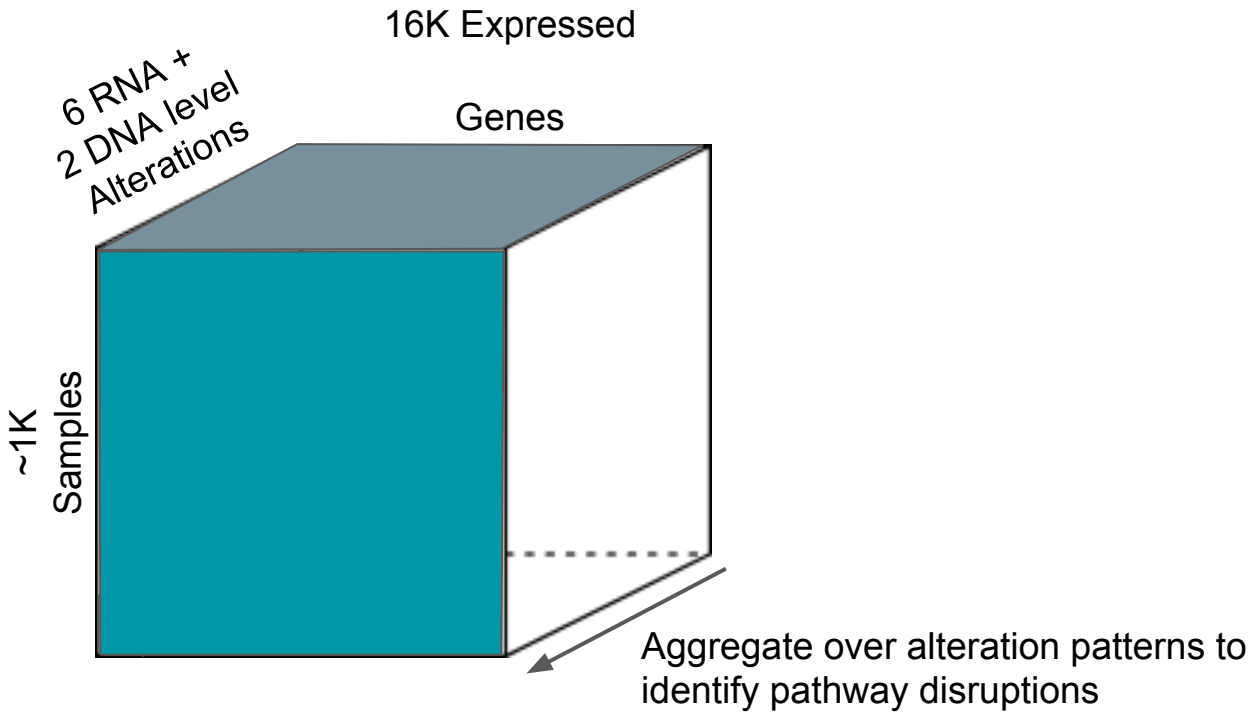
Summarize Alterations Across Genes or Samples



Aggregate over samples to perform a recurrence analysis

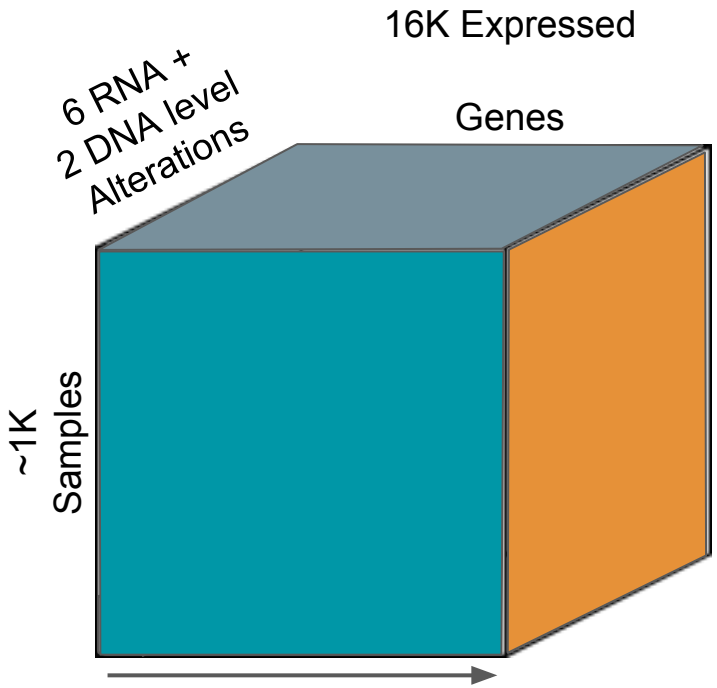
Courtesy of Natalie Davidson

Summarize Alterations Across Genes or Samples



Courtesy of Natalie Davidson

Summarize Alterations Across Genes or Samples

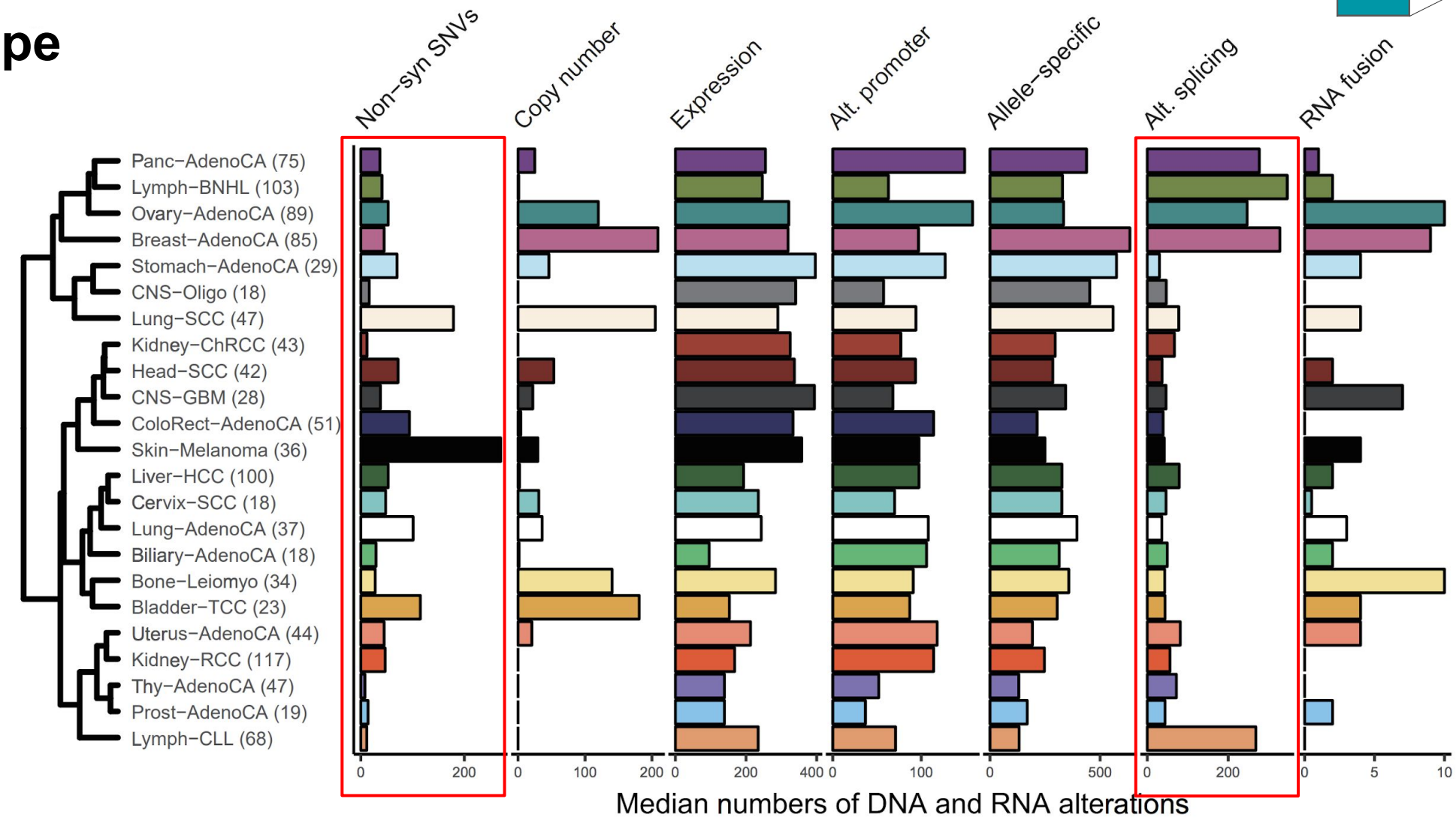


Aggregate over genes to identify sample specific transcriptomic patterns

Courtesy of Natalie Davidson

Compare and contrast samples

➡ Cancer Type



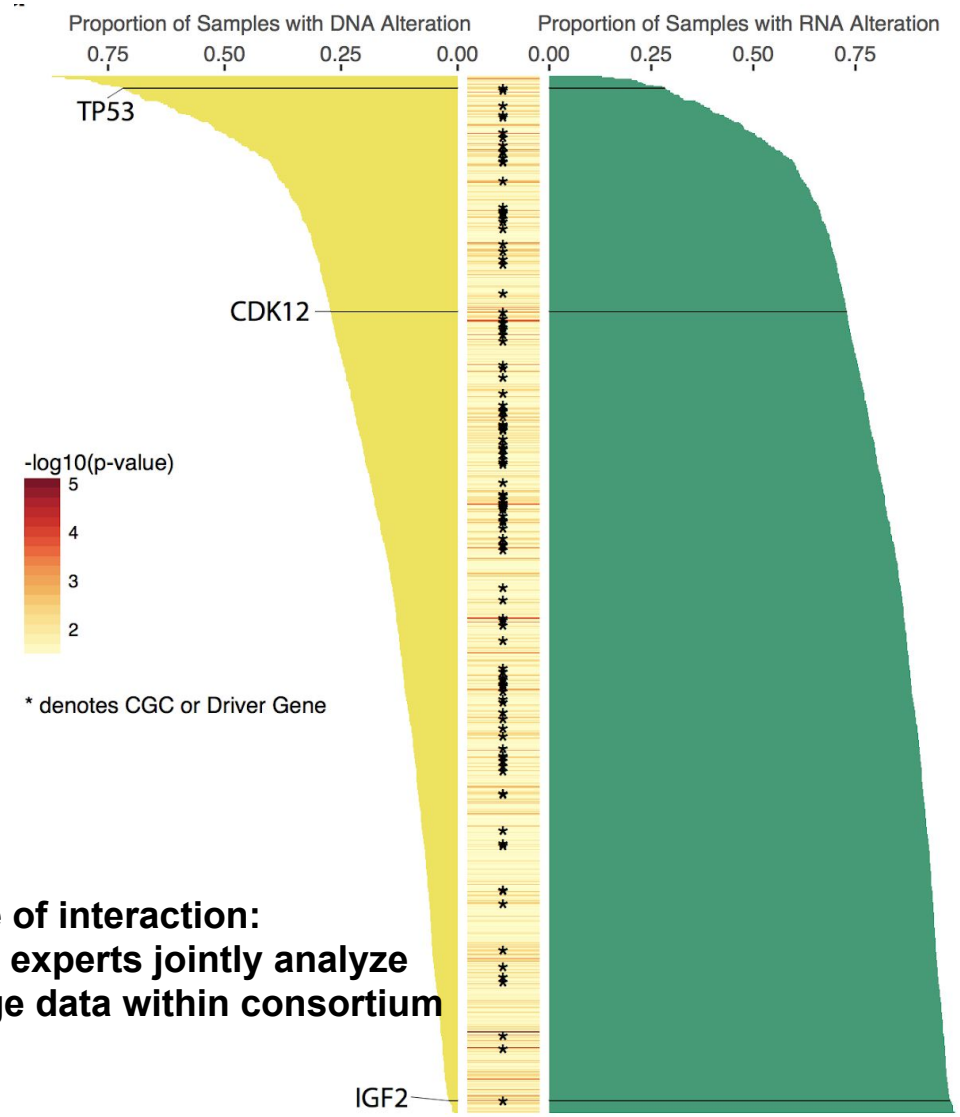
Source: PCAWG-3 Working group

Identify known and novel recurrently altered genes

DNA alterations

Copy-Number

Non-synonymous Mutations



RNA alterations

RNA-Editing

Fusions

Expression Outlier

Alternative Promoters

Alternative Splicing

Allele Specific Expression

1,012 genes depicted have a significant recurrence score.

This score:

- Guards against frequent alterations dominating
- Prioritizes genes that are heterogeneously altered

Mode of interaction:
Many experts jointly analyze a large data within consortium

All Recurrently Altered Genes (empirical p-value < 0.05) Source: PCAWG-3 Working group

BIOMEDICAL INFORMATICS

Project 2: Cancer-Specific Splicing & Implications in 8,705 Samples

Goals:

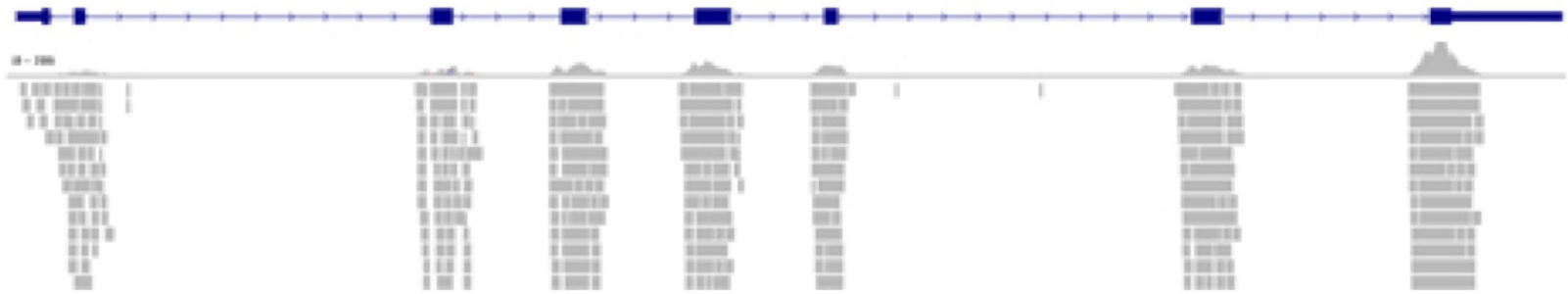
- Identify cancer-specific splicing patterns
- Identify variants regulating splicing in the same gene (cis)
- Identify variants regulating splicing in other genes (trans)
- Is splicing relevant for cancer treatments?

The Cancer Genome Atlas provides RNA-seq and matching exome data

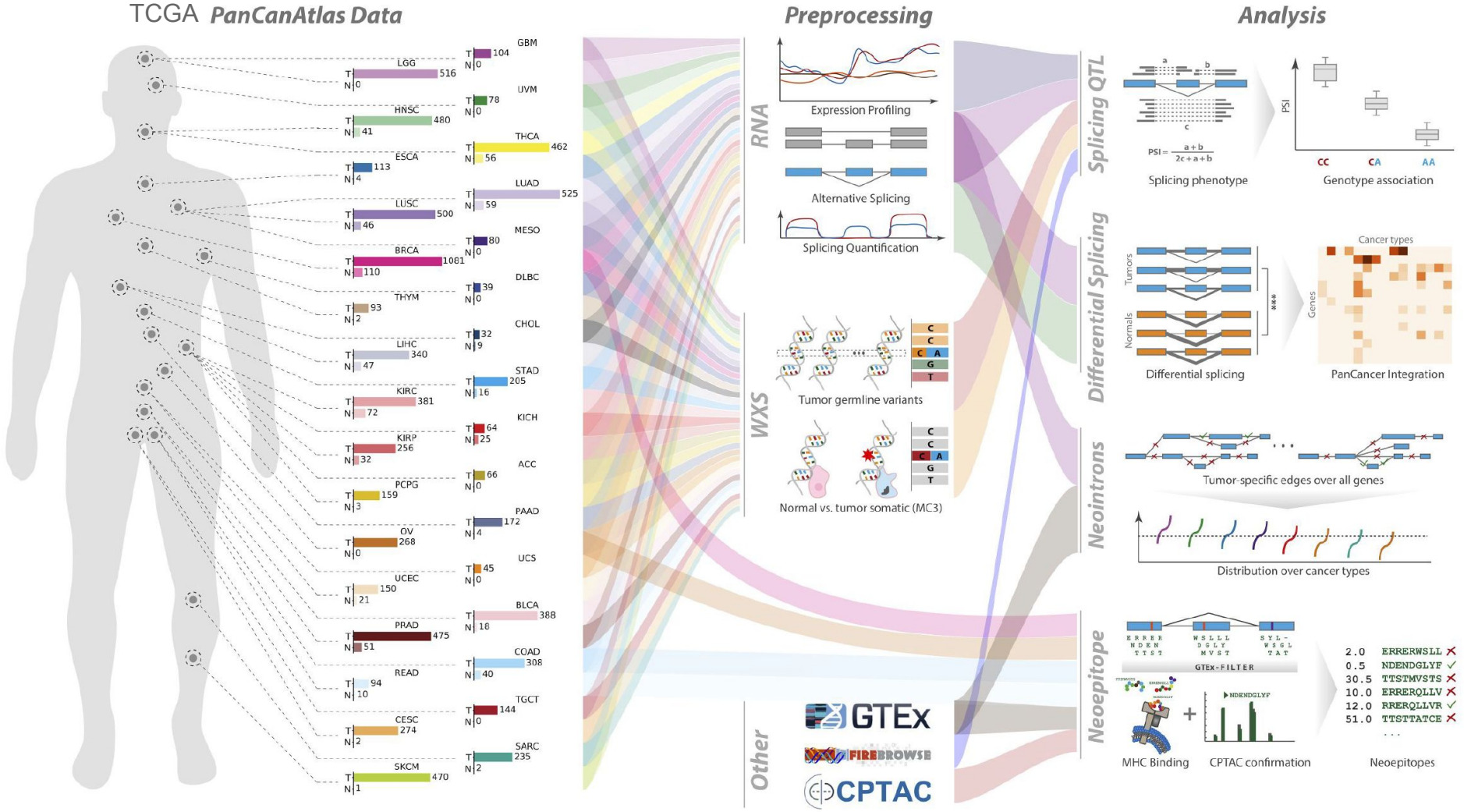
- RNA-seq => Find & quantify splicing events
- Exome => Identify variants in exons and flanking intronic regions

Mode of interaction:
One group of experts jointly analyzes data from multiple source.

Would not be possible without data sharing.

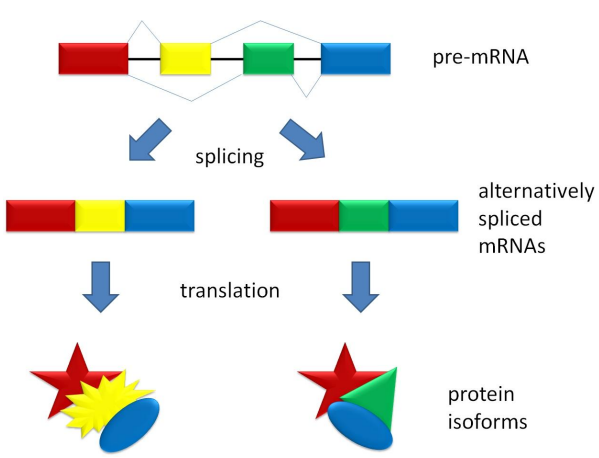


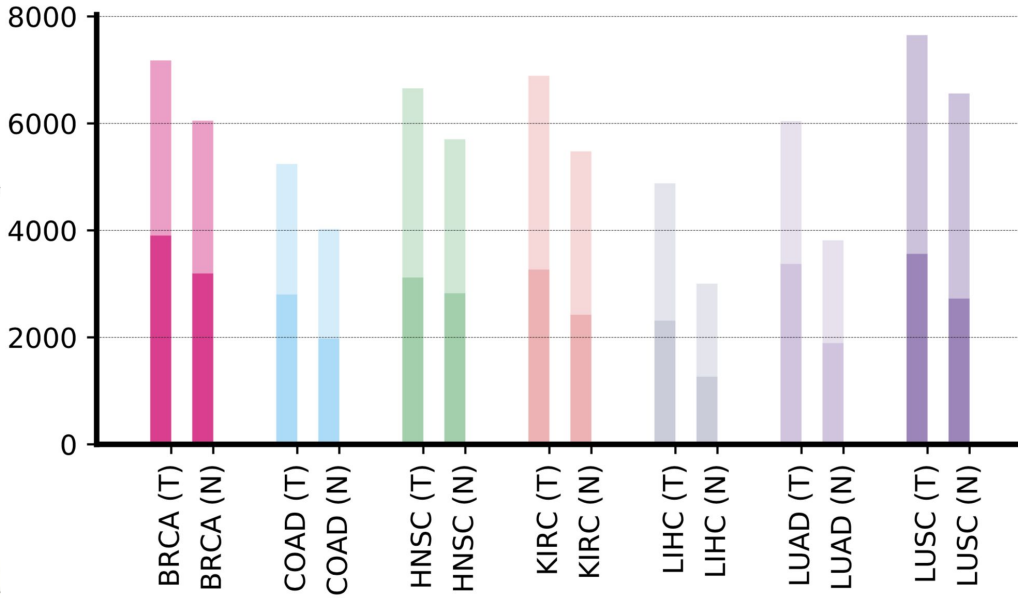
Analysis of Aberrant Splicing in Cancer in 8,512 tumors

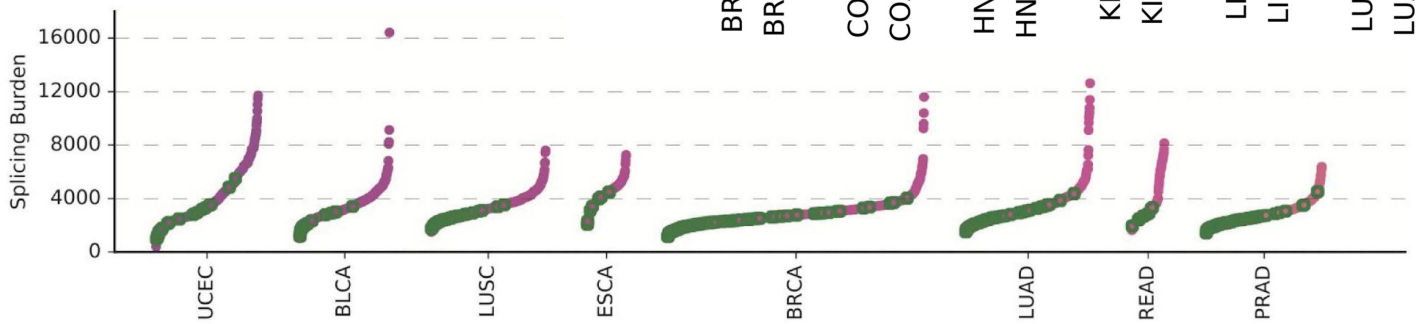


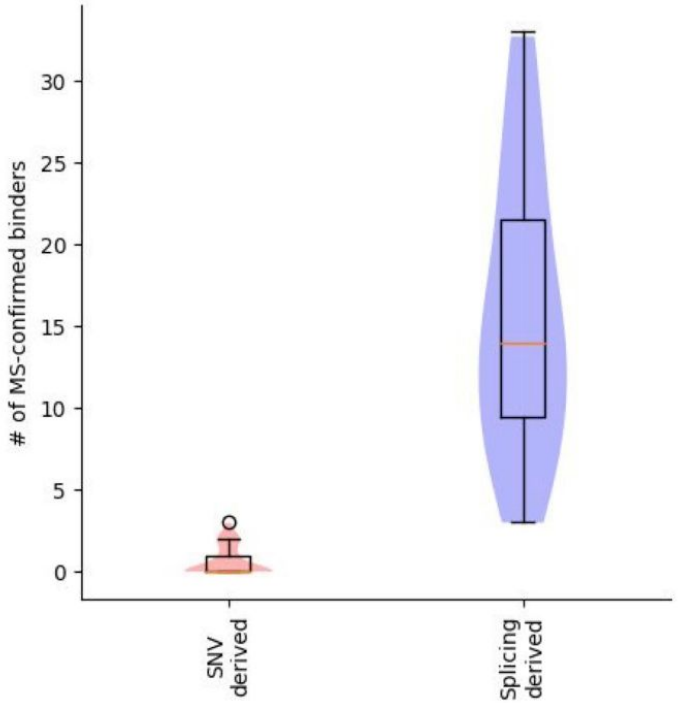
Kahles, Lehmann, et al., Rätsch, (Cancer Cell, 2018)

Analysis of Aberrant Splicing in Cancer in 8,512 tumors









Question: Can aberrant splicing be exploited in immunotherapies?

BIOMEDICAL INFORMATICS

Kahles, Lehmann, et al., Rätsch, (Cancer Cell, 2018)

Project 3: BRCA Exchange

The BRCA EXCHANGE

Sharing Global Knowledge about BRCA1/2

Current State, Challenges & Opportunities

Gunnar Rätsch
ETH Zürich
(MSKCC New York)

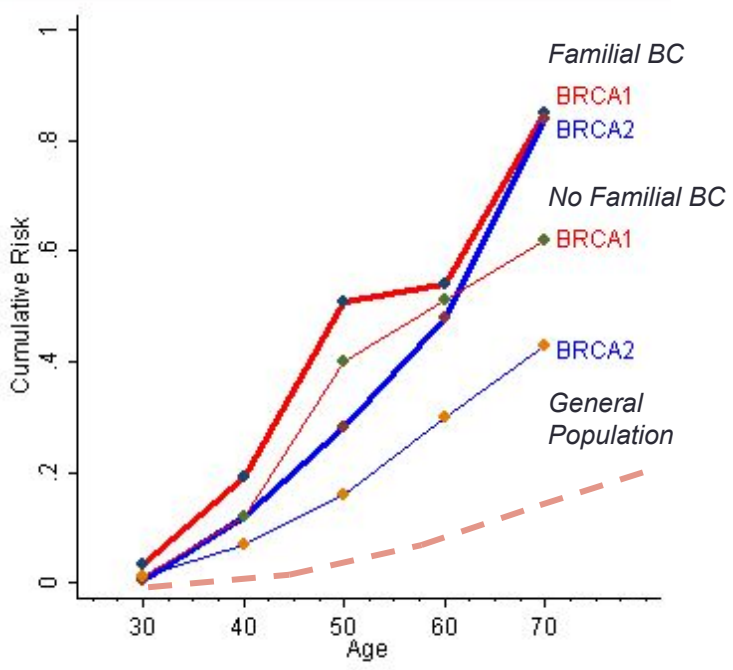
Motivation for the BRCA Challenge

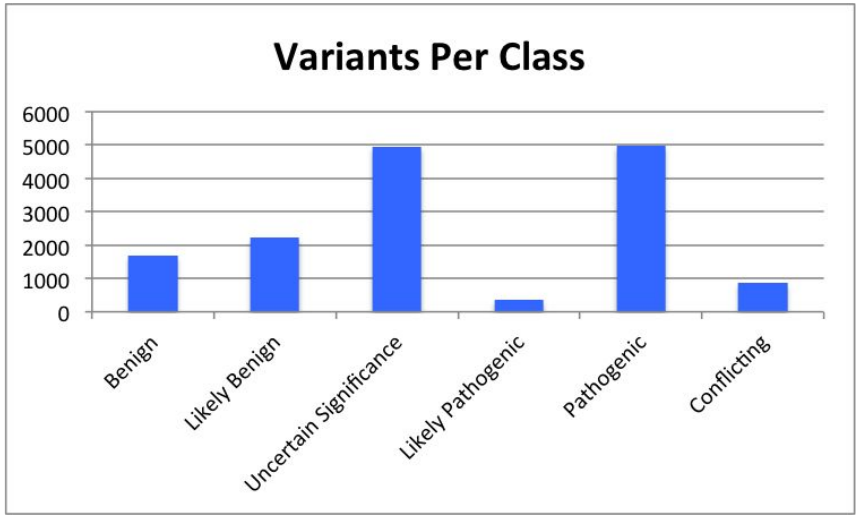
BRCA variation is relatively common with well known medical implications

Problem 1: Many variants lack clear interpretation

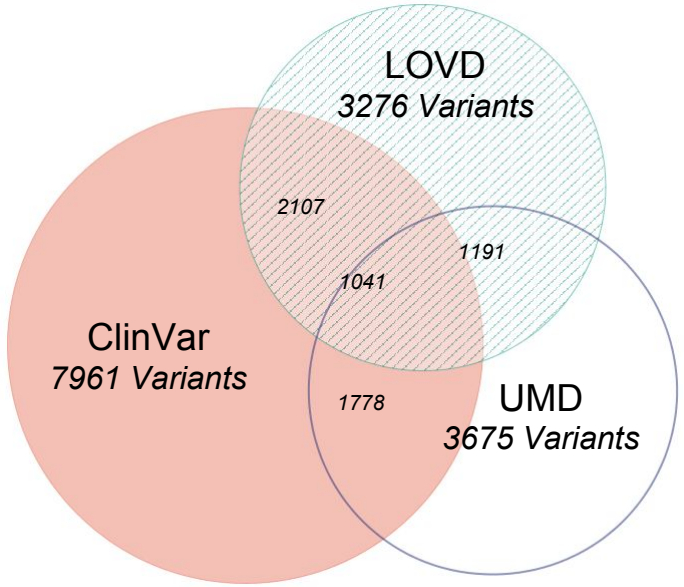
Problem 2: Variation databases are disjoint

Problem 3: Too little available data for effective curation





Shown: *the BRCA variants in ClinVar as of 6/22/17*

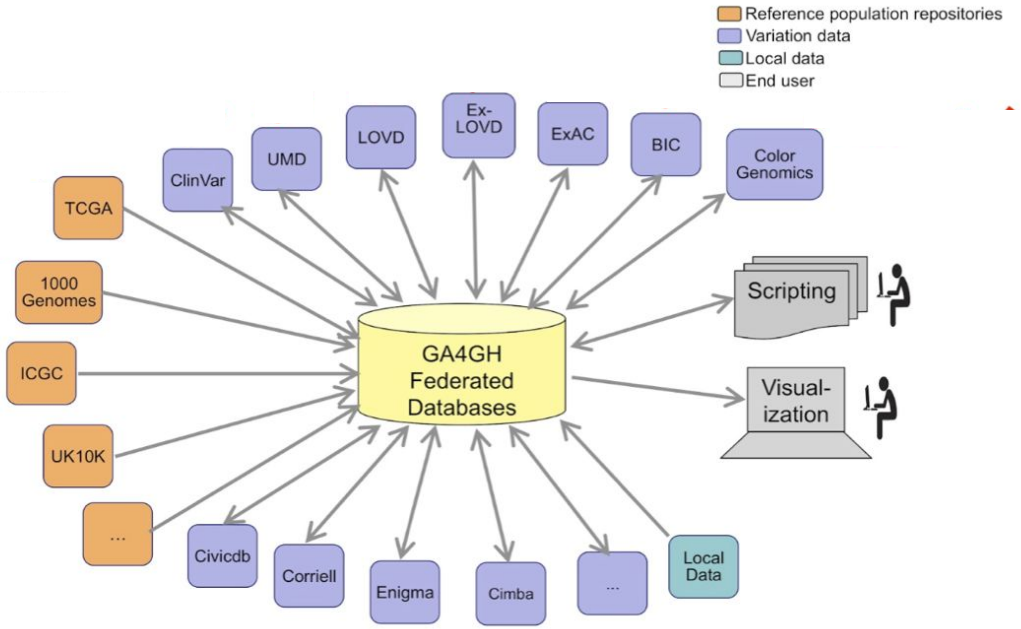


Source: BRCA Challenge working group

Wouldn't it be nice, if ...

HGVS Variant Lookup

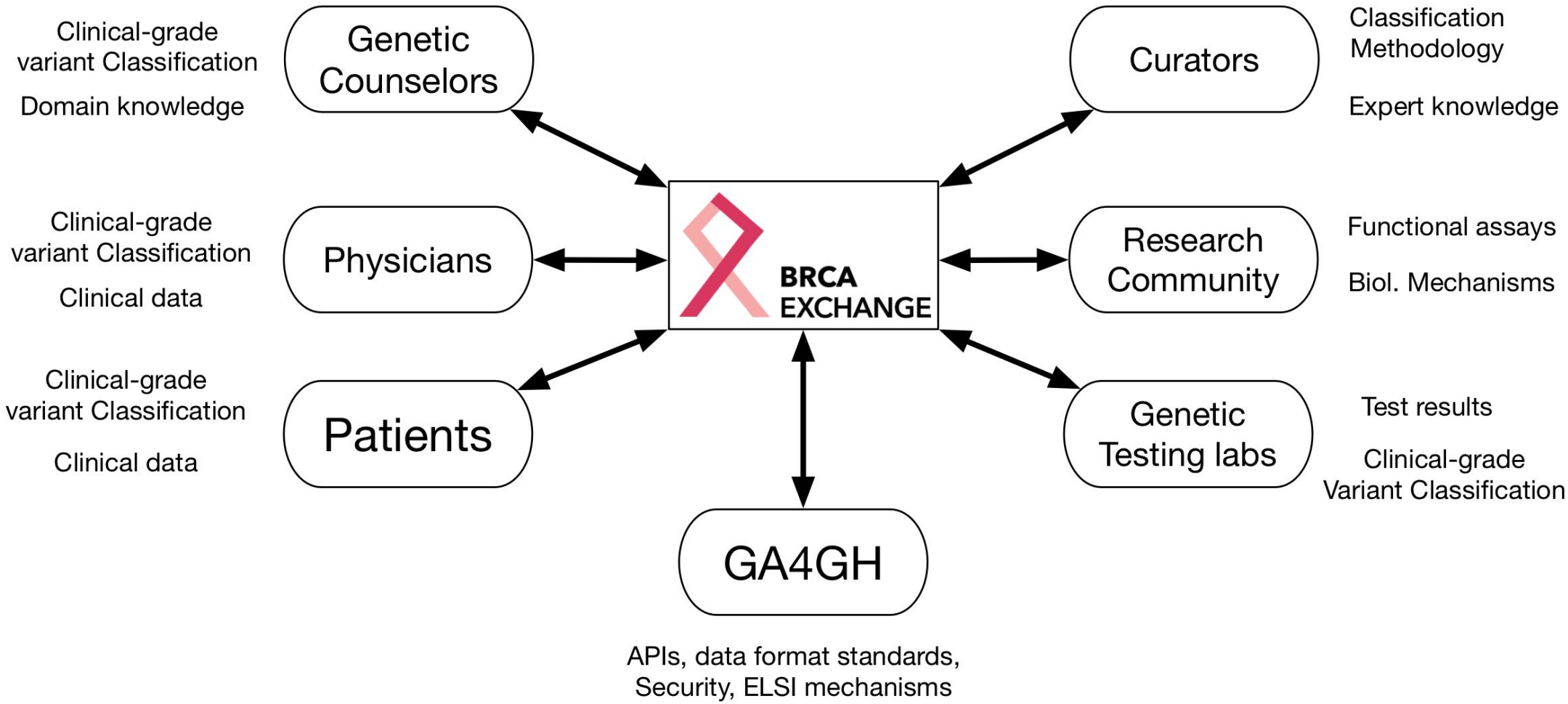
Federated Database



Variant Curation

Source: BRCA Challenge working group

Goal: One-Stop Shop for BRCA1/2 Variant Data



Want everything! But only for two well-studied genes. GA4GH driver project!

Types of Data

Variant-level (data annotated to a variant)

- Genotype, classification, allele frequencies, etc.; the majority of our data.
- Well-structured, most problems from inaccurate/ambiguous variant specs.
- Easy to share.

Case-level (data annotated to a case/patient)

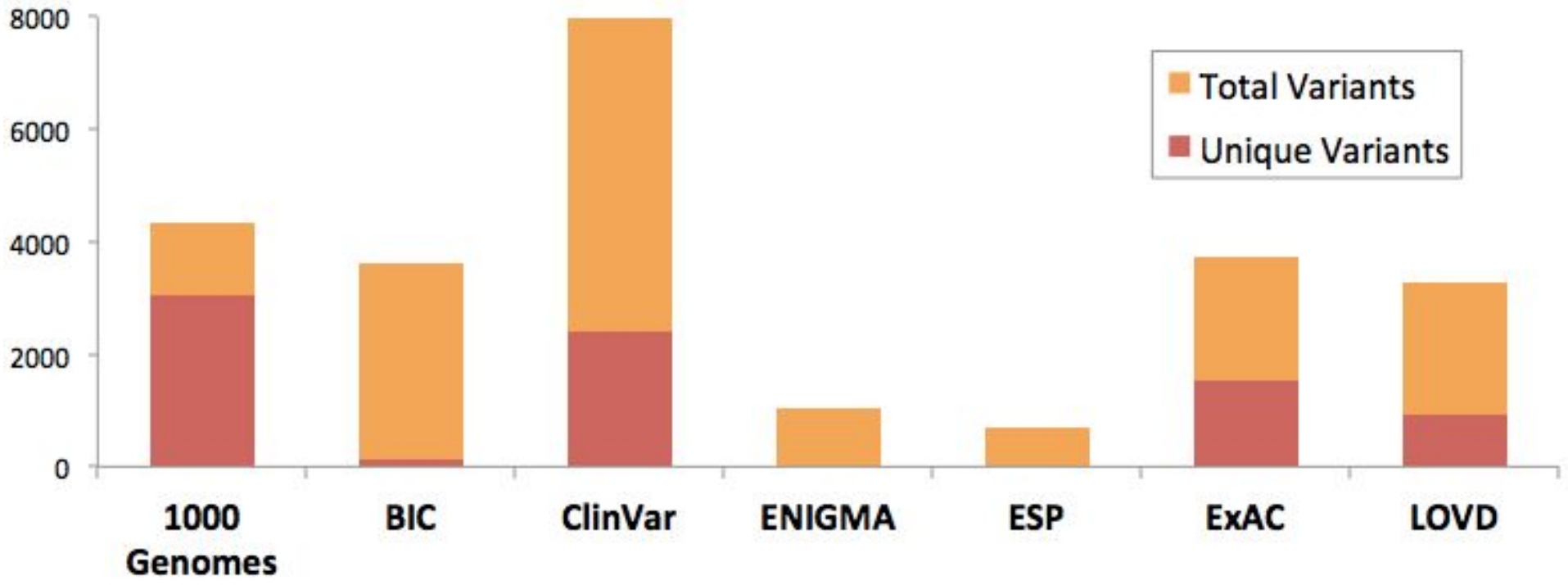
- Currently a small percentage of our data. Hoping to grow!
- Detailed data of cancer history, molecular features, family history, pedigrees.
- Heterogeneous clinical data.
- Privacy concerns, may need controlled-access mechanisms or other privacy enhancing mechanisms.

BRCA Exchange: Variant Exchange Platform

Highlights:

- Federated network for variant data exchange: ClinVar, LOVD, BIC, ExAC, ...
- Uniform variant processing and identification.
- Open source, cloud based, fully automatic (https://github.com/BRCAChallenge/brca-exchange).
- Public access, monthly releases and versioning support.
- Programmatic access via GA4GH interfaces.
- Largest public (federated) repository of BRCA1/2 variants.

Each repository contributes distinct information on BRCA1/2 variation



(As of 10/2016)

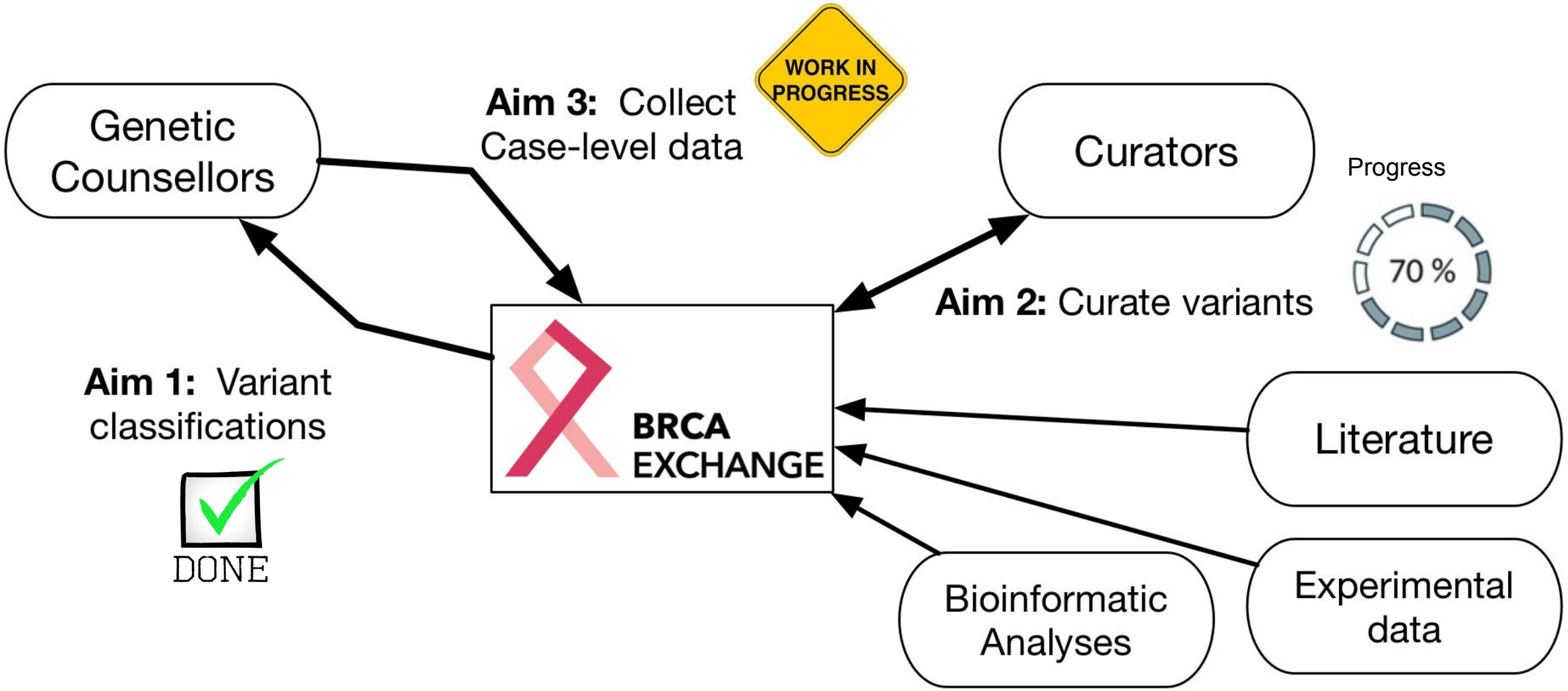
Combined, BRCA Exchange has **21,691** individual deduplicated variants (11/2018, monthly release)

Largest BRCA1/2 public variant database worldwide.

Cline et al., PLoS Genetics, December 2018, in press.

Source: BRCA Challenge working group

Data Flow and Project Aims



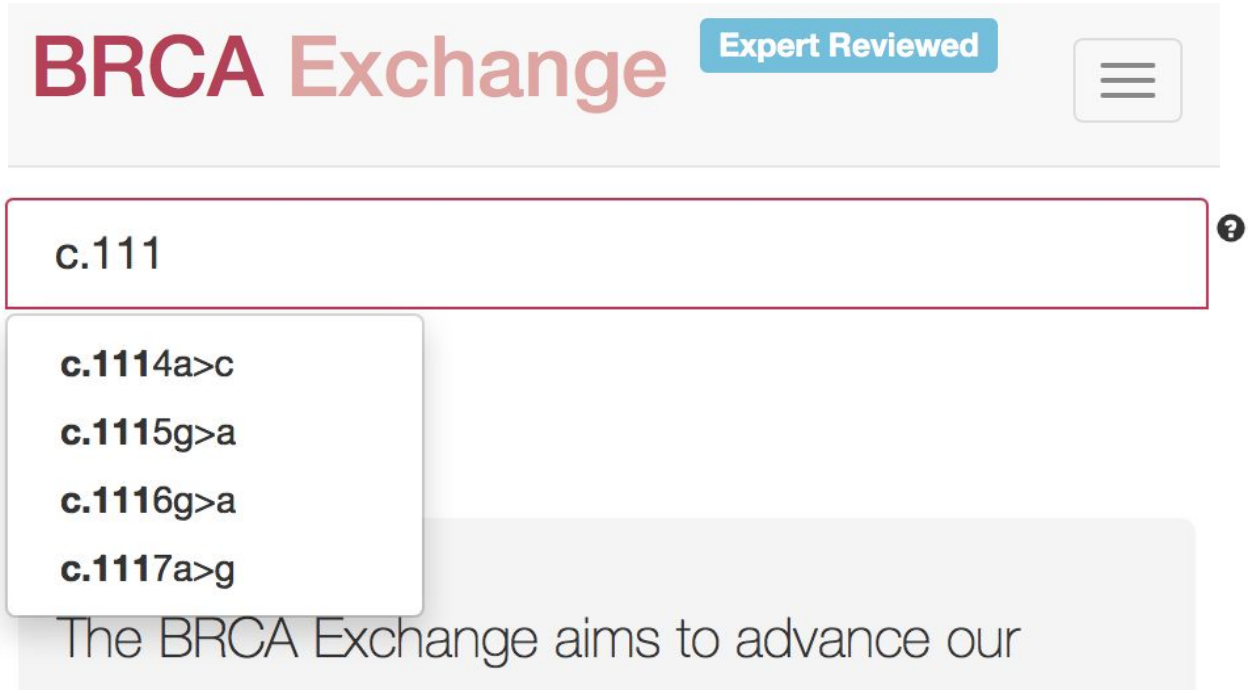
Aim 1: Enable Finding Variant Classifications

- One place for all known BRCA1/2 variants
- Highlight expert-panel reviewed variant interpretations for clinical use
- Simple to use user interface

BIC

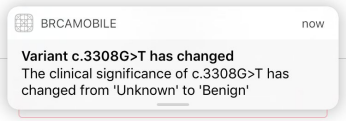
CIMBA

(The Consortium of Investigators of Modifiers of BRCA1/2)



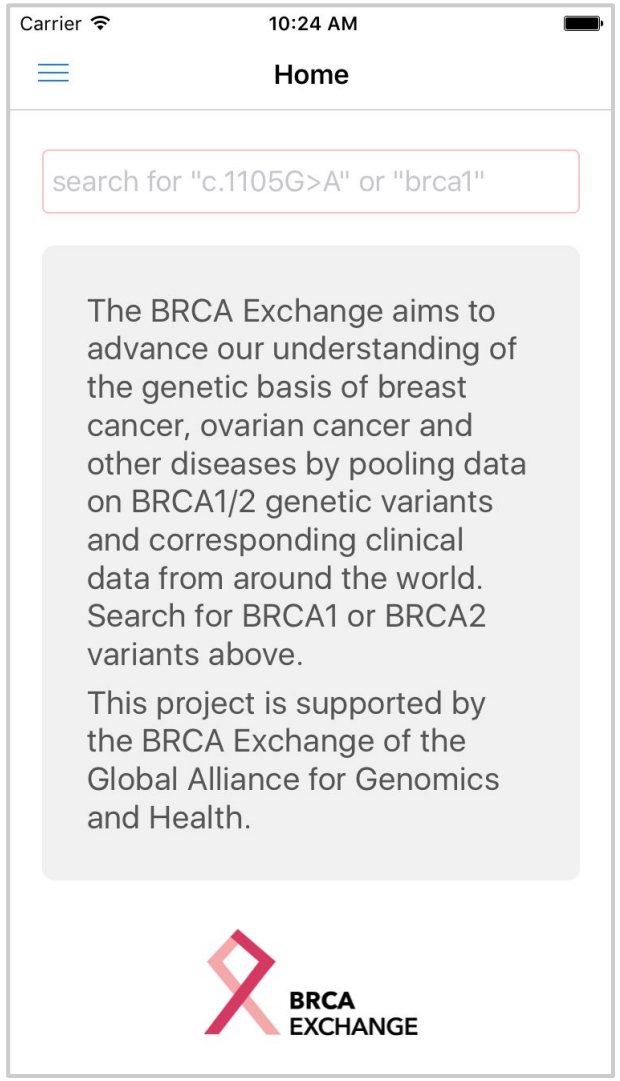
Variant Lookup via BRCA Exchange App

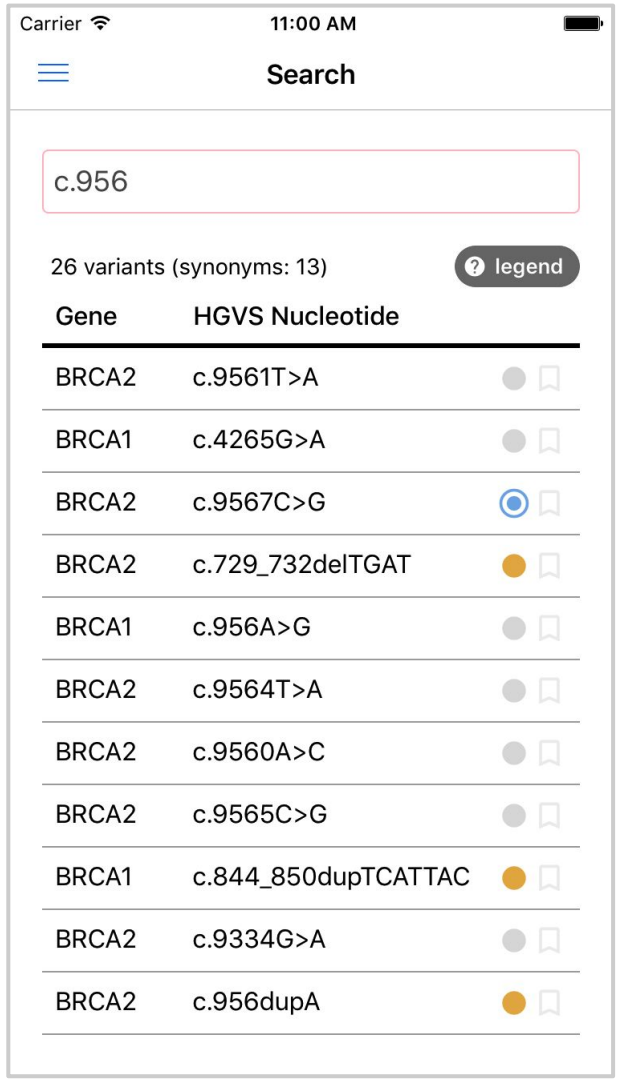
BRCA Exchange Expert Reviewed

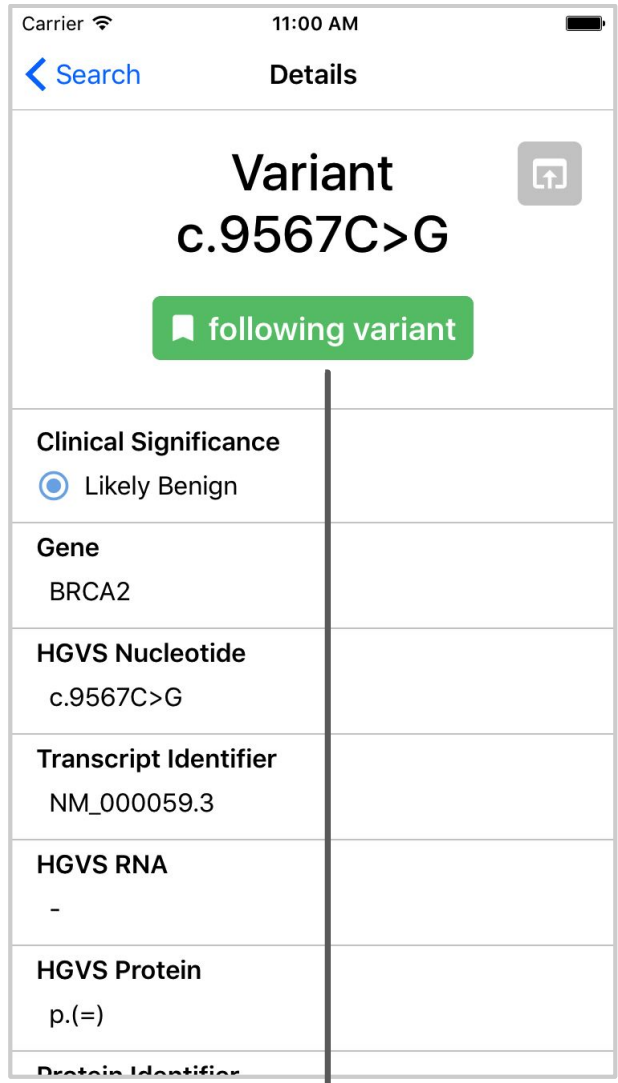


Push notification

Variant update
+
following variant







Aim 2: Research Data & Curation Environment

- Information necessary for variant classification (allele frequencies, priors, …)
- Data from many different, possibly disagreeing sources
- Curation tools & partially automatic variant classification
- All public data!

BIC

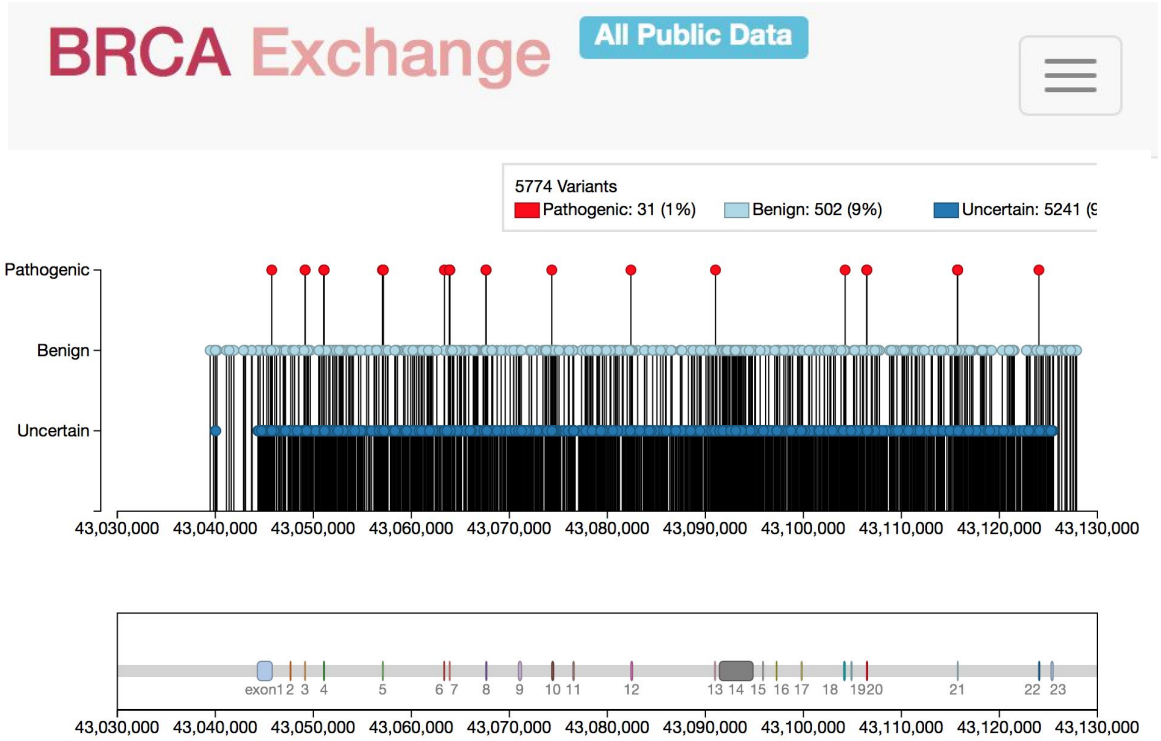
ClinVar
Clinically relevant variation

ENIGMA

LOVD
Leiden Open Variation Database

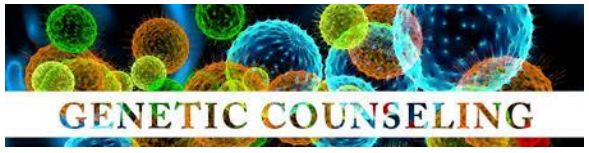
1000 Genomes Project

ExAC



Aim 3: Case Level Data Exchange

- Provide infrastructure to collect & store case-level data
- Genotypes, clinical data, family history, etc.
- Analysis tools: based on family history; multi-factorial
- Controlled access mechanisms



Health Trains!?

BRCA Exchange -- Global Data Sharing demonstrated

Aims:

- Share variant information to clinicians/physicians
- Provide platform to facilitate research & variant curation
- Collect data from case-level data repositories help curation of VUS

We need your help!

- Help connect us to large case-level repositories. National initiatives/consortia.
- Come talk to me or write email **raetsch@ethz.ch**

Technical, legal, organizational challenges are similar for other diseases:

- Relatively easy to replicate BRCA Exchange for other diseases/genes
  - MMR/InSIGHT variant database
  - Lynch Syndrome
  - Other hereditary cancers

Mode of interaction:
Groups of experts solve a global challenge to make clinical variant interpretation more effective.

Data Sharing is key!

Summary

Data sharing is key to progress in biomedicine.

- Project 1: Effective collaboration at the highest level
- Project 2: Advanced data aggregation across different technologies and cohorts
- Project 3: Leverage global expert knowledge to make medical genetics more efficient

Data sharing has to be thought about globally, but implemented locally.

Global Alliance for Genomics & Health

Acknowledgements to Collaborators

Biomedical Informatics
Cristóbal Esteban
Andre Kahles
Kjong Lehmann
Stephanie Hyland
Natalie Davidson
Gideon Dresdner
Stefan Stark
Xinrui Liu
Matthias Hüser
Vipin Sreedharan
David Kuo
Francesco Locatello

Alumni
Julia Vogt
Yi Zhong
Linda Sundermann
Melanie Fernandez
Theofanis Karaletsos
Katherine Redfield-Chan

MSKCC Cancer Biology
Guido Wendel
Kamini Singh

MSKCC Molecular Oncology Center
Niki Schultz
David Solit
David Hyman

MSKCC IT Services
Chris Crosbie
Stuart Gardos
Juan Perin

ETH IT Services
Bernd Rinn
Olivier Byrde
Stefan Walter

Global Alliance for Genomics and Health
David Haussler/UCSC
Benedict Paten/UCSC
Melissa Cline/UCSC
Stephen Chanock/NCI
John Burn/University of Newcastle

International Cancer Genome Consortium
Angela Brooks/UCSC
Alvis Brazma/EBI
Oliver Stegle/EBI

NEXUS@ETH
Nora Toussaint
Daniel Stekhoven

ETH BSSE
Dean Bodenham
Karsten Borgwardt

University of Tübingen
Oliver Kohlbacher

Funding: ETH Zürich, Sloan Kettering Institute, Memorial Hospital, National Institute of Health, National Cancer Institute, Swiss National Science Foundation, Max Planck Society, German Research Foundation, European Union, Geoffrey Beene Foundation, Lucille Castori Center

Thank You!

Courtesy of Mark Rubin

Questions?