Welcome to the...

NIH Cloud Platforms Interoperability Fall 2020 Workshop

We'll be starting shortly!

NIH National Heart, Lung, and Blood Institute | BioData CATALYST

AnVIL

Gabriella Miller Kids First Pediatric Research Program Data Resource Center

NIH National Cancer Institute Cancer Research Data Commons

Welcome & Introduction to Day 1

Adam Resnick
Children's Hospital of Philadelphia
Valerie Cotton
Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), NIH

Introduction & Congratulations!

https://datascience.nih.gov/nih-cloud-platform-interoperability

NIH National Institutes of Health
Office of Data Science Strategy

https://datascience.nih.gov/nih-cloud-platform-interoperability

NIH Cloud Platform Interop...

NIH Cloud Platform Interoperability Effort

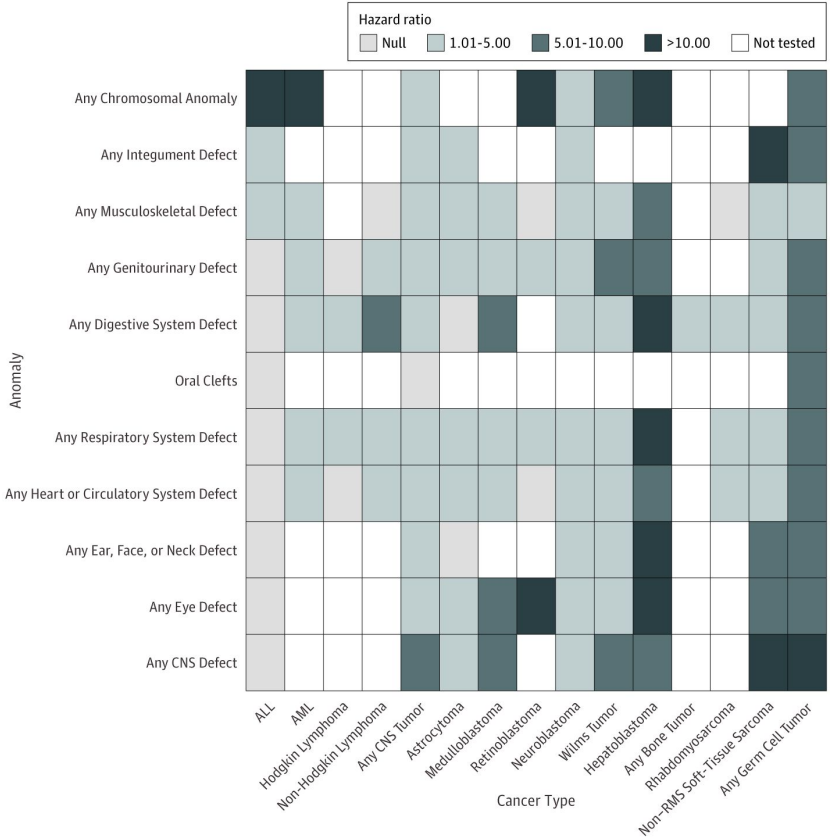
About the NIH Cloud Platform Interoperability (NCPI) Effort

Connecting NIH's various data systems is a critical step toward improving researchers' access to all types of data. The NIH Cloud Platform Interoperability (NCPI) effort seeks to create a federated genomic data ecosystem and is a collaborative project between NIH and external partners comprising five working groups.

When researchers obtain data from a specific platform, there is no guarantee that the data will be readily usable alongside data from a different platform. By focusing on interoperability, the NCPI effort is ensuring that researchers can both find and integrate data more easily from the following four participating platforms:

Gabriella Miller Kids First Pediatric Research Program: Interoperability of *childhood cancer & structural birth defects*

Birth defects are associated with increased risk of cancer among children… ***suggesting shared genetic pathways***



From: **Association Between Birth Defects and Cancer Risk Among Children and Adolescents in a Population-Based Assessment of 10 Million Live Births**

Lupo et al, JAMA Oncol. 2019;5(8):1150-1158. doi:10.1001/jamaoncol.2019.1215

Kids First Sequencing Cohorts 2015-2020

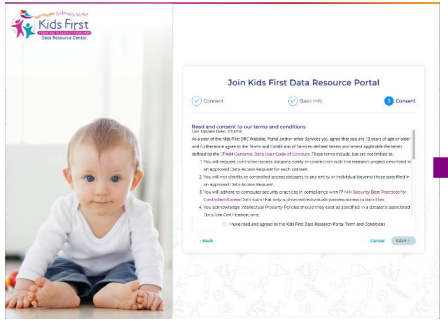
40 projects | 40,000 genomes | 16,000 cases | 14 released datasets

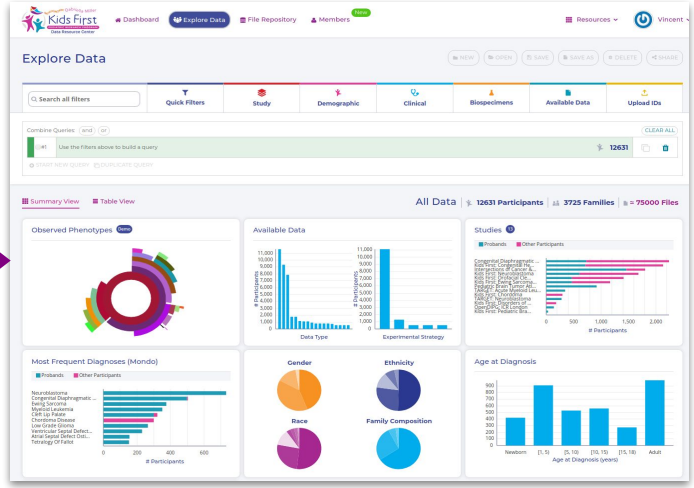
BROAD INSTITUTE | HUDSONALPHA INSTITUTE FOR BIOTECHNOLOGY | St. Jude Children's Research Hospital

- Disorders of Sex Development
- Congenital Diaphragmatic Hernia
- Ewing Sarcoma
- Structural Heart & Other Defects
- Syndromic Cranial Dysinnervation Disorders
- Cancer Susceptibility
- Adolescent Idiopathic Scoliosis
- Neuroblastomas
- Enchondromatoses
- Orofacial Clefts in Caucasian, Latin American, Asian & African, Filipino populations
- Osteosarcoma
- Familial Leukemia
- Hemangiomas, Vascular Anomalies & Overgrowth
- Craniofacial Microsomia
- Intersection of childhood cancer & birth defects
- Microtia
- Esophageal Atresia and Tracheoesophageal Fistulas
- Kidney and Urinary Tract Defects
- Nonsyndromic Craniosynostosis
- Bladder Exstrophy
- Hearing Loss
- Cornelia de Lange Syndrome
- Intracranial & Extracranial Germ Cell Tumors
- Fetal Alcohol Spectrum Disorders
- Myeloid Malignancies + overlap with Down syndrome
- CHD & ALL in Children with Down Syndrome
- Structural Brain Defects
- Structural Defects of the Neural Tube (Myelomeningocele)
- CHARGE Syndrome
- Laterality Birth Defects
- T-cell Acute Lymphoblastic Leukemia
- Pediatric Rhabdomyosarcoma
- Valvar Pulmonary Stenosis

Use Case: Compare genetic variants of congenital heart defects & neuroblastoma

Anyone can register & login to the portal (via ORCID, Google). User agrees to terms

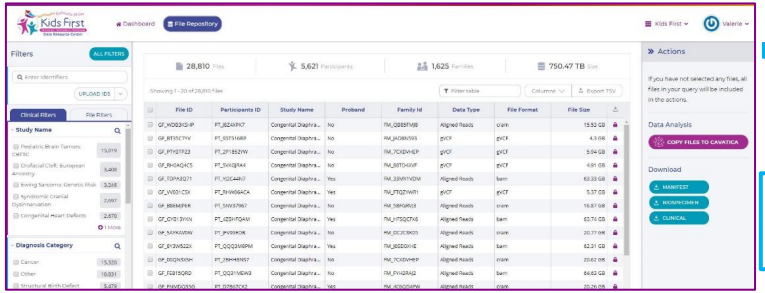




In ***Explore Data***, user searches the terms "heart" and "neuroblastoma". Discovers data from children with congenital heart disease (KF & BDC data) & neuroblastoma (KF & NCI TARGET)

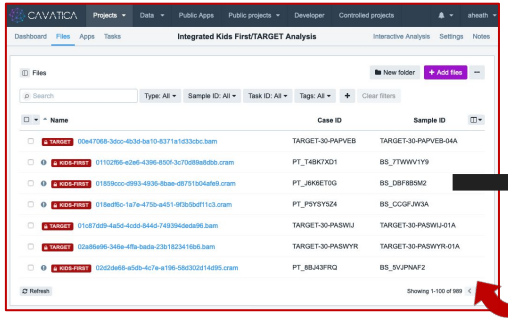
User builds a synthetic cohort based on these criteria and can view summary & deidentified individual-level clinical, demographic, and phenotypic information.

Synthetic cohort is ported to the ***File Repository*** where user selects which **genomic** and **histology image** files they want to analyze.

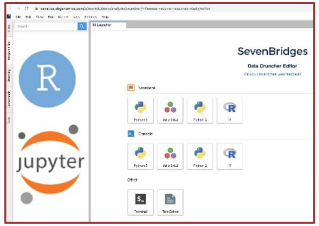


User has or applies for dbGaP access for genomic data

User pushes genomic, clinical, and image data into Cavatica for analysis & visualization



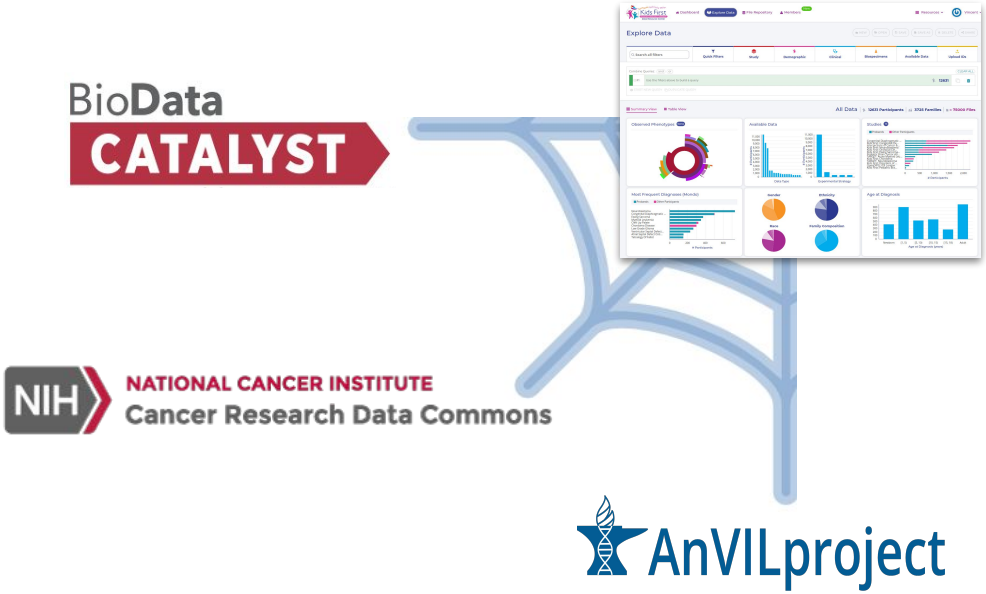
User runs statistical analyses in notebooks



User iterates through genomic workflows

Childhood Cancer & Structural Birth Defects Use Cases:

- Childhood Cancer data from TARGET in the CRDC
- Congenital Heart Disease data from TOPMed/PCGC in BioData Catalyst
- Structural Birth Defects data from the CMGs in AnVIL

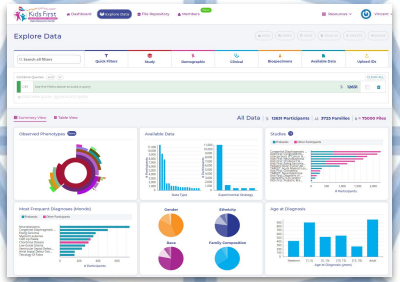


Additional Use Cases for Pediatric Federation

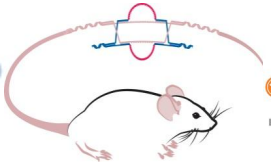
COMMON FUND DATA ECOSYSTEM

INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndrome (INCLUDE)

db GaP GENOTYPES and PHENOTYPES



BioData CATALYST



IMPC International Mouse Phenotyping Consortium

KOMP2

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

IMMPORT Shared Data
Your site for searching and downloading shared data

AnVILproject

Tackling Multiple Layers of Interoperability

Challenge	Working Group	NCPI Activities
Operational barriers to trans-platform data sharing	**Community Governance**	Establish principles for promoting interoperability across multiple platforms; evaluate operational barriers
Inability to search & access data across platforms	**Systems Interoperation**	Test & implement technical standards for auth (RAS) & data exchange (e.g. GA4GH DRS) based on key use cases
Transitioning researchers to use the cloud	**Outreach & Training**	Create public “knowledge base”; create training materials
Lack of standards for clinical data exchange	**FHIR**	Pilot and assess FHIR resources to model and share complex clinical and phenotypic data

Additional Challenges for Potential NCPI Roadmap

Challenge	Potential NCPI Activities?
Users don't want to use the cloud if their favorite tools and workflows are not there	Potential new WG to port workflows to the cloud? New activity of Systems Interop and/or Outreach/Training group?
New programs, platforms, and databases want to play in the sandbox	How do we onboard new programs or development teams to NCPI?
How to estimate cloud costs for researcher analyses	Benchmark pipelines? Create public cloud cost guide?
Complex clinical and phenotypic data (that don't map to CDMs/CDEs)	FHIR as a flexible structure for clinical data interoperability (even if not derived from EHRs)

INTEROPERABILITY

interoperability definition

All | News | Images | Shopping | Videos | More | Settings | Tools

About 22,600,000 results (0.44 seconds)

Dictionary

Search for a word

in·ter·op·er·a·bil·i·ty

/ˌin(t)ərˌäp(ə)rəˈbilədē/

noun

the ability of computer systems or software to exchange and make use of information.
"interoperability between devices made by different manufacturers"

- the ability of military equipment or groups to operate in conjunction with each other.
  "staff believe interoperability between forces is crucial to effectiveness"

Does Mark Have an Interoperability Problem?

[External] H3F3B G34W variant

MC

Mark Cowley <MCowley@ccia.org.au> Wednesday, September 23, 2020 at 10:00 PM

To: Resnick, Adam C; **Cc:** Pamela Ajuyah; Paul Ekert; Paulette Barahona; Loretta Lau (External)

You forwarded this message on 9/24/20, 6:44 AM. Show Forward

You replied to this message on 9/24/20, 8:16 AM. Show Reply

This message is flagged for follow up.

Dear Adam,
We have a diagnostic dilemma that I hope you could help us with?

The case has been challenging to diagnose by histopathology and also molecularly due to low tumour purity. The patient has a left thalamic/midbrain lesion and it is unclear whether it is a low grade or high grade glioma (which dictates the treatment the patient will receive). She has the canonical BRAF:p.V600E mutation which would be a clear driver in the tumour but of less certainty is a H3F3B G35W variant in the tumour (NM_005324(H3F3B):c.103G>T (p.Gly35Trp) - which if deemed pathogenic would bump up the grade). The MNP classifier failed to classify the tumour (maybe due to low purity), but was confidently MGMT methylated.

Literature supports H3F3A G34W in Giant cell tumours of bone (GCTB), but not in brain tumours. We've never seen H3F3A G34W, but we have seen G34R four times, all reported as pathogenic. All the G34R's had MGMT methylation, an association reported in the literature (28966033, 25752754), and thus pushing us towards saying the variant is pathogenic.

Have you seen this and can make a comment?

Thanks,
Mark

Mark Cowley, PhD BSc (Bioinf, Hons 1)
Computational Biology Group Leader
Conjoint Associate Professor, School of Women's and Children's Health, UNSW Medicine

Children's Cancer Institute
Lowy Cancer Research Centre, UNSW Australia
PO Box 81 Randwick 2031 Australia
P: 02 9385 2074 | **M:** 0413 481 017 | **E:** MCowley@ccia.org.au | **W:** www.ccia.org.au | **T:** @markjcowley

Does Mark Have an Interoperability Problem?

Dear Adam,
We have a diagnostic dilemma that I hope you could help us with?

The case has been challenging to diagnose by histopathology and also molecularly due to low tumour purity. The patient has a left thalamic/midbrain lesion and it is unclear whether it is a low grade or high grade glioma (which dictates the treatment the patient will receive). She has the canonical BRAF:p.V600E mutation which would be a clear driver in the tumour but of less certainty is a H3F3B G35W variant in the tumour (NM_005324(H3F3B):c.103G>T (p.Gly35Trp) - which if deemed pathogenic would bump up the grade). The MNP classifier failed to classify the tumour (maybe due to low purity), but was confidently MGMT methylated.

Literature supports H3F3A G34W in Giant cell tumours of bone (GCTB), but not in brain tumours. We've never seen H3F3A G34W, but we have seen G34R four times, all reported as pathogenic. All the G34R's had MGMT methylation, an association reported in the literature (28966033, 25752754), and thus pushing us towards saying the variant is pathogenic.

Have you seen this and can make a comment?

Thanks,
Mark

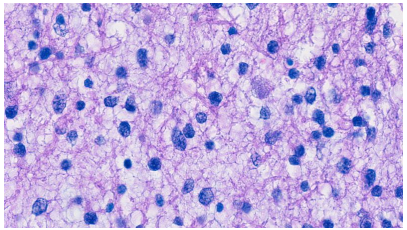
Mark Cowley, PhD BSc (Bioinf, Hons 1)
Computational Biology Group Leader
Conjoint Associate Professor, School of Women's and Children's Health, UNSW Medicine

Children's Cancer Institute
Lowy Cancer Research Centre, UNSW Australia
PO Box 81 Randwick 2031 Australia
P: 02 9385 2074 | **M:** 0413 481 017 | **E:** MCowley@ccia.org.au | **W:** www.ccia.org.au | **T:** @markjcowley

Does Mark Have an Interoperability Problem?

Dear Adam,
We have a diagnostic dilemma that I hope you could help us with?

The case has been challenging to diagnose by histopathology and also molecularly due to low tumour purity. The patient has a left thalamic/midbrain lesion and it is unclear whether it is a low grade or high grade glioma (which dictates the treatment the patient will receive). She has the canonical BRAF:p.V600E mutation which would be a clear driver in the tumour but of less certainty is a H3F3B G35W variant in the tumour (NM_005324(H3F3B):c.103G>T (p.Gly35Trp) - which if deemed pathogenic would bump up the grade). The MNP classifier failed to classify the tumour (maybe due to low purity), but was confidently MGMT methylated.



Literature supports H3F3A G34W in Giant cell tumours of bone (GCTB), but not in brain tumours. We've never seen H3F3A G34W, but we have seen G34R four times, all reported as pathogenic. All the G34R's had MGMT methylation, an association reported in the literature (28966033, 25752754), and thus pushing us towards saying the variant is pathogenic.

Have you seen this and can make a comment?

Thanks,
Mark

Mark Cowley, PhD BSc (Bioinf, Hons 1)
Computational Biology Group Leader
Conjoint Associate Professor, School of Women's and Children's Health, UNSW Medicine

Children's Cancer Institute
Lowy Cancer Research Centre, UNSW Australia
PO Box 81 Randwick 2031 Australia
P: 02 9385 2074 | **M:** 0413 481 017 | **E:** MCowley@ccia.org.au | **W:** www.ccia.org.au | **T:** @markjcowley

Does Mark Have an Interoperability Problem?

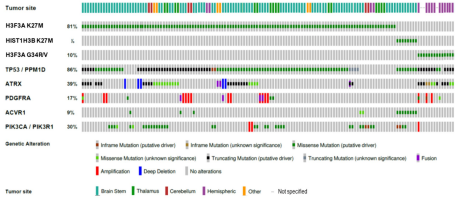
Dear Adam,
We have a diagnostic dilemma that I hope you could help us with?

The case has been challenging to diagnose by histopathology and also molecularly due to low tumour purity. The patient has a left thalamic/midbrain lesion and it is unclear whether it is a low grade or high grade glioma (which dictates the treatment the patient will receive). She has the canonical BRAF:p.V600E mutation which would be a clear driver in the tumour but of less certainty is a H3F3B G35W variant in the tumour (NM_005324(H3F3B):c.103G>T (p.Gly35Trp) - which if deemed pathogenic would bump up the grade). The MNP classifier failed to classify the tumour (maybe due to low purity), but was confidently MGMT methylated.

Literature supports H3F3A G34W in Giant cell tumours of bone (GCTB), but not in brain tumours. We've never seen H3F3A G34W, but we have seen G34R four times, all reported as pathogenic. All the G34R's had MGMT methylation, an association reported in the literature (28966033, 25752754), and thus pushing us towards saying the variant is pathogenic.

Have you seen this and can make a comment?

Thanks,
Mark



Mark Cowley, PhD BSc (Bioinf, Hons 1)
Computational Biology Group Leader
Conjoint Associate Professor, School of Women's and Children's Health, UNSW Medicine

Children's Cancer Institute
Lowy Cancer Research Centre, UNSW Australia
PO Box 81 Randwick 2031 Australia
P: 02 9385 2074 | **M:** 0413 481 017 | **E:** MCowley@ccia.org.au | **W:** www.ccia.org.au | **T:** @markjcowley

Does Mark Have an Interoperability Problem?

Dear Adam,
We have a diagnostic dilemma that I hope you could help us with?

The case has been challenging to diagnose by histopathology and also molecularly due to low tumour purity. The patient has a left thalamic/midbrain lesion and it is unclear whether it is a low grade or high grade glioma (which dictates the treatment the patient will receive). She has the canonical BRAF:p.V600E mutation which would be a clear driver in the tumour but of less certainty is a H3F3B G35W variant in the tumour (NM_005324(H3F3B):c.103G>T (p.Gly35Trp) - which if deemed pathogenic would bump up the grade). The MNP classifier failed to classify the tumour (maybe due to low purity), but was confidently MGMT methylated.

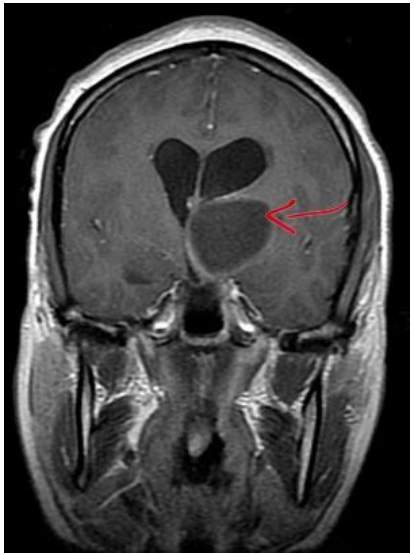
Literature supports H3F3A G34W in Giant cell tumours of bone (GCTB), but not in brain tumours. We've never seen H3F3A G34W, but we have seen G34R four times, all reported as pathogenic. All the G34R's had MGMT methylation, an association reported in the literature (28966033, 25752754), and thus pushing us towards saying the variant is pathogenic.

Have you seen this and can make a comment?

Thanks,
Mark

Mark Cowley, PhD BSc (Bioinf, Hons 1)
Computational Biology Group Leader
Conjoint Associate Professor, School of Women's and Children's Health, UNSW Medicine

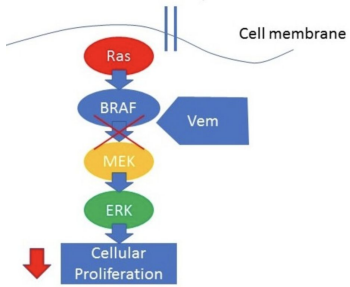
Children's Cancer Institute
Lowy Cancer Research Centre, UNSW Australia
PO Box 81 Randwick 2031 Australia
P: 02 9385 2074 | **M:** 0413 481 017 | **E:** MCowley@ccia.org.au | **W:** www.ccia.org.au | **T:** @markjcowley

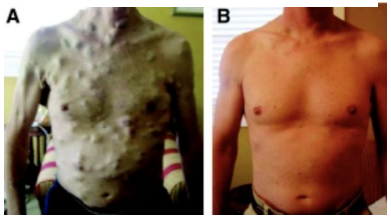


Does Mark Have an Interoperability Problem?

Dear Adam,
We have a diagnostic dilemma that I hope you could help us with?

The case has been challenging to diagnose by histopathology and also molecularly due to low tumour purity. The patient has a left thalamic/midbrain lesion and it is unclear whether it is a low grade or high grade glioma (which dictates the treatment the patient will receive). She has the canonical BRAF:p.V600E mutation which would be a clear driver in the tumour but of less certainty is a H3F3B G35W variant in the tumour (NM_005324(H3F3B):c.103G>T (p.Gly35Trp) - which if deemed pathogenic would bump up the grade). The MNP classifier failed to classify the tumour (maybe due to low purity), but was confidently MGMT methylated.





Literature supports H3F3A G34W in Giant cell tumours of bone (GCTB), but not in brain tumours. We've never seen H3F3A G34W, but we have seen G34R four times, all reported as pathogenic. All the G34R's had MGMT methylation, an association reported in the literature (28966033, 25752754), and thus pushing us towards saying the variant is pathogenic.

Have you seen this and can make a comment?

Thanks,
Mark

Mark Cowley, PhD BSc (Bioinf, Hons 1)
Computational Biology Group Leader
Conjoint Associate Professor, School of Women's and Children's Health, UNSW Medicine

Children's Cancer Institute
Lowy Cancer Research Centre, UNSW Australia
PO Box 81 Randwick 2031 Australia
P: 02 9385 2074 | **M:** 0413 481 017 | **E:** MCowley@ccia.org.au | **W:** www.ccia.org.au | **T:** @markjcowley

Does Mark Have an Interoperability Problem?

Dear Adam,
We have a diagnostic dilemma that I hope you could help us with?

The case has been challenging to diagnose by histopathology and also molecularly due to low tumour purity. The patient has a left thalamic/midbrain lesion and it is unclear whether it is a low grade or high grade glioma (which dictates the treatment the patient will receive). She has the canonical BRAF:p.V600E mutation which would be a clear driver in the tumour but of less certainty is a H3F3B G35W variant in the tumour (NM_005324(H3F3B):c.103G>T (p.Gly35Trp) - which if deemed pathogenic would bump up the grade). The MNP classifier failed to classify the tumour (maybe due to low purity), but was confidently MGMT methylated.

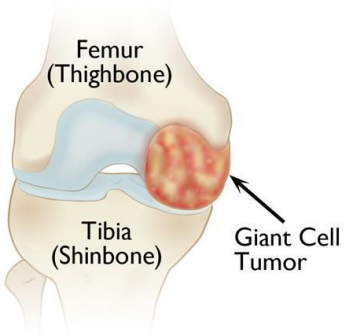
Literature supports H3F3A G34W in Giant cell tumours of bone (GCTB), but not in brain tumours. We've never seen H3F3A G34W, but we have seen G34R four times, all reported as pathogenic. All the G34R's had MGMT methylation, an association reported in the literature (28966033, 25752754), and thus pushing us towards saying the variant is pathogenic.

Have you seen this and can make a comment?

Thanks,
Mark

Mark Cowley, PhD BSc (Bioinf, Hons 1)
Computational Biology Group Leader
Conjoint Associate Professor, School of Women's and Children's Health, UNSW Medicine

Children's Cancer Institute
Lowy Cancer Research Centre, UNSW Australia
PO Box 81 Randwick 2031 Australia
P: 02 9385 2074 | **M:** 0413 481 017 | **E:** MCowley@ccia.org.au | **W:** www.ccia.org.au | **T:** @markjcowley



Does Mark Have an Interoperability Problem?

Dear Adam,
We have a diagnostic dilemma that I hope you could help us with?

The case has been challenging to diagnose by histopathology and also molecularly due to low tumour purity. The patient has a left thalamic/midbrain lesion and it is unclear whether it is a low grade or high grade glioma (which dictates the treatment the patient will receive). She has the canonical BRAF:p.V600E mutation which would be a clear driver in the tumour but of less certainty is a H3F3B G35W variant in the tumour (NM_005324(H3F3B):c.103G>T (p.Gly35Trp) - which if deemed pathogenic would bump up the grade). The MNP classifier failed to classify the tumour (maybe due to low purity), but was confidently MGMT methylated.

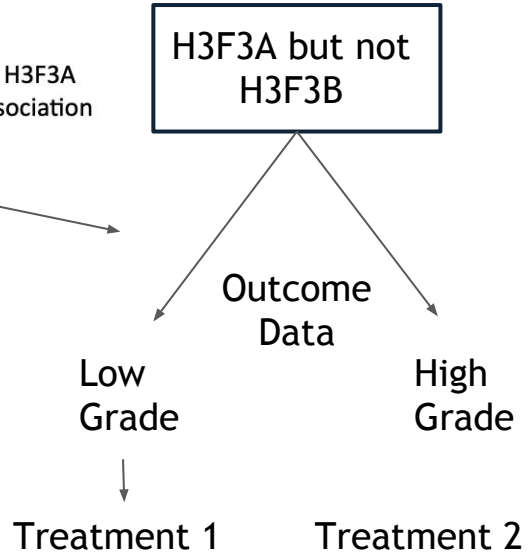
Literature supports H3F3A G34W in Giant cell tumours of bone (GCTB), but not in brain tumours. We've never seen H3F3A G34W, but we have seen G34R four times, all reported as pathogenic. All the G34R's had MGMT methylation, an association reported in the literature (28966033, 25752754), and thus pushing us towards saying the variant is pathogenic.

Have you seen this and can make a comment?

Thanks,
Mark

Mark Cowley, PhD BSc (Bioinf, Hons 1)
Computational Biology Group Leader
Conjoint Associate Professor, School of Women's and Children's Health, UNSW Medicine

Children's Cancer Institute
Lowy Cancer Research Centre, UNSW Australia
PO Box 81 Randwick 2031 Australia
P: 02 9385 2074 | **M:** 0413 481 017 | **E:** MCowley@ccia.org.au | **W:** www.ccia.org.au | **T:** @markjcowley



Does Mark Have an Interoperability Problem?

Dear Adam,
We have a diagnostic dilemma that I hope you could help us with?

The case has been challenging to diagnose by histopathology and also molecularly due to low tumour purity. The patient has a left thalamic/midbrain lesion and it is unclear whether it is a low grade or high grade glioma (which dictates the treatment the patient will receive). She has the canonical BRAF:p.V600E mutation which would be a clear driver in the tumour but of less certainty is a H3F3B G35W variant in the tumour (NM_005324(H3F3B):c.103G>T (p.Gly35Trp) - which if deemed pathogenic would bump up the grade). The MNP classifier failed to classify the tumour (maybe due to low purity), but was confidently MGMT methylated.

Literature supports H3F3A G34W in Giant cell tumours of bone (GCTB), but not in brain tumours. We've never seen H3F3A G34W, but we have seen G34R four times, all reported as pathogenic. All the G34R's had MGMT methylation, an association reported in the literature (28966033, 25752754), and thus pushing us towards saying the variant is pathogenic.

Have you seen this and can make a comment?

Thanks,
Mark

Mark Cowley, PhD BSc (Bioinf, Hons 1)
Computational Biology Group Leader
Conjoint Associate Professor, School of Women's and Children's Health, UNSW Medicine

Children's Cancer Institute
Lowy Cancer Research Centre, UNSW Australia
PO Box 81 Randwick 2031 Australia
P: 02 9385 2074 | **M:** 0413 481 017 | **E:** MCowley@ccia.org.au | **W:** www.ccia.org.au | **T:** @markjcowley

jmb
Journal of Molecular Biology
Volume 430, Issue 11, 25 May 2018, Pages 1562-1565

Brevia

Histone H3.3 G34 Mutations Alter Histone H3K36 and H3K27 Methylation *In Cis*

Leilei Shi [1,†], Jiejun Shi [2,†], Xiaobing Shi [1], Wei Li [2], Hong Wen [1,3]

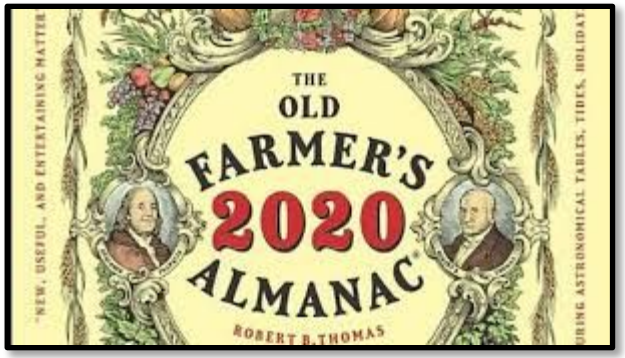
Show more

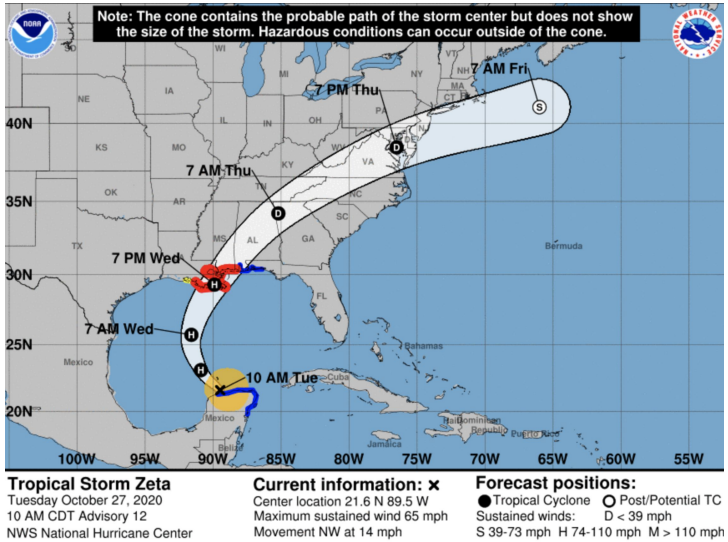
https://doi.org/10.1016/j.jmb.2018.04.014

Highlights

- Giant cell tumors of the bone (GCTB) H3.3G34 mutations (G34L/W) only affect histone H3K36 and H3K27 methylation on the same mutated histone tails (*in cis*).

Almanacs versus Weather Forecasts





Tropical Storm Zeta
Tuesday October 27, 2020
10 AM CDT Advisory 12
NWS National Hurricane Center

Current information: ×
Center location 21.6 N 89.5 W
Maximum sustained wind 65 mph
Movement NW at 14 mph

Forecast positions:
● Tropical Cyclone ○ Post/Potential TC
Sustained winds: D < 39 mph
S 39-73 mph H 74-110 mph M > 110 mph

Almanacs versus Weather Forecasts

Dow -1.27%	Nasdaq -2.27%
26,321.00 / -338.11	10,931.91 / -253.68

Most Popular Stocks

Apple Inc	109.54	-4.37%
Citigroup Inc	41.00	-0.36%
General Electric Co	7.50	1.76%
Alphabet Inc	1,627.97	4.68%
Microsoft Corp	200.96	-1.66%

Updated: 10:29:18am ET

Key Stats

10-year yield	0.84%	+0.01
Oil	$35.41	-2.10
Yen	¥104.58	+0.02
Euro	$1.17	-0.00
Gold	$1,881.10	+0.70%

FAIR VERUS IF-AR

IF YOU CAN INTEROPERATE THEN YOU CAN FIND, ACCESS AND EVENTUALLY REPRODUCE

Knowledbases versus a Commons

WHAT AND HOW vs WHERE WHEN

Knowledbases versus a Commons

CAVATICA

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

PedcBioPortal
for Integrated
Childhood Cancer Genomics

AACR
American Association for Cancer Research
FINDING CURES TOGETHER

PROJECTGENIE
Genomics Evidence Neoplasia Information Exchange

SYNAPSE

UCSF
University of California San Francisco

2 PATIENTS -- OUT OF 80,000

How will we know when we succeed?

WHEN USERS TALK ABOUT **FAR OUT** DATA

Morning Session Key Messages

1. **Awesome, impactful, accelerated** science can *actually* happen by harnessing the multi-platform cloud setting!

2. Both "expert" users and "new" users are able to leverage the advantages of cloud platforms when supported.

3. Users still face "binaries" in decision making that limit their full potential for harnessing platforms/cloud:

    a. Costs/platforms→ On Prem vs. Cloud (and which cloud?), where and from whom do I have my credits, how do I support "other" data (see b.) -- help with cost optimization.

    b. Terra vs. SBG vs. ISB vs. "X" →

        i. What data do I have to move where since I not only am accessing multiply hosted datasets, **but have some of my own data, own cohorts, or other existing studies that I need to intersect with the cloud-based cohorts (relates to the multiple cohort creation processes users will engage when navigating interop)**.

    c. CWL vs. WDL → where should I either invest in transforming my pipelines or are the "right" combinations of multiple pipelines available? Is there a way not to be "locked in" by this?

Experience Analyzing Human Genomes on the Cloud

Harrison Brand
Assistant Professor in Neurology
MGH, Harvard Medical School, & Broad Institute

Introduction

PhD in Human Genetics from the University of Pittsburgh (Advisors: Drs. Eleanor Feingold and Brenda Diergaarde)

- Focus in Statistical Genetics

Postdoc at Center for Genomic Medicine at MGH, Harvard Medical School, and Broad Institute (Advisor: Dr. Michael Talkowski)

- Applied novel WGS techniques to better detect structural variation (SV) in the human genome

Assistant Professor in the Department of Neurology at MGH, Harvard Medical School

- Assessing the impact of SV across a wide range of complex disorders
- Leading pipeline development and disease association studies in the Broad SV group

Experience with Relevant Platforms

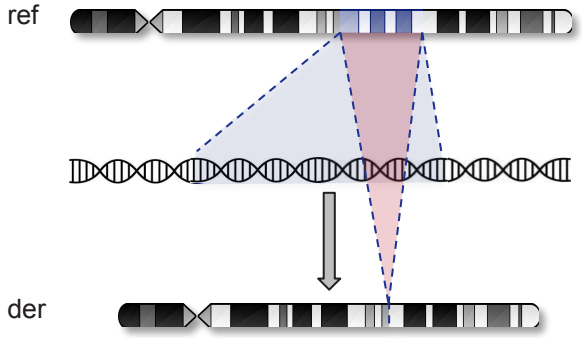
- NHLBI Biodata Catalyst – Fellow working on SV in Type 2 Diabetes and Glycemic Traits
- NHRGI's Analysis, Visualization, and Informatics Lab-space (AnVIL) – Member of the Broad CCDG and CMG teams
- Kids First Data Resource Center (KFDRC) - Member of the Broad GMFK Sequencing & Analysis Team. Part of several GMKF disease specific working groups
- Simons Simplex Collection – Member of Autism Sequencing Consortia
- The Genome Aggregation Database (gnomAD) – SV group

SV Background

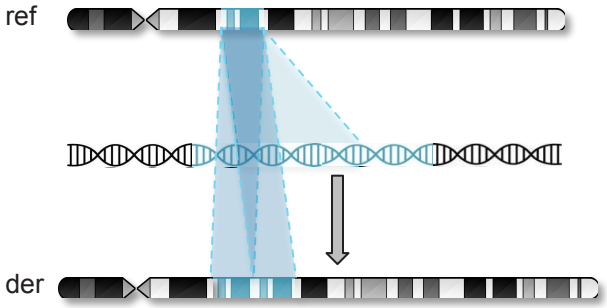
Structural Variation

Four basic classes of structural variation (SV) in the human genome

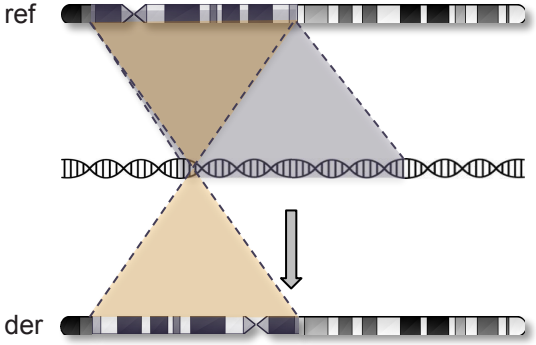
Deletion



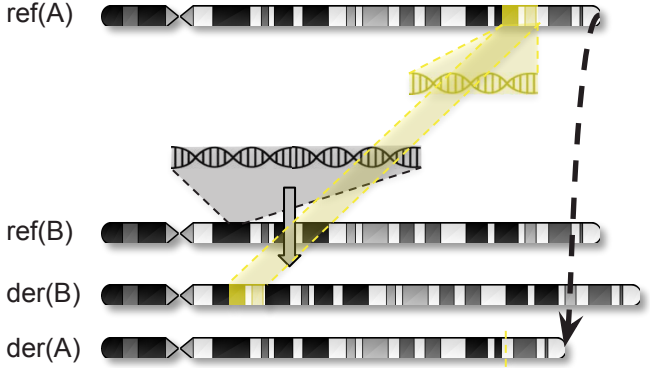
Duplication



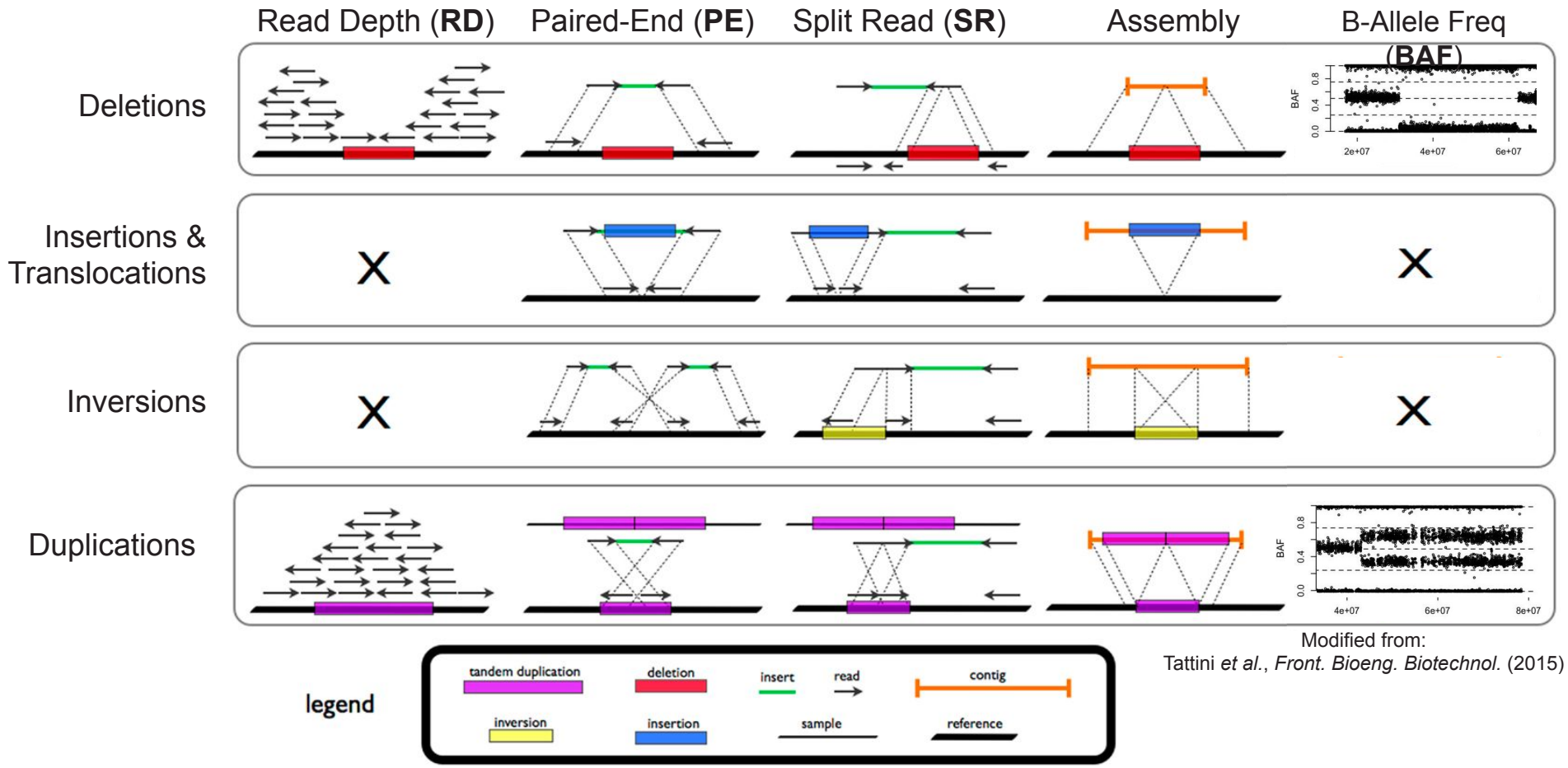
Inversion



Insertion/Translocation



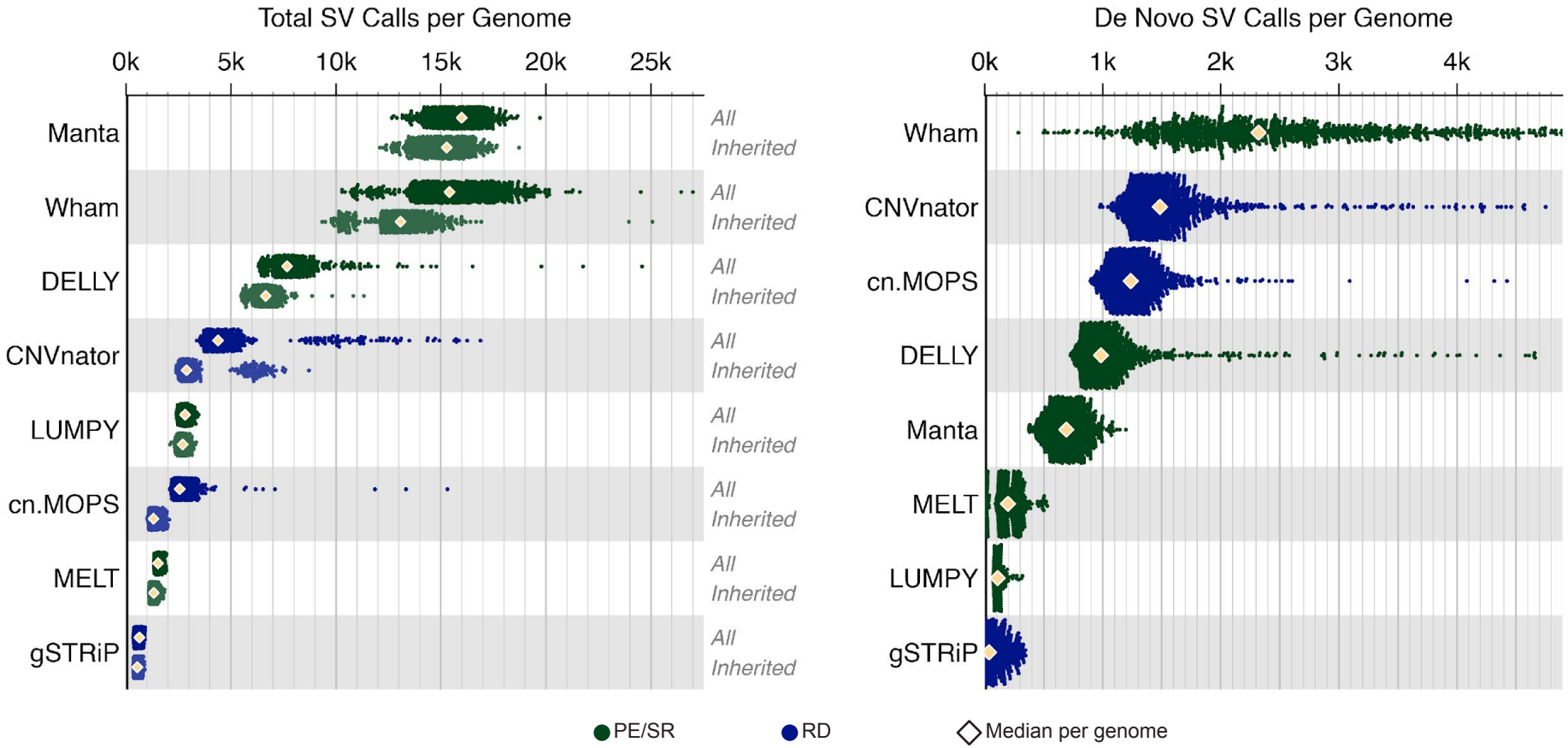
SV Discovery in Whole Genome Sequencing



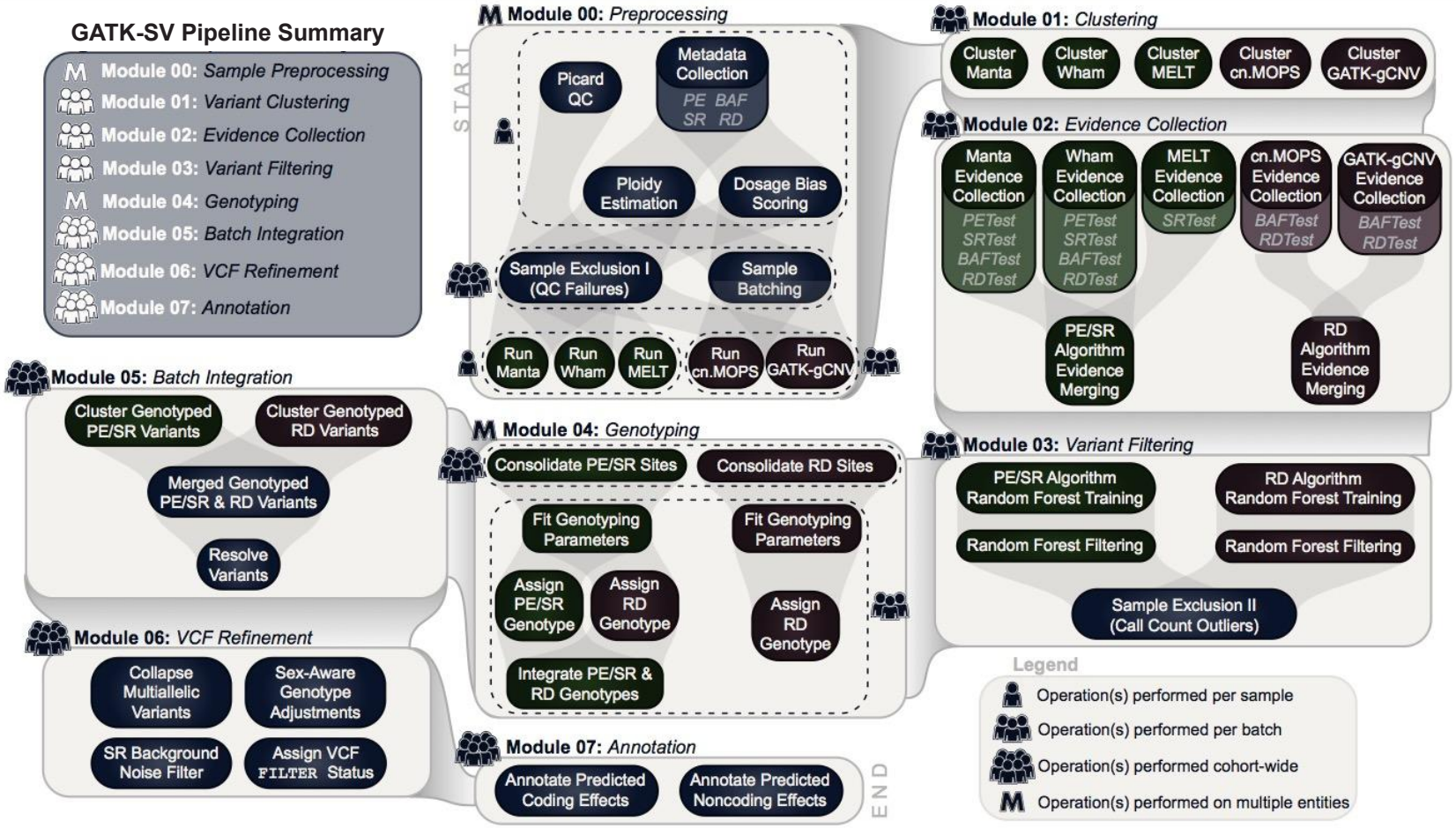
Modified from:
Tattini *et al.*, *Front. Bioeng. Biotechnol.* (2015)

Many SV Algorithms, but No Silver Bullet

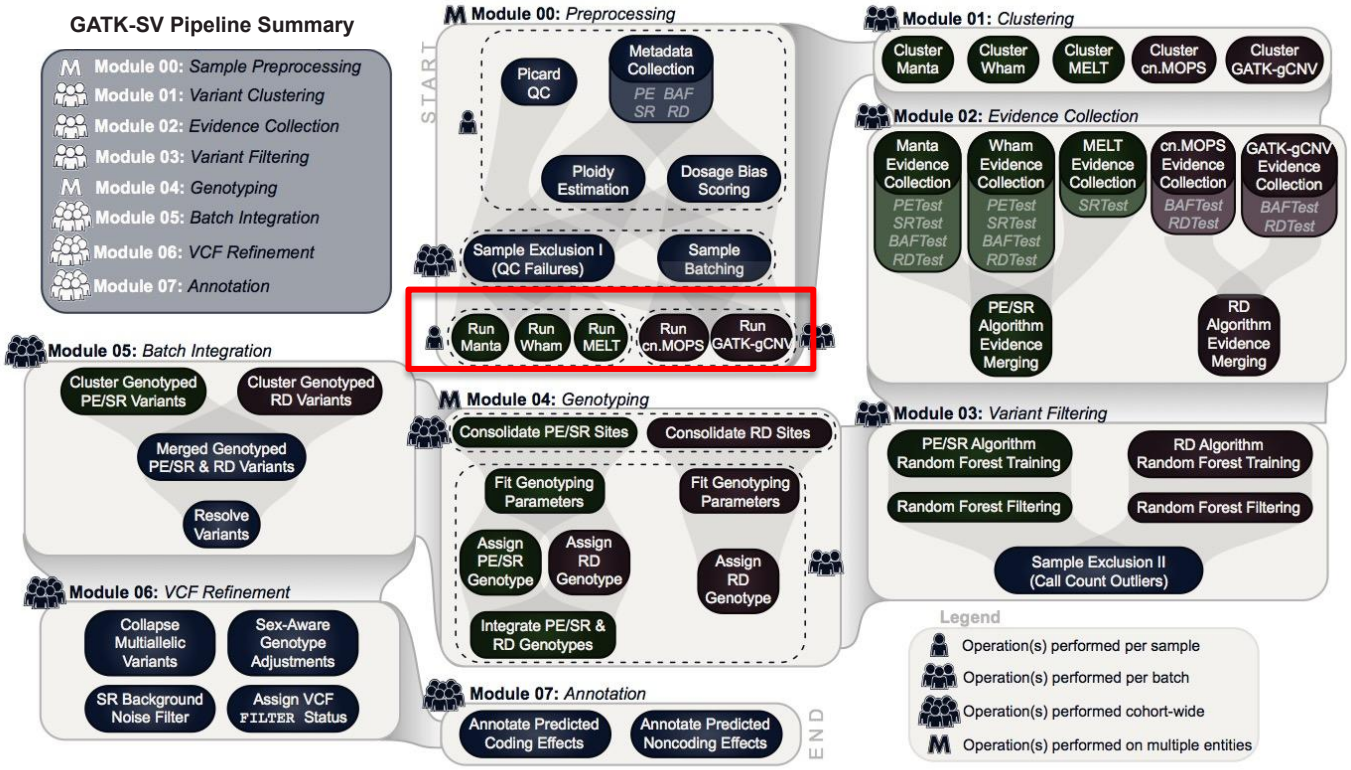
Raw algorithms yield >200-fold more *de novo* SV than expected (~0.2/genome)



GATK-SV: Cloud Enabled SV Pipeline

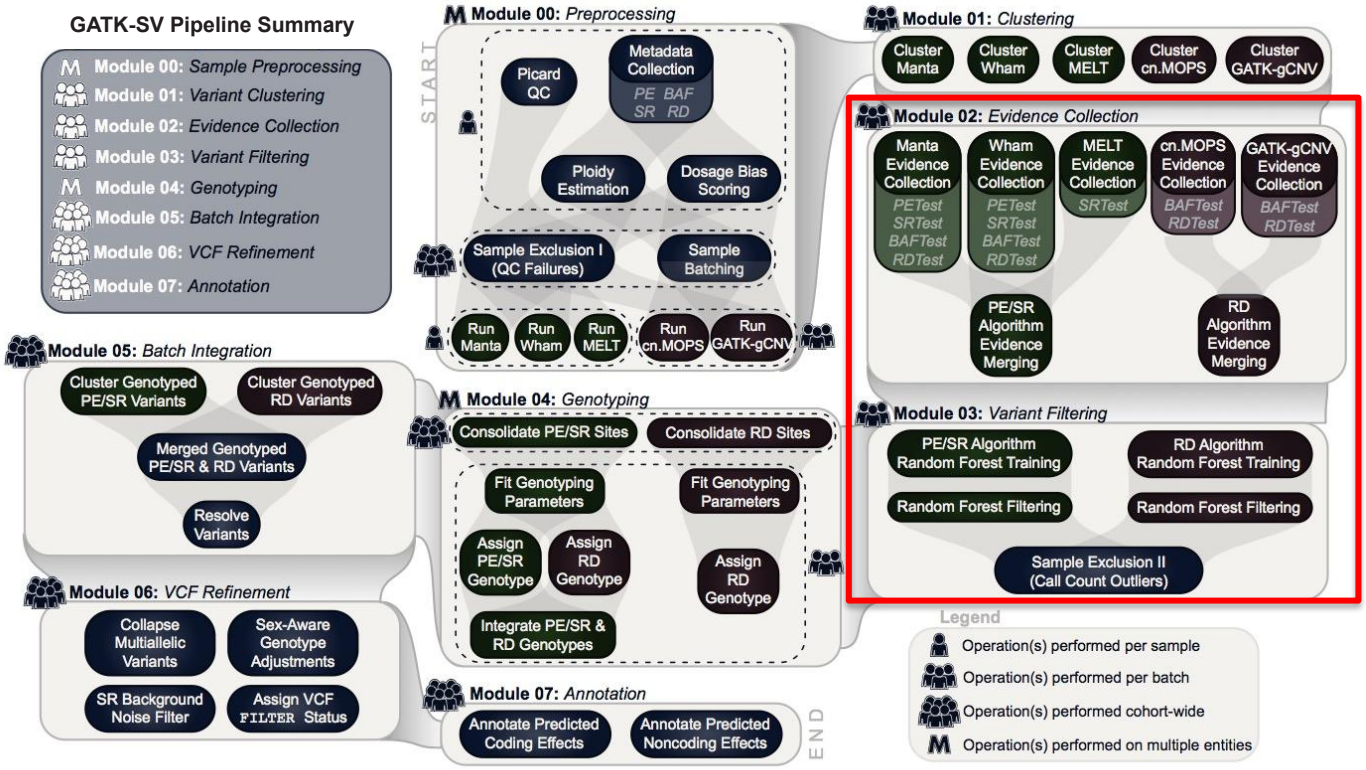


GATK-SV: Cloud Enabled SV Pipeline



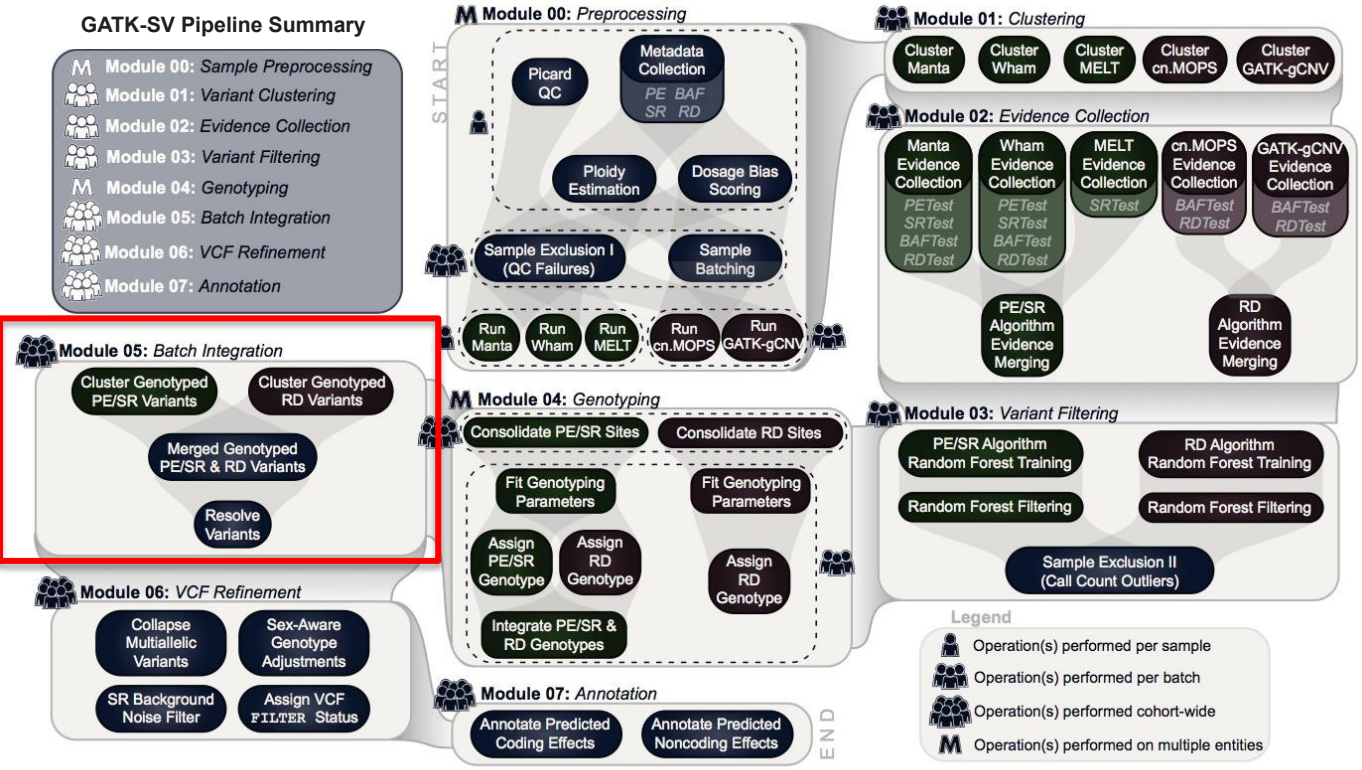
- Run several unfiltered algorithms to **maximize sensitivity**
- Re-evaluate evidence directly from BAMs to improve specificity
- Captures both unbalanced (CNV) and balanced (inversion, translocation) SV
- Integrates SV signatures to resolve complex events
- Modular design provides flexibility for improvements

GATK-SV: Cloud Enabled SV Pipeline



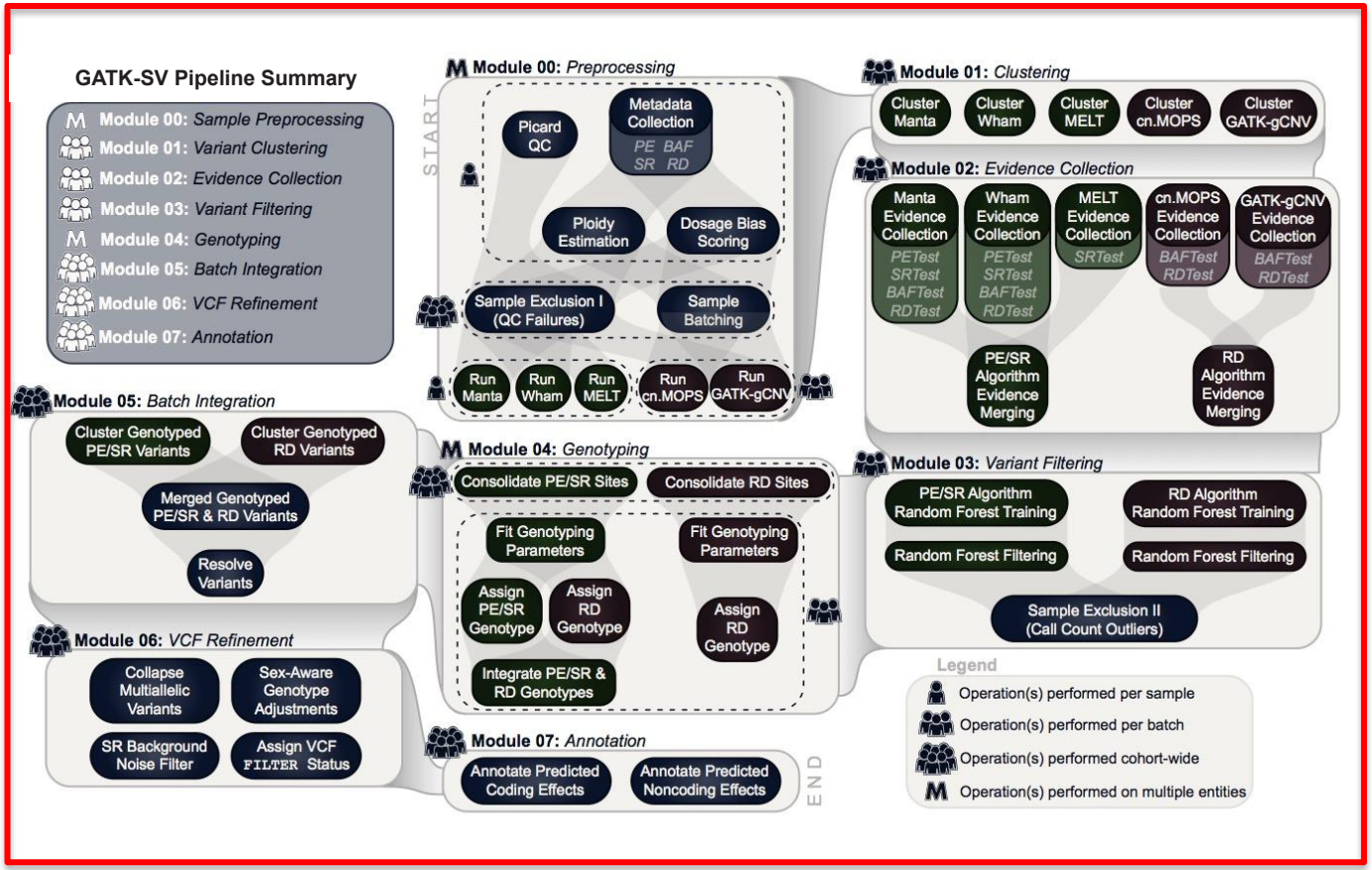
- Run several unfiltered algorithms to maximize sensitivity
- Re-evaluate evidence directly from BAMs to **improve specificity**
- Captures both unbalanced (CNV) and balanced (inversion, translocation) SV
- Integrates SV signatures to resolve complex events
- Modular design provides flexibility for improvements

GATK-SV: Cloud Enabled SV Pipeline



- Run several unfiltered algorithms to maximize sensitivity
- Re-evaluate evidence directly from BAMs to improve specificity
- Captures both unbalanced (CNV) and balanced (inversion, translocation) SV
- Integrates SV signatures to resolve complex events
- Modular design provides flexibility for improvements

GATK-SV: Cloud Enabled SV Pipeline



- Run several unfiltered algorithms to maximize sensitivity
- Re-evaluate evidence directly from BAMs to improve specificity
- Captures both unbalanced (CNV) and balanced (inversion, translocation) SV
- Integrates SV signatures to resolve complex events
- Modular design provides flexibility for improvements

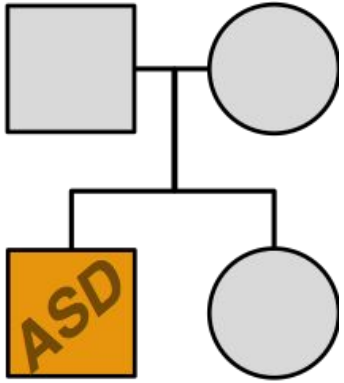
Experience in the Cloud

https://innovationatwork.ieee.org

First Experience

- Pilot study involving 40 Autism Spectrum Disorder (ASD) families (n = 160) from SFARI
- Data hosted on AWS
- Pulled down BAMs to local computing cluster ~16 TB
- Ran SV detection locally
- Quickly realized the challenge of handling WGS on local computing cluster

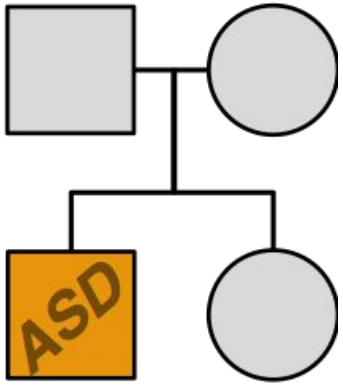
SFARI
SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE



Hybrid Approach

- Phase 1 increased to 519 families (n = 2,076) from SFARI
- Raw algorithms run on AWS
- Lots of issues with cloud stability
- Pulled down raw SV VCFs to local computing cluster
- Ran SV pipeline on local compute cluster

SFARI
SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

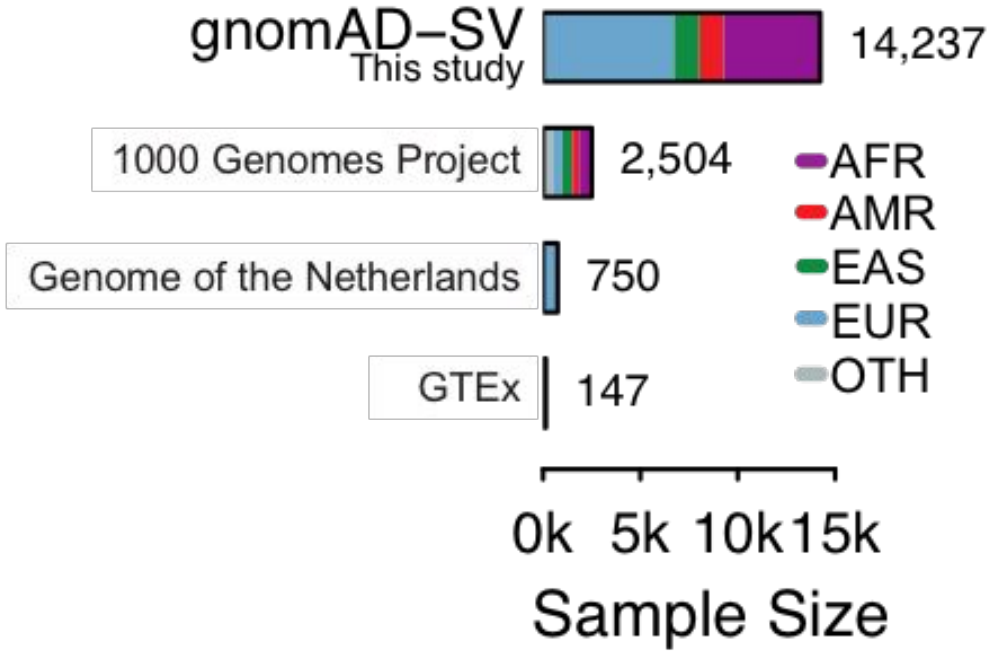


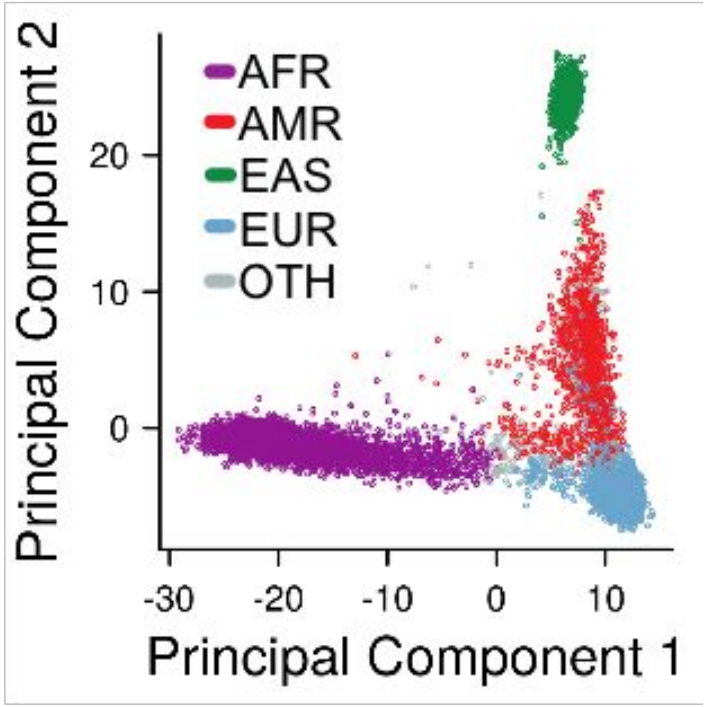
The Value of Population Variation References

Variant Class	*Current Gold-Standard Reference*	*Advances Catalyzed*
SNVs **InDels**	**ExAC** (60,706 exomes) ARTICLE OPEN doi:10.1038/nature19057 Analysis of protein-coding genetic variation in 60,706 humans **gnomAD** (125,748 exomes + 15,708 genomes) New Results Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes	• Improved understanding of human demography • Mutational constraint • Refined clinical interpretation • Power for disease association • Frequency filter for rare diseases • Human “knockout” identification
SVs	**1000 Genomes Project** (2,504 low-coverage genomes) ARTICLE OPEN doi:10.1038/nature15394 An integrated map of structural variation in 2,504 human genomes A list of authors and their affiliations appears at the end of the paper.	???

GnomAD-SV Dataset

Illumina WGS on 14,891 samples (14,237 passed quality control). Majority (54%) non-European.



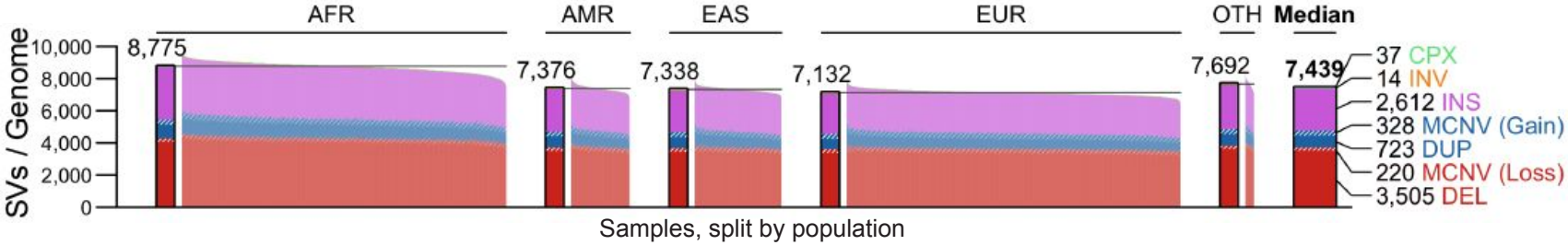


Collins, Brand* et al., Nature 2020*

Shift to Cloud

- Large sample in gnomAD necessitated compete shift to cloud
- Set up pipeline on google cloud (GCP) using firecloud/terra platform from the Broad Institute
- Processed and ran QC on all 15,000 samples

The average genome harbored 7.5k SVs



What I Have Learned using the Cloud for Genomics

My Experience - Benefits of the Cloud

- Data sharing
- Ability to massively parallelize due to incredible resources
- Reproducibility of code for groups outside one’s home institution
- Technical Support

What Terrifies Me?

- Financial issues
  - Cost tracking lag (24 hours)
  - Intermediate data file storage
  - Infrastructure changes that break code
  - Surprise preemptible VM bills

- Scalability issues
  - Making sure to run parallel jobs to optimize both time and cost
  - Cost monitoring

Challenges of Interoperability

GATK-SV has only been adapted for the Terra system on GCP

- Can’t directly access data on AWS without pulling to google cloud
- If adapted for AWS do I need to support two provide support for both AWS and GCP
- Resource optimization likely to differ between AWS and GCP

Conclusions

- I have helped build a cloud-based SV pipeline that has been applied on tens of thousands of samples
- These studies would not have been possible on a standard high-performance computing cluster
- The cloud holds great promise for sharing data and reduces barriers for reproducibility
- Cost tracking is still a little terrifying

Acknowledgements

Michael Talkowski

Ryan Collins

Daniel MacArthur

The Broad Institute

Talkowski Lab

Xuefang Zhao
Harold Wang
Chelsea Lowther
Jack Fu
Isaac Wong
Elise Valkanas
Isaac Wong
Matt Stone

gnomAD

Jessica Alföldi
Konrad Karczewski
Laurent Francioli
Mark Daly
Nick Watts
Matt Solomonson
Anne O'Donnell
Grace Tiao

Broad-SV Team

Eric Banks
Laura Gauthier
Chris Whelan
Mark Walker
Ted Brookings
Emma Pierce-Hoffman
Ted Sharpe
Steve Huang
Samuel Lee
Andrey Smirnov

Proof of concept of interoperable approaches for improving outcomes of pediatric diseases.

Alisa Manning
Assistant Investigator, Massachusetts General Hospital
Instructor, Harvard Medical School

Tim Majarian
Computational Biologist, Broad Institute

Background: Using the Cloud for Complex Trait Genetics Analysis

2017 - 2018:
First researchers to perform a GWAS using FireCloud

TOPMed Diabetes working group
- Genome-wide association studies
- Rare variant association tests
- Writing our first WDLs
- Deploying our first cloud-based workflows

TOPMed Cloud Computing Pilots
- FireCloud

2018 - 2019:
Collaborative Development of Cloud-based Workflows

Rare variant analysis workflows:
- Collaboration on github
- Analysis Commons hosted by DNANexus
- TOPMed Diabetes working group analysis on Terra

Large-scale Gene-environment Interaction
- Principle Investigator (MGH)
- Open-source statistical software tools
- WDL workflows
- WDLs in DockStore

User resources: GWAS in the cloud
- Featured Workspace in Terra
- Workshop at ASHG 2019

2020:
Collaborative analysis in NHLBI's BioData Catalyst

Biodata Catalyst
- Principle Investigator (Broad Institute)

Biodata Catalyst - Fellows Cohort 1
- Postdoc with Gene-environment Interaction study including TOPMed WGS and 'Omics Data

CICI Interoperability Project
- Pilot process for cross-platform analysis

Genetics of CHD: improving outcomes of pediatric diseases

Study aims:

1. Identify, access, and summarize available genetic and phenotypic data on native cloud platforms
2. Leverage individual-level data from multiple cloud platforms to assess rare variants contributing to CHD risk

Framework:

Internal cases (KFDR CHD)
External controls (FHS/JHS)
Gene expression follow-up (GTEx)

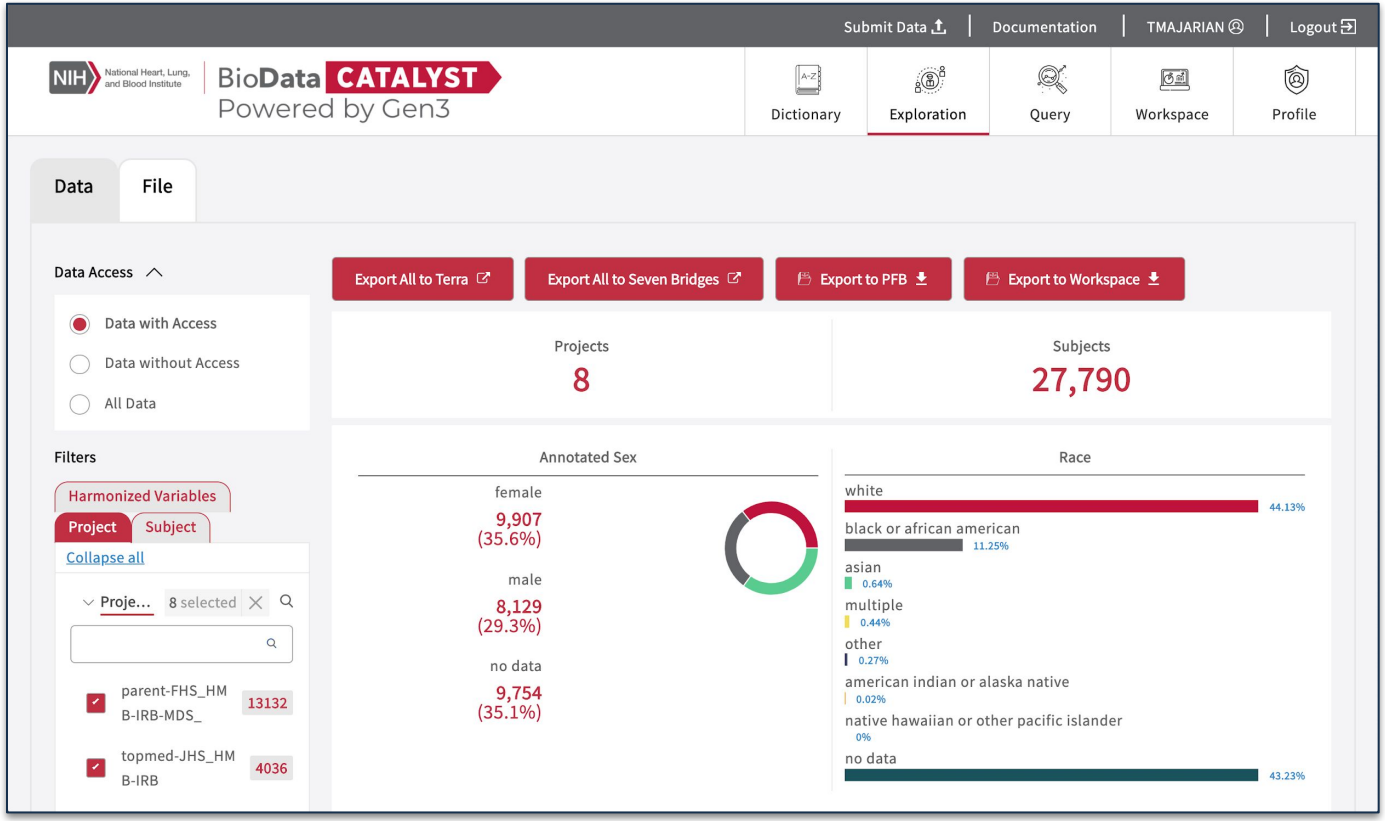
Method: Proxy External Controls Association Test (ProxECAT)

Compare ratio of rare, synonymous and nonsynonymous variants per gene between cases and controls

Platform	Datasets	dbGaP	Sample	Use
AnVIL	GTEx	phs000424.v8.p2	980	Not used
Kids First	PCGC	phs001138.v3.p2	699	Case
BioData Catalyst	TOPMed PCGC	phs001735, phs001194.v2.p2	1,901	Not used
	FHS	phs000974.v4.p3, phs000007.v30.p11	4,155	Control
	JHS	phs000964.v4.p1	2,777	Control

Hendricks et al (2018). doi: 10.1371/journal.pgen.1007591

Export to native cloud platforms



Export to native cloud platforms

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

Dashboard | Explore Data | File Repository | Members | Resources | Timothy

Filter Browse All

Clinical Filters | File Filters

Study Name

	# FILES
[x] Kids First: Congenital Heart Defects	699

Diagnosis Category

	# FILES
[x] Structural Birth Defect	699
[ ] No Data	697

Diagnosis (Source Text)

	# FILES
[ ] Atrial septal defect, secundum	127
[ ] Tetralogy of Fallot	100
[ ] Right aortic arch with mirror image branching pattern	89
[ ] Hypoplastic left heart	63

CLEAR **FILE FORMAT** is vcf and **HARMONIZED DATA** is true and **DIAGNOSIS CATEGORY** is Structural Birth Defect and **FAMILY SHARED DATA TYPES** is Variant Calls and **PARTICIPANTS ID** is Uploaded List and **STUDY NAME** is Kids First: Congenital Hear... and **EXPERIMENT STRATEGY** is WGS

share | Save

699 Files | 2,096 Participants | 699 Families | 698.23 GB Size

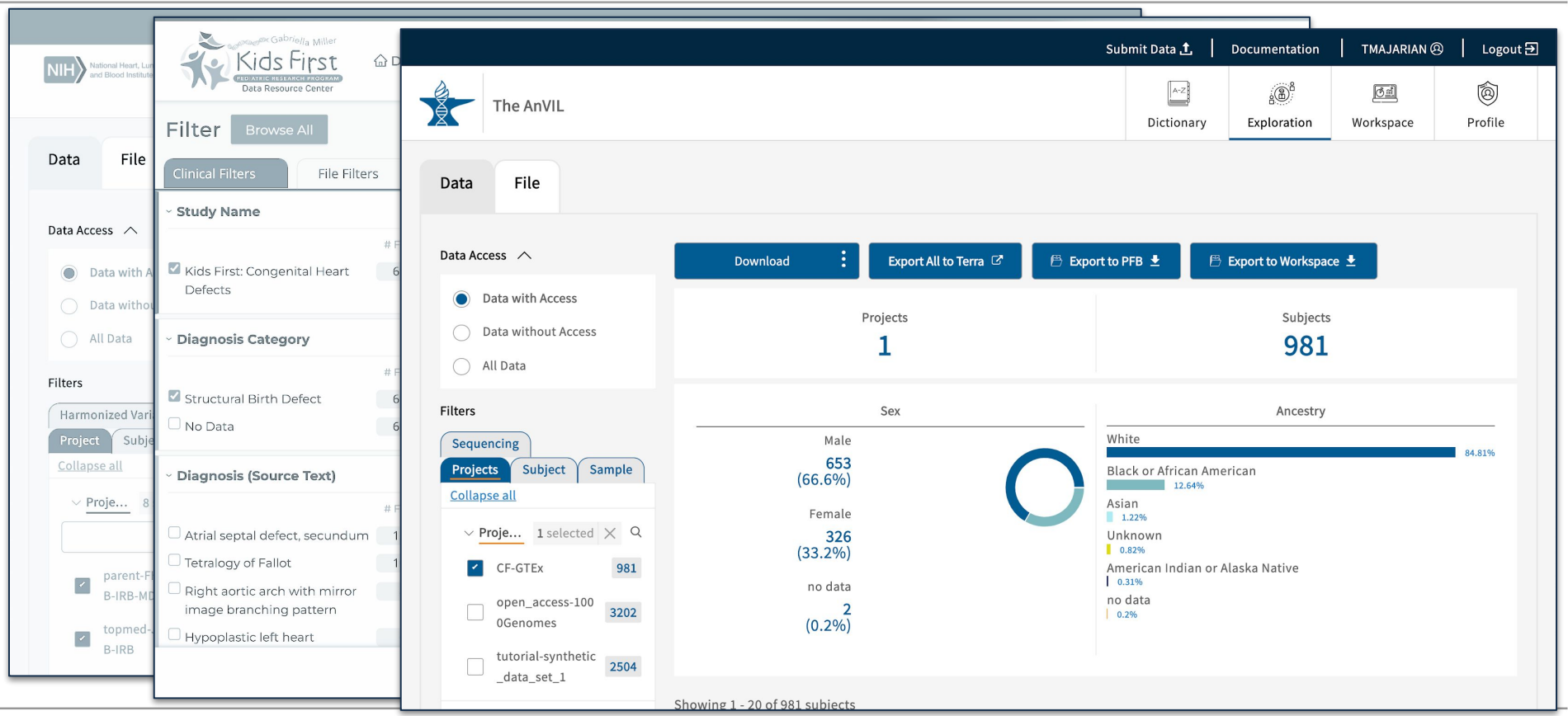
Showing 1 - 20 of 699 files | ANALYZE IN CAVATICA | Download | File Manifest | Columns | Export TSV

File ID	Participant...	Study Name	Proband	Family Id	Data Type	File Format	File Size	Actions
GF_8NRDWD...	PT_BWPJWA...	Kids First: Congenital Heart Defects ...	Yes, No, No	FM_HMNBRF... ...	Variant Calls	vcf	1009.66 MB	
GF_TJRD7P4H	PT_DWBNND...	Kids First: Congenital Heart Defects	No, No, Yes	FM_8MMCZC... ...	Variant Calls	vcf	1.23 GB	

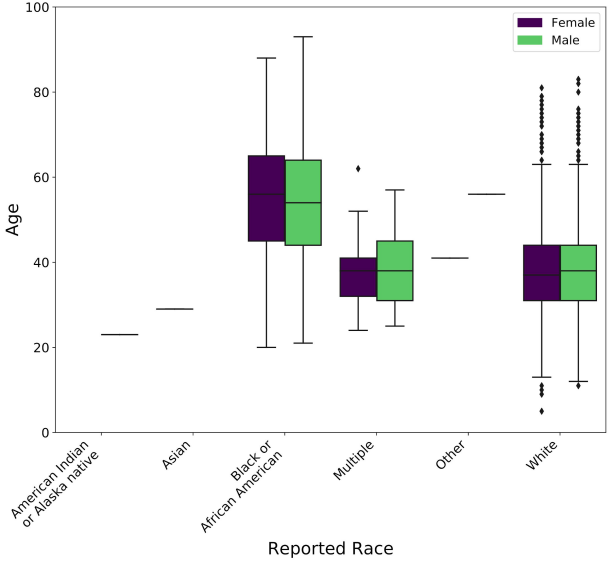
Show 20 rows | << < 1 2 3 4 5 > >>

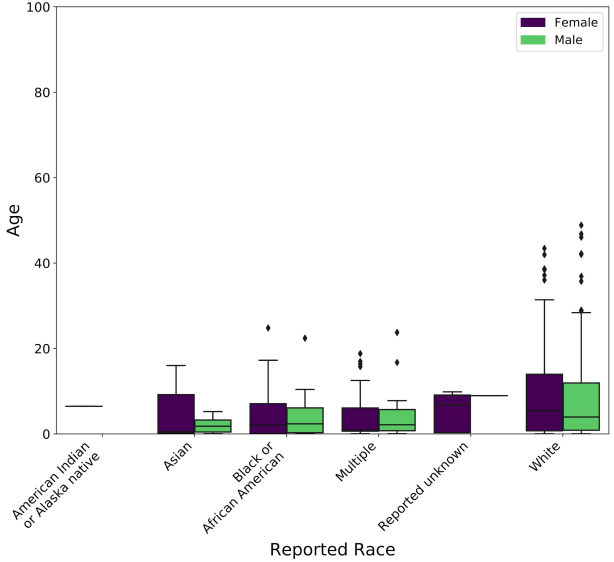
kidsfirstdrc.org | About the Portal | Policies | Support | Contact | UI: 2.26.1, Data Release: 5.42.0 | Follow Us

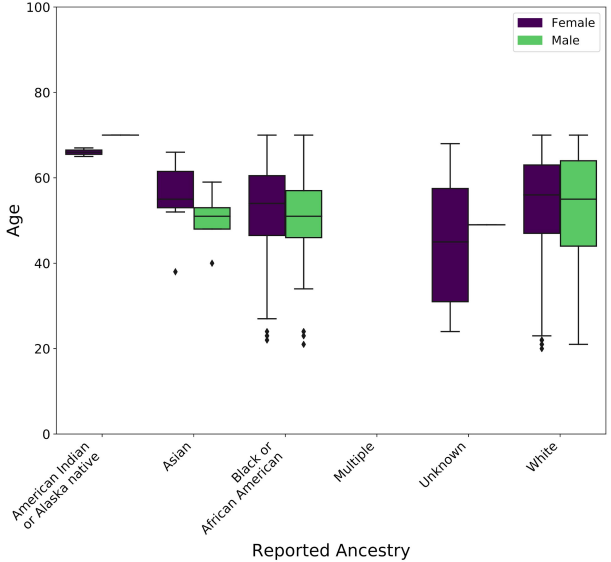
Export to native cloud platforms



Platform-specific summaries







Preparation of genetic data for association analysis

All preparation steps were performed within separate ecosystems

1. **KFDR - Cavatica**
2. **BioData Catalyst - Terra**
3. **AnVIL - Terra**

Variants included in analysis:

- MAF < 1%
- Protein coding exonic

Variant annotation - Synonymous and non-synonymous

- ANN field in VCF files for KFDR
- DBSNFP for JHS and FHS

For each protein coding gene

- Count synonymous and non-synonymous variants
- Separated by cases (KFDR) and controls (JHS and FHS)

ANN: *annotation* field

- Predicted variant effect on gene expression or protein function

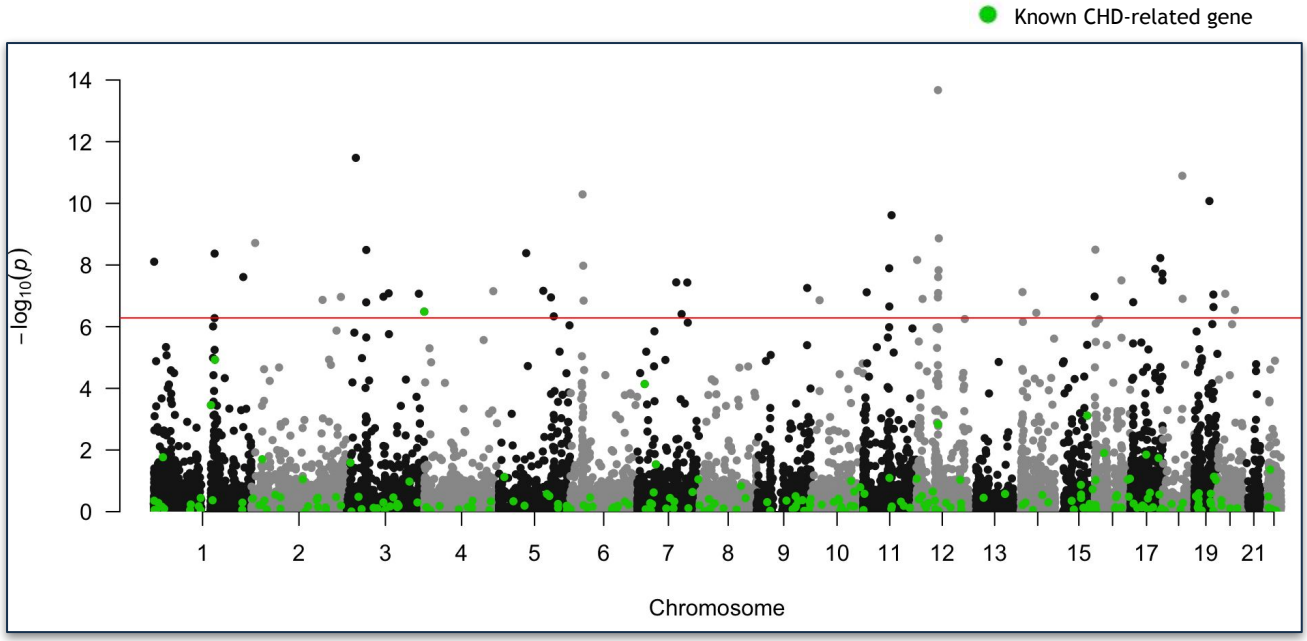
DBSNFP:

- Database of functional predictions for all coding variants
- Includes same variant effect predictions as ANN field

ProxECAT results

Association analyses were performed within the BioData Catalyst ecosystem
KFDR data was manually downloaded and uploaded to a BDC workspace

- 17,285 genes tested
- 55 genes with P<5e-7
- 1 known CHD gene with P<5e-7



Then vs Now vs Future

Pre-interoperability effort

Data authorization
- Obtain dbGaP access
- Log into dbGaP
- Create download request

Access and localization to cloud platform
- Start GCS VM
- Download data via Aspera
- Upload data to GCS bucket
- Access through Terra workspace

Data preprocessing & Final analysis
- Single Terra workspace

Current paradigms

Data authorization
- Obtain dbGaP access

Access and localization to cloud platform
- ERA credentials through Gen3 or KFDR
- Export data links (DRS) within a individual ecosystems

Data preprocessing
- Separate workspaces within individual ecosystems

Final analysis
- Single BDC workspace
- Download & upload KFDR data for analysis

Future

Data authorization
- Obtain dbGaP access

Access and localization to cloud platform
- Single sign in within a BDC ecosystem

Data preprocessing
- One BDC workspace for all data

Final analysis
- One BDC workspace
- No download and upload

Stumbles and roadblocks

Data availability across platforms - KFDR (Cavatica) to BDC (Terra)

PFB import to Terra - TOPMed PCGC (BDC) [**SOLVED**]

DRS links - GTEx (AnVIL) [**SOLVED**]

Workflow compatibility - CWL (Cavatica) vs. WDL (Terra)

Data documentation: Data are easy to access but finding exactly how the data were generated remains difficult

- Ex: Why is the ANN field missing in the TOPMed cohort-level VCFs?
- Ex: What fields are included in genetics data and what do they mean?
- Ex: What methods were used for genotype calling? (KFDR vs. TOPMed)

Acknowledgements

Brian O'Connor
Asia Mieczkowska
Becky Boyles
Patrick Patton
Steven Cox
Michael Baumann
Andrew Rula
Alex Baumann
Allison Heath
David Higgins
Maia Nguyen

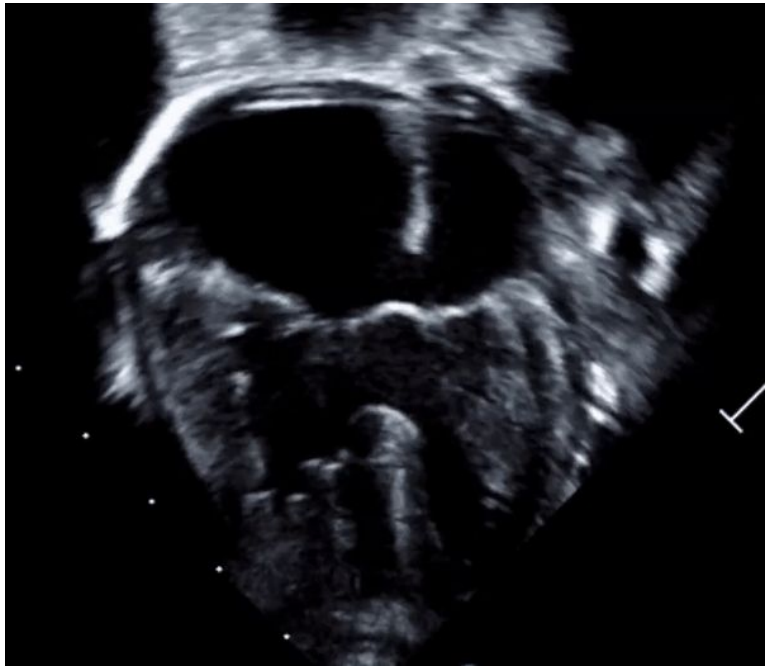
Gabriella Miller Kids First Pediatric Research Program of the Pediatric Cardiac Genetics Consortium (PCGC)
Pediatric Cardiac Genomics Consortium (PCGC)
Genotype-Tissue Expression (GTEx) project
TOPMed's PCGC's Congenital Heart Disease Biobank
Framingham Heart Study
Jackson Heart Study

Use of cloud computing to study structural variation in congenital heart disease

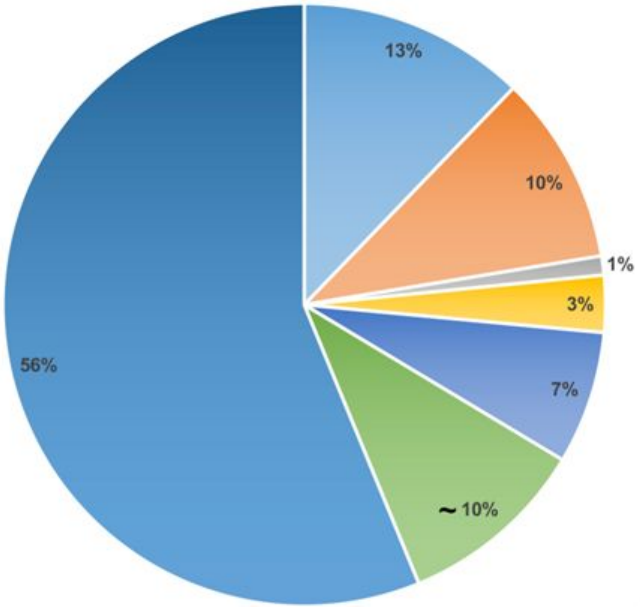
Daniel Quiat M.D., Ph.D
Attending in Cardiology - Boston Children's Hospital
Postdoctoral Fellow - Seidman Lab - Harvard Medical School

Congenital Heart Disease

- Congenital Heart Disease (CHD)
  - Most common congenital anomaly
    - 7-8/1000 live births
  - Leading cause of mortality due to a birth defect
  - Strong genetic basis
    - Association with genetic syndromes and chromosomal abnormalities



Genetics of CHD



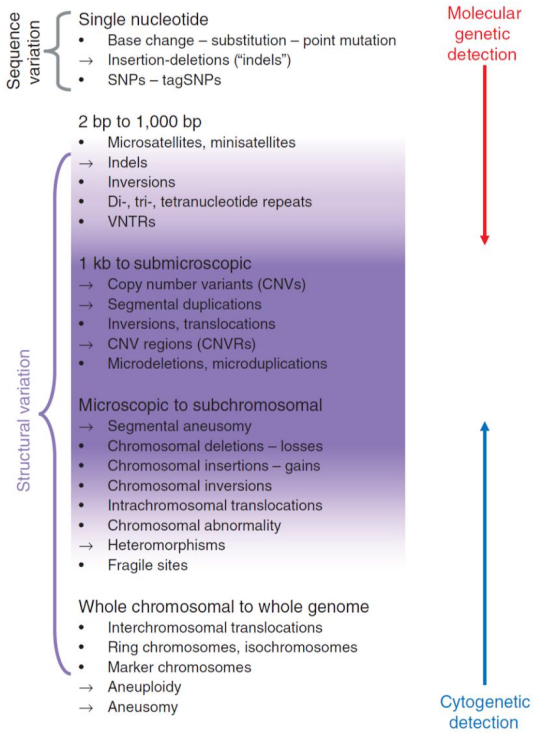
- **aneuploidy** (Hartman et el, *Pediatric Cardiology* 2011)
- **CNV** (Kim et al, *J Thorac Cardiovasc Surg*, 2016; Glessner et al, *Circ Res*, 2014)
- **known gene inherited**
- **de-novo chromatin SNV** (Zaidi et al, *Nature*, 2013)
- **other de-novo SNV** (Zaidi et al, *Nature*, 2013; Homsy et al, *Science*, 2015; Hitz et al, *Nat Gen*, 2016)
- **environmental** (Jenkins et al, *Circ*, 2007)
- **unknown**

Aim

Can we use WGS to identify previously undetected genetic variants responsible for CHD?

Pediatric Cardiac Genomics Consortium

Genomic structural variants as a class of undetected variation

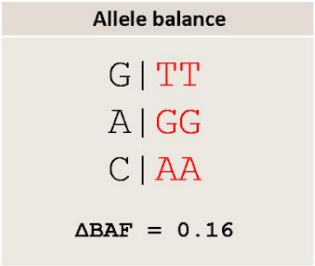


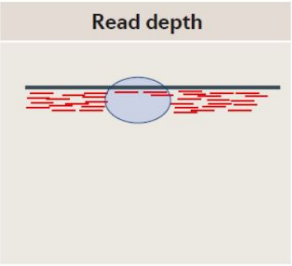
Structural variant (SV) – Any genetic change > 50 bp in size that alters the structure of the genome

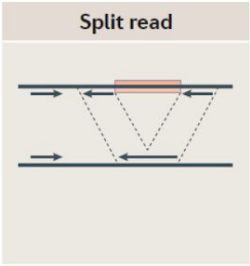
- Unbalanced: duplications, deletions, insertions
- Balanced: translocations, inversions

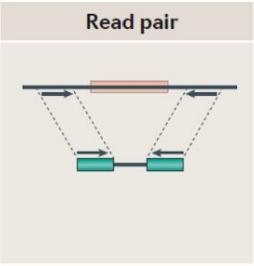
Scherer et al. Nature Genetics - Supplement 2007

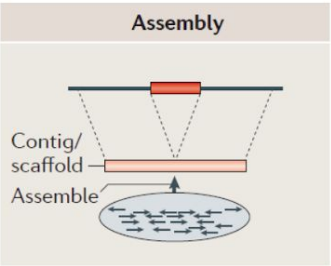
Detection of genomic SVs by WGS is resource intensive











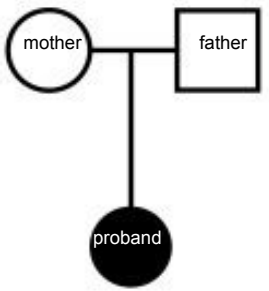
- Utilize multiple tools to collect a variety of evidence genome-wide
- Resource requirements pushed our group to consider computing in the cloud

Important factors that eased transition to the cloud

- Concerns about unknowns surrounding cost of analyses vs no additional cost associated with computing on HPC cluster
  - $$$ available for pilot studies
- Learning curve
  - User-friendly tool editors on Cavatica and help from Seven Bridges bioinformatics team when necessary
- Data availability
  - GMKF generated WGS data on Cavatica

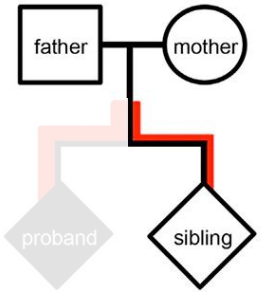
Experimental Approach

Cases



- **716 CHD trios**

Controls



- **1650 non-CHD 'trios'**

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

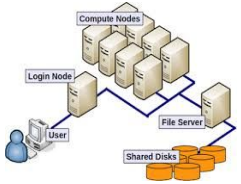
SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

Experimental Approach

Run SV tools
(SvABA, Manta, Delly, etc)

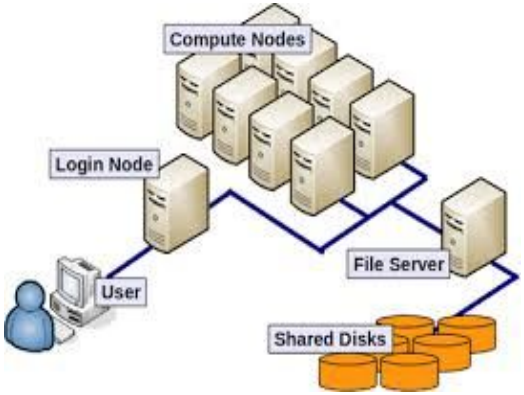
CHD Cases

Non-CHD Controls



HPC Cluster

Merging SV calls, joint genotyping of trios (SVTYPER), filtering



HPC Cluster

Initial Study Results

- **SV genotyping identifies pathogenic loss-of-function SVs in known CHD genes, and a burden of *de novo* loss-of-function variants in constrained genes**
  - **Example: Patients with tetralogy of Fallot harbor rare loss-of-function variants in genes associated with the diagnosis ranging from 57bp to 8kb in size**
    - ***TBX1, KDR, FLT1, NOTCH1***

Expansion of CHD WGS dataset and population level SV genotype data

- **892 trios sequenced by GMKF**
- **1067 trios sequenced by TOPMED**
- **Population level SV data from gnomAD-SV**

Current Approach

- **Genotype SVs in 1950+ CHD trios using GATK-SV in collaboration with Drs. Brand and Talkowski (ongoing)**

- **GMKF WGS data manually uploaded to Terra platform for this analysis**

Importance of Interoperability

- **PCGC Cohort split between two platforms**
  - **A problem for major analyses and minor tasks**

- **In addition to CHD, we are applying GATK-SV workflow in Terra to other cardiovascular and developmental datasets: TOPMED (cardiomyopathy) and GMKF (microtia)**

- **As our lab is starting to perform additional analyses in the cloud and location of workflows and datasets is a major considerations as we make this transition**

Positive experiences computing in cloud ecosystems

- **Acceleration of research through use of ‘on demand’ cloud compute resources**
- **Ease of data sharing**
  - **Access to more control WGS data**

Barriers encountered while computing in cloud ecosystems

- **Datasets of interest split between two platforms**
- **Difficulty estimating cost upfront / difficulty monitoring cost**
- **Expensive mistakes / backend errors**
- **WDL vs CWL, and lack of workflow portability**

Acknowledgements

Seidman Lab

Kricket and Jon Seidman

Sarah Morton

Steve DePalma

Jon Willcox

Alex Pereira

Josh Gorham

Alireza Haghighi

Barbara McDonough

BCH

Jane Newburger

Amy Roberts

Talkowski & Brand Labs

GATK-SV team

Mark Walker

National Heart Lung and Blood Institute

PCGC

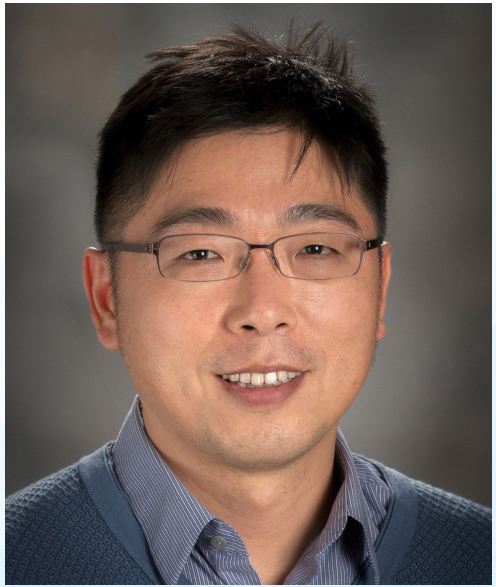
CVDC

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

HARVARD MEDICAL SCHOOL

Boston Children's Hospital

Studies on extrachromosomal DNA alterations using cloud computing over multiple tumor types



Hoon Kim
Senior Research Scientist
Jackson Laboratory

Our two studies made possible through the Cancer Genomics Cloud of the Institute for Systems Biology (ISB-CGC) and Amazon Web Service (AWS)

ARTICLES
https://doi.org/10.1038/s41588-018-0105-0
nature genetics

Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma

Whole-genome sequencing (WGS) from 53 TCGA-GBM & LGG samples

Ana, Kim* et al, 2018*

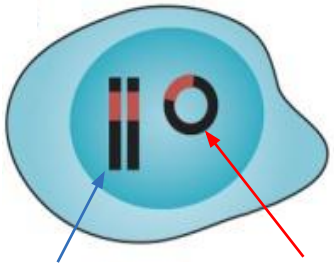
nature genetics
LETTERS
https://doi.org/10.1038/s41588-020-0678-2
Check for updates

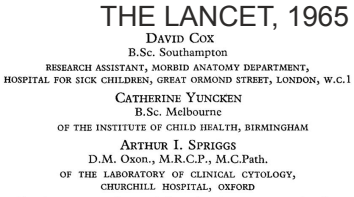
Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers

Whole-genome sequencing from >5000 samples (tumor & normal)

Kim et al, 2020

Extrachromosomal DNA (ecDNA) elements in cancer were first described in 1965

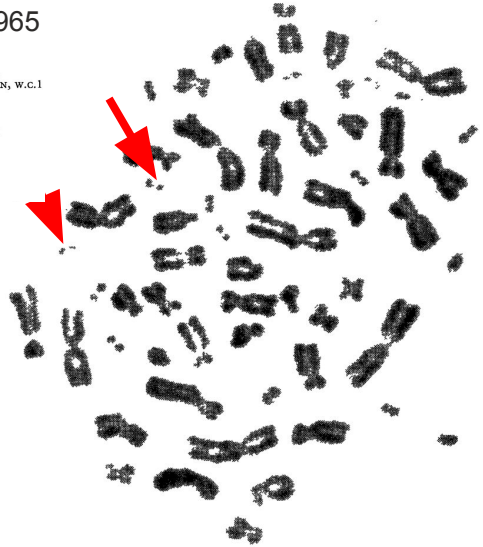



THE LANCET, 1965

DAVID COX
B.Sc. Southampton
RESEARCH ASSISTANT, MORBID ANATOMY DEPARTMENT,
HOSPITAL FOR SICK CHILDREN, GREAT ORMOND STREET, LONDON, W.C.1

CATHERINE YUNCKEN
B.Sc. Melbourne
OF THE INSTITUTE OF CHILD HEALTH, BIRMINGHAM

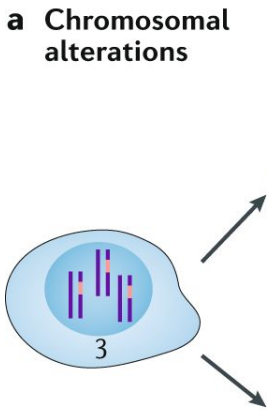
ARTHUR I. SPRIGGS
D.M. Oxon., M.R.C.P., M.C.Path.
OF THE LABORATORY OF CLINICAL CYTOLOGY,
CHURCHILL HOSPITAL, OXFORD



Metaphase chromosome spreads from neuroblastoma cell

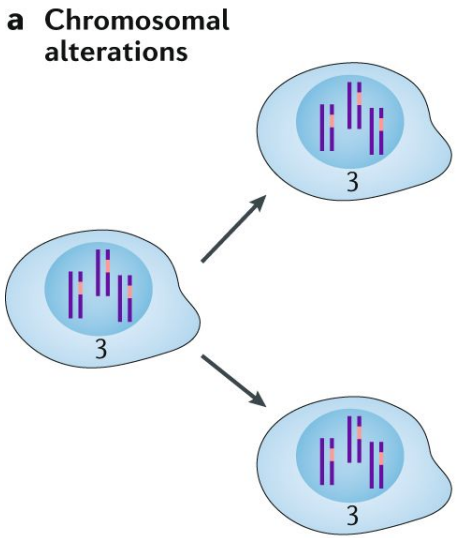
- Circular DNA
- Also referred to as “minute bodies” or “double minutes”
- **Previously**, it was reported to be in **only 1.4% of tumors** (Mitelman, 2007)

Uneven segregation of ecDNAs during cell division



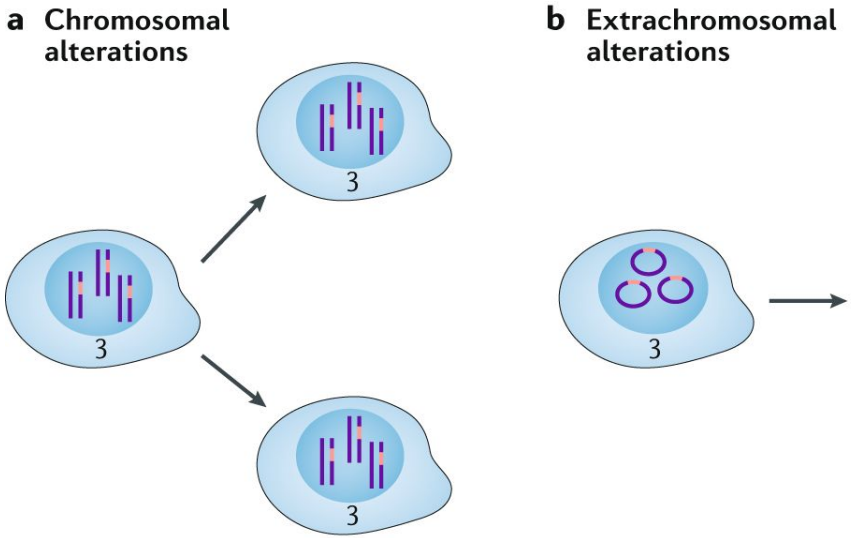
Verhaak et al, Nat Rev Cancer, 2019

Chromosomal alterations are equally segregated during cell division



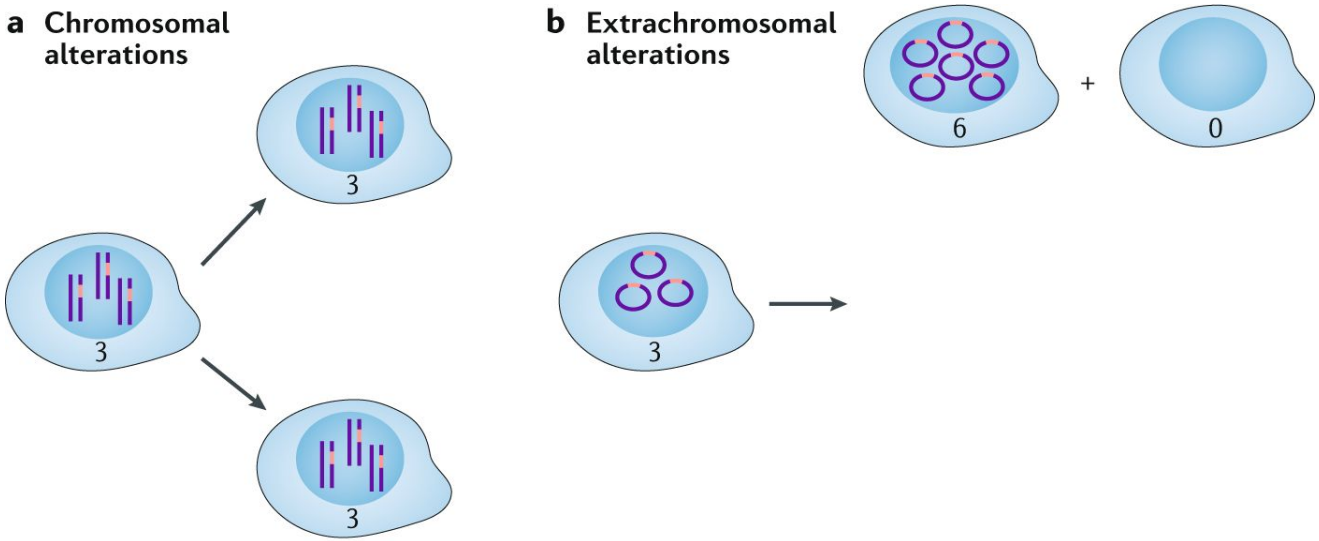
Verhaak et al, Nat Rev Cancer, 2019

The segregation patterns of ecDNAs during cell division are different from chromosomal DNA.



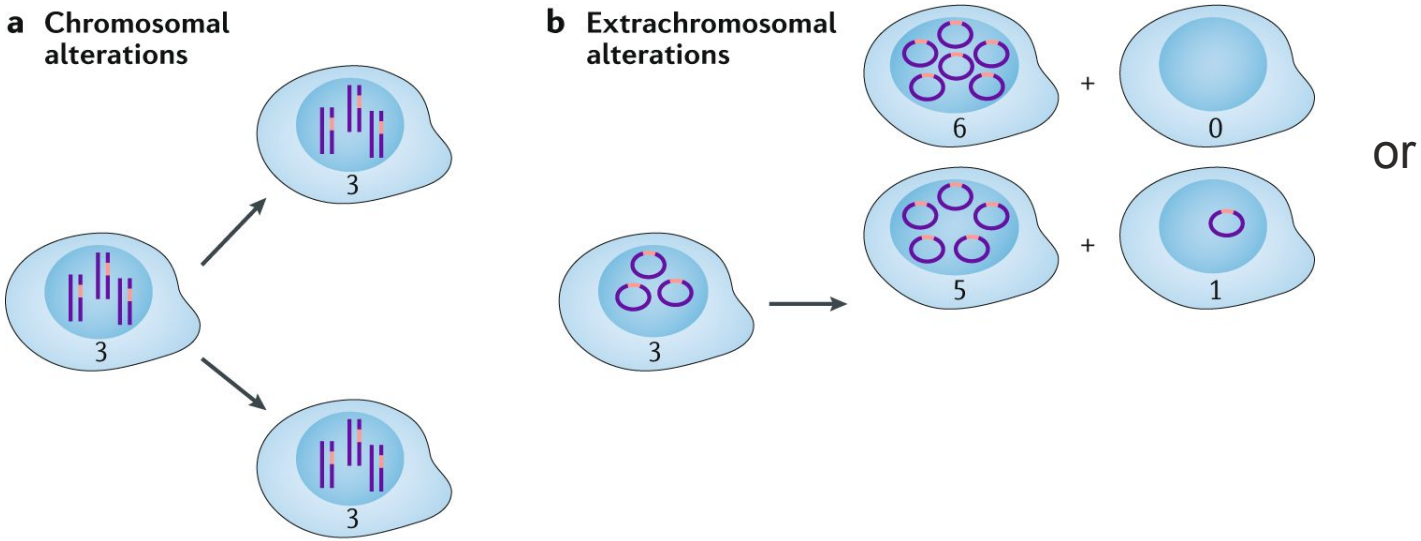
Verhaak et al, Nat Rev Cancer, 2019

Cont'd



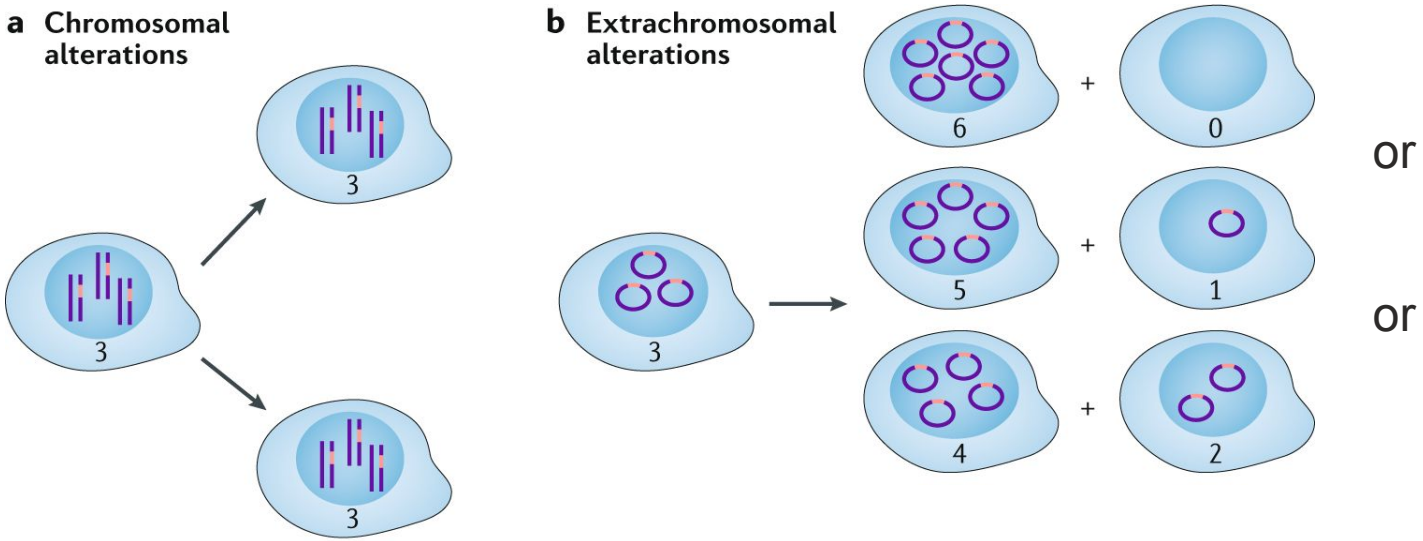
Verhaak et al, Nat Rev Cancer, 2019

Cont'd



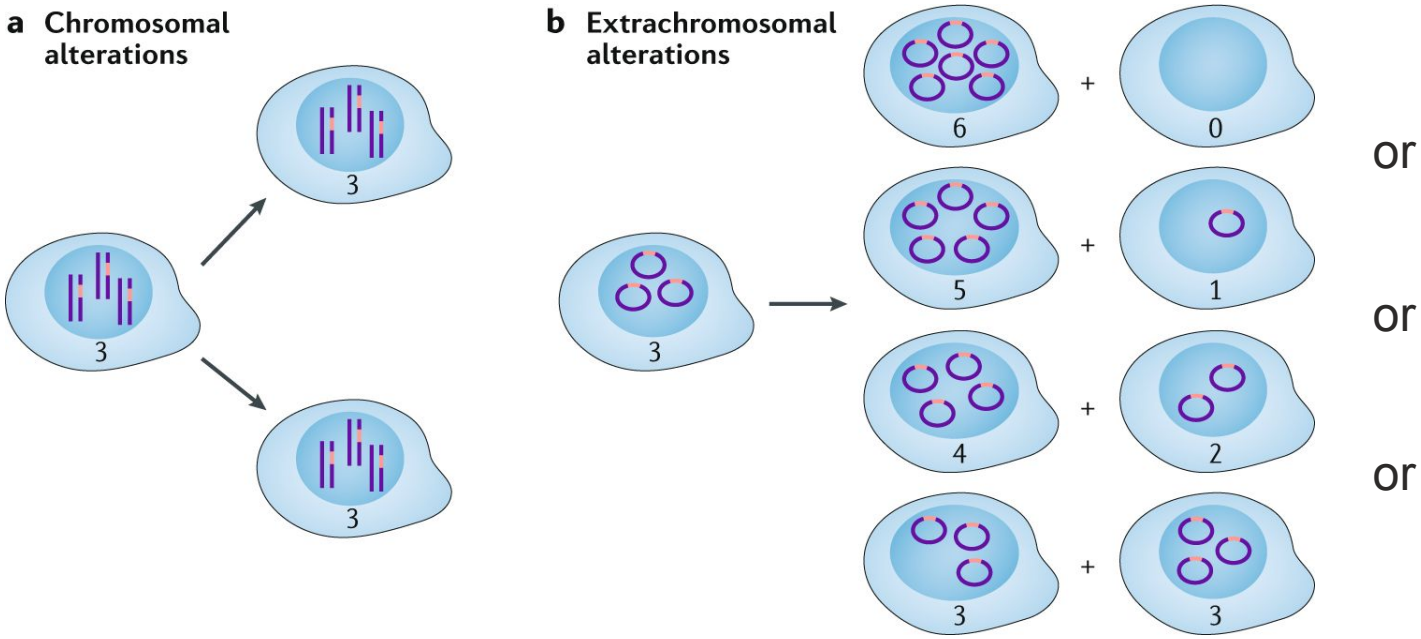
Verhaak et al, Nat Rev Cancer, 2019

Cont'd



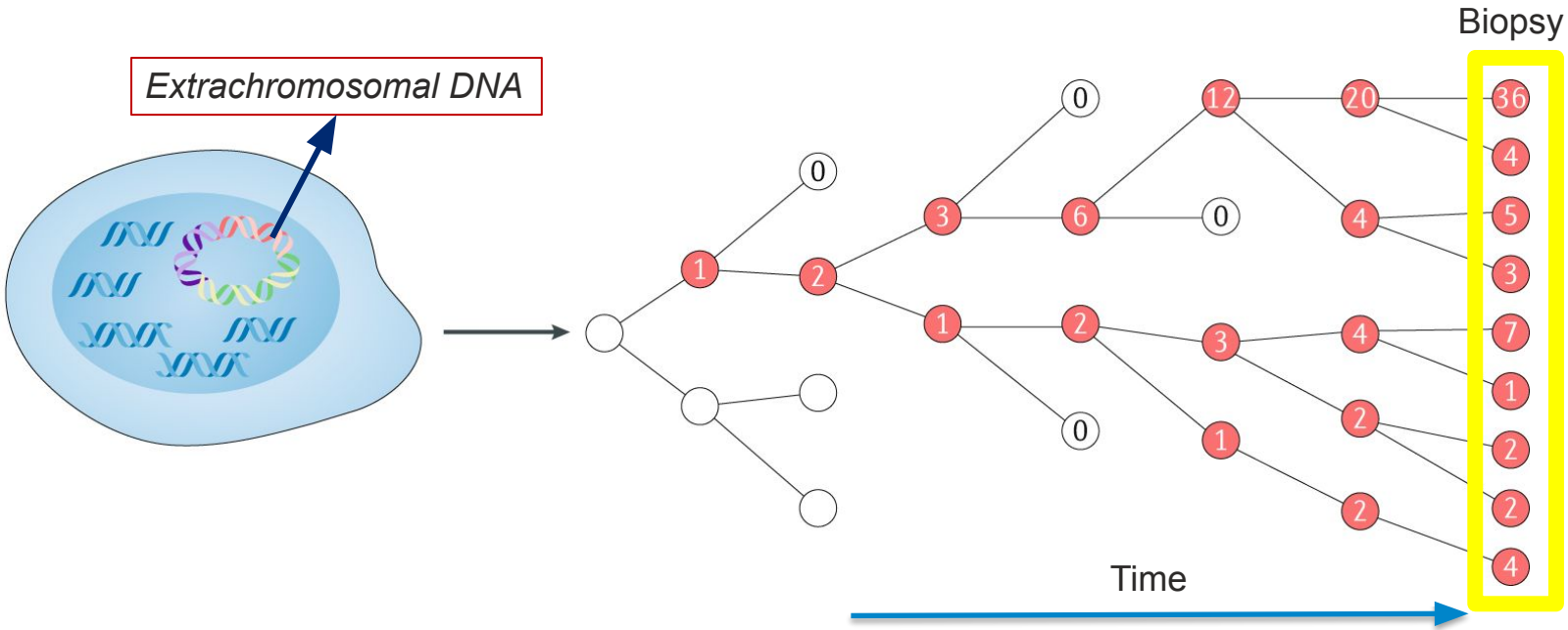
Verhaak et al, Nat Rev Cancer, 2019

Uneven and random segregation of ecDNA during cell division



Verhaak et al, Nat Rev Cancer, 2019

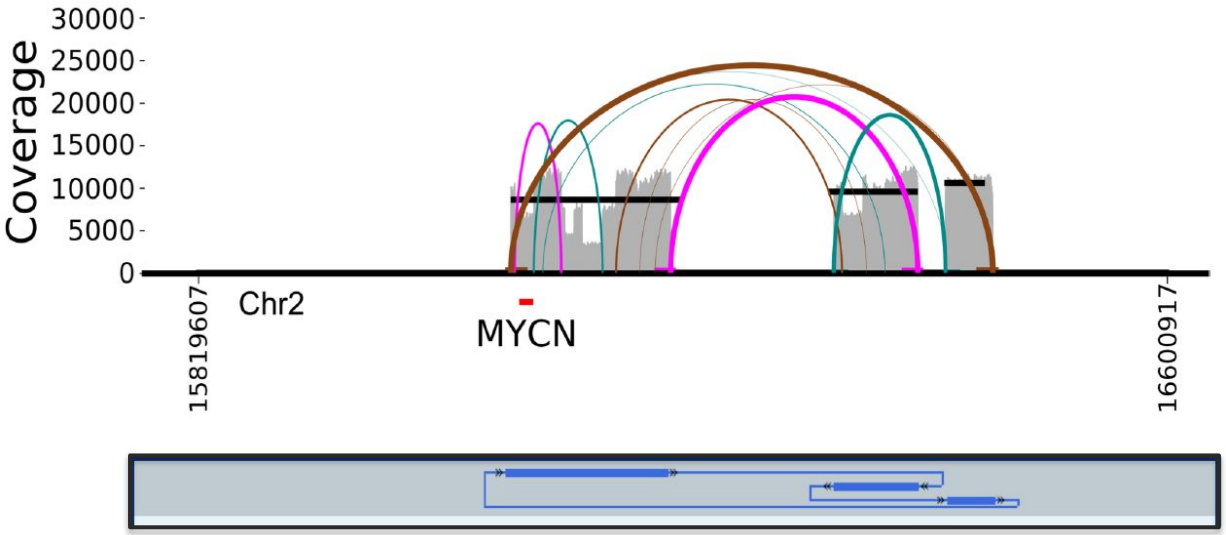
Extrachromosomal oncogenic DNA elements rapidly accumulate, driving tumor heterogeneity.



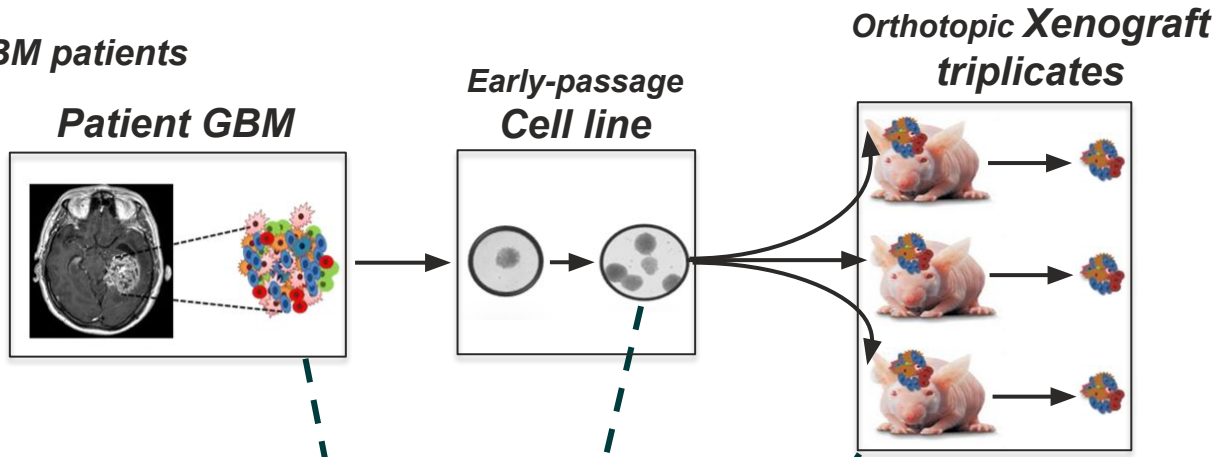
Rapid ecDNA-driven tumor heterogeneity associated with uneven ecDNA segregation.

Verhaak et al, Nat Rev Cancer, 2019

We can computationally predict ecDNA from whole-genome sequencing



Study I - Modeling GBM evolution *in vitro and in vivo*

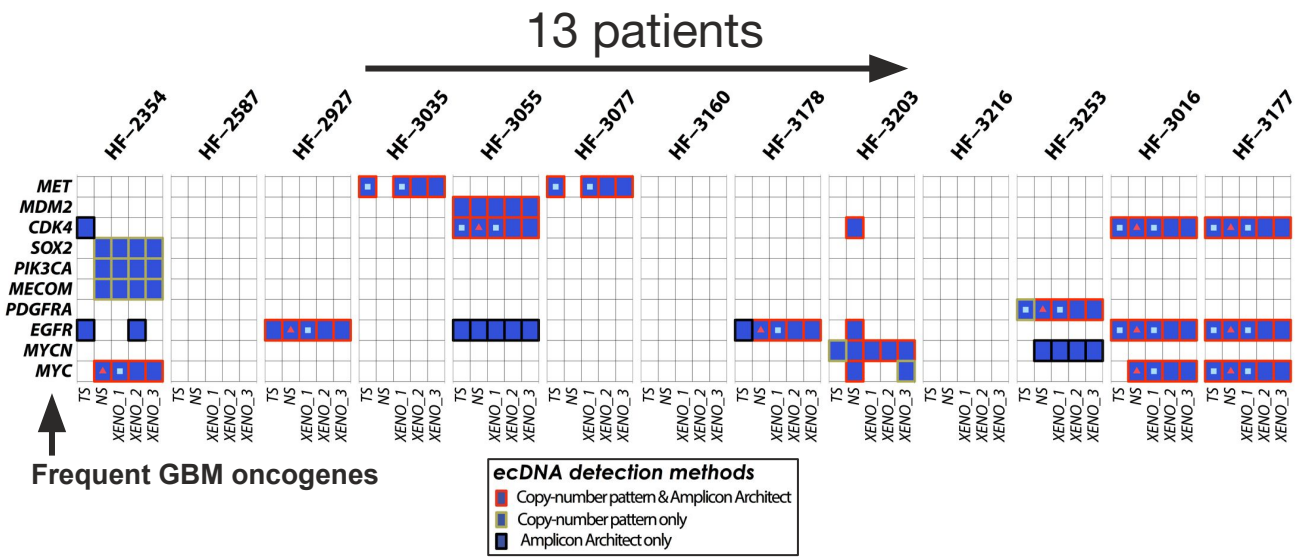


Anna, HF hospital

Whole Exome Sequencing (N = 65)
Whole Genome Sequencing (N = 65)
RNA Sequencing (N = 65)
Long Read Sequencing (N = 2)

Anna, Kim* et al*
Nature Genetics, 2018

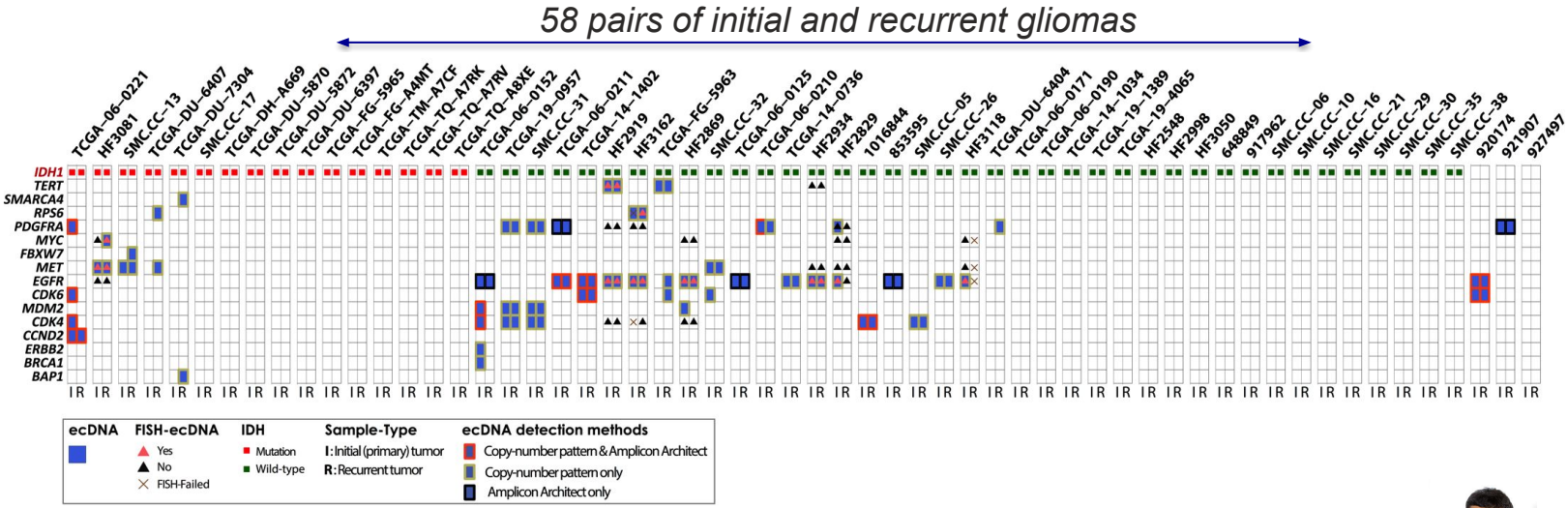
All oncogenic amplifications were ecDNAs in our data.



- We reconstructed one or more ecDNAs in most of the glioblastoma samples.
- EcDNAs are highly frequent in glioblastoma.
- **The previous ecDNA incidence rate (1.4%) may be wrong.**

DA, Kim*, Nature genetics, 2018*

Analyze 58 pairs of initial and recurrent gliomas to detect ecDNAs



- ***27 pairs of gliomas from TCGA*** *were analyzed* ***through ISB-CGC***
- *38 patients were predicted to contain at least one ecDNA.*
- *~70% of the ecDNA driver genes were preserved.*
- *High level CNV amplifications that disappeared at relapse were most likely to be ecDNAs.*


Sandeep Namburi

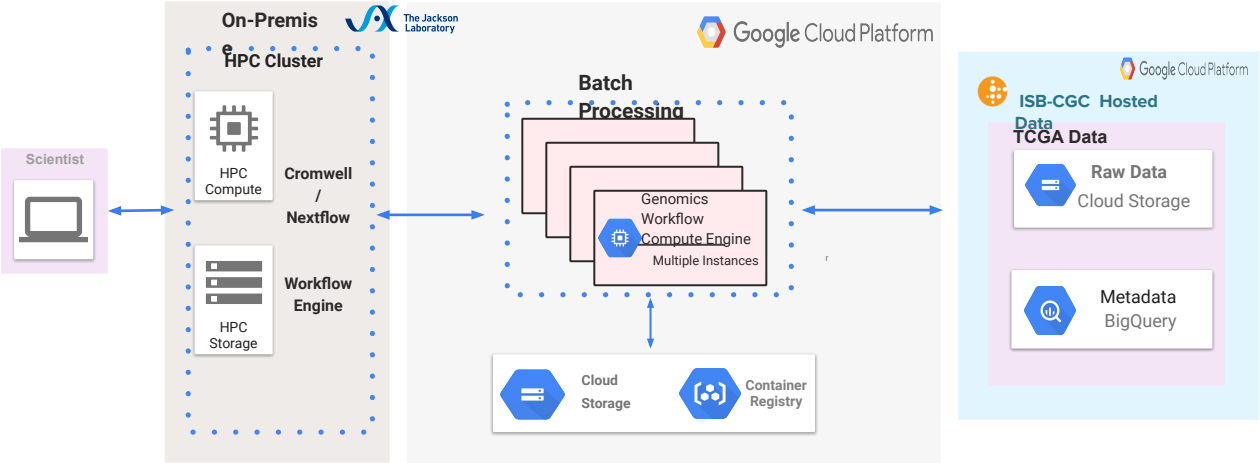
Study II - Pan-cancer survey of ecDNAs

Lineage	Tumor	Normal	Total
Prostate	301	116	417
Liver	252	52	304
Pancreatic	213		213
Renal	198	129	327
Pediatric Brain	188		188
Skin	164	137	301
Breast	159	111	270
Head and Neck	153	137	290
Gastric	145	124	269
Lung Adeno	143	145	288
Uterine Corpus Endometrial	143	137	280
Thyroid papillary	136	130	266
Bladder	112	95	207
Esophageal	112	60	172
Lymphoid leukemia	95		95
Lower Grade Glioma	85	89	174
B-cell lymphoma	83	7	90
Colorectal	74	70	144
Ovarian	70	45	115
Cervical	66	64	130
Lung Squamous cell	50	49	99
Uveal melanoma	50	51	101
Myeloid leukemia	48	39	87
Glioblastoma	47	45	92
Ewing Sarcoma	37		37
Sarcoma	36	37	73
Myeloid Disorders	30		30
Biliary tract	11		11
Oral	11		11
Total	**3212**	**1869**	**5081**

The Challenge: Large-scale data analysis in hybrid, multi cloud system

- Leverage on-premise HPC system and public cloud platforms
  - TCGA data is hosted on Google Cloud Platform (GCP)
  - ICGC data is hosted on Amazon Web Services (AWS)
  - Initial and subsequent analysis on the on-premise HPC cluster
- Use a workflow engine that supports multiple backend environments, thus avoiding reengineering of the workflow
- Minimize data transfer between the systems and avoid local storage issue.

Analysis of TCGA WGS on Google Cloud Platform



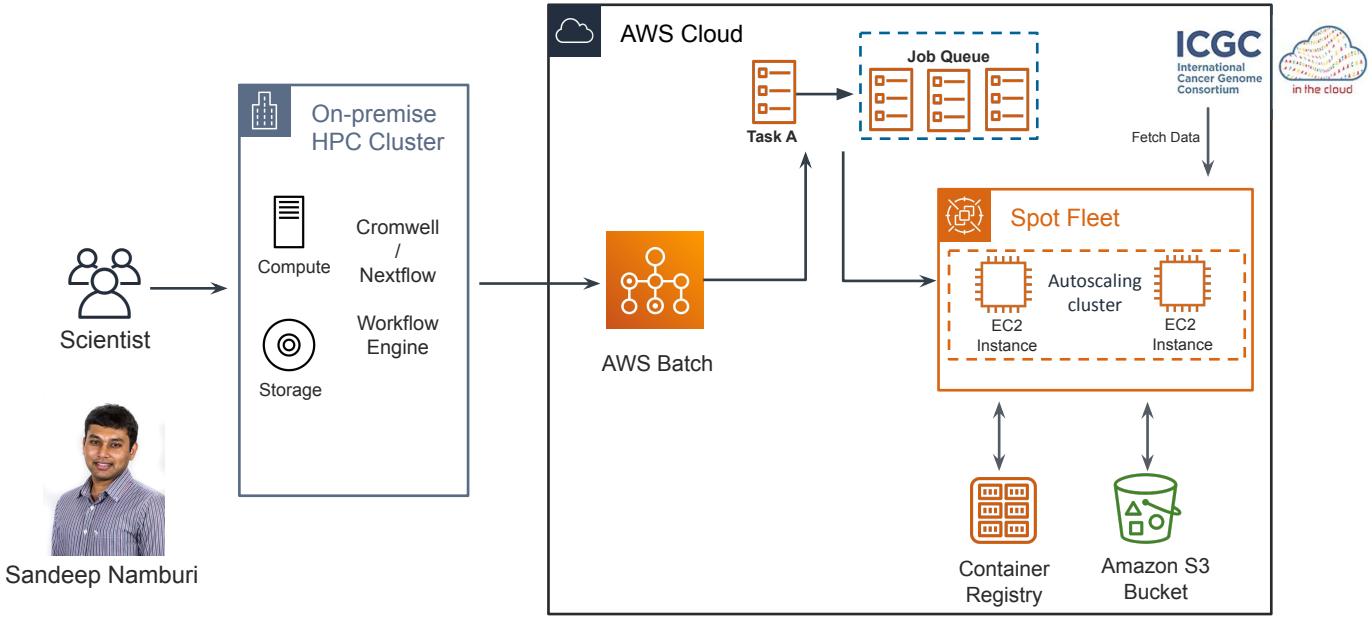
- ISB Cancer Genomics Cloud (ISB-CGC) hosts the TCGA data in the cloud
- Cromwell workflow was used.
- Co-localization of the compute and data for the computation.
- Scalable, short-lived batch analysis
- Google's Preemptible VMs to save costs (~90% discount)


Sandeep Namburi


Sheila Reynolds

ISB-CGC can be accessed at: www.isb-cgc.org

Analysis of ICGC WGS on Amazon Web Services



- ICGC data is hosted on Amazon Web Services (AWS).
- Cromwell workflow was used.
- Unlike the GCP preemptible VMs (lasting 24hours), spot instances have no such limit.
- Ability to auto-scale disks attached to an AWS instance.

We were able to predict ecDNAs and non-ecDNA types through clouds

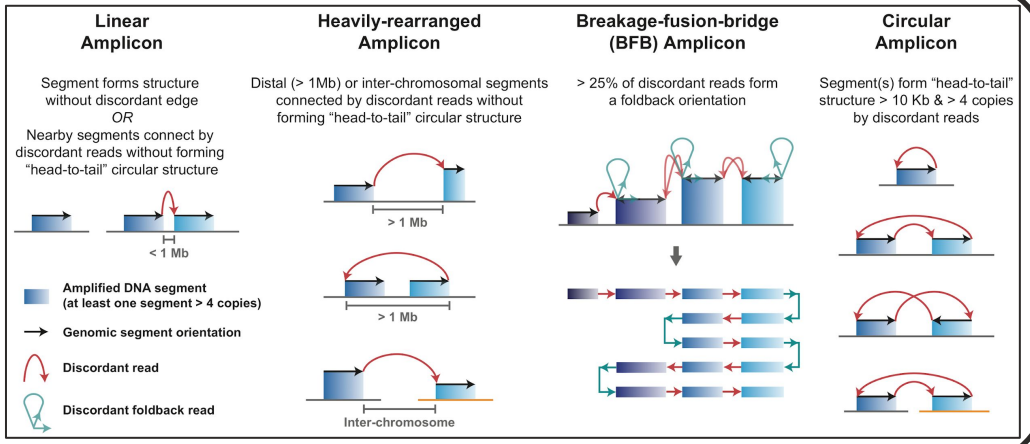
Whole Genome Sequence from ~5,000 samples

ISB-CGC
Cancer Genomics Cloud

docker

amazon web services

Google Cloud Platform



In collaboration with UCSD, Stanford, Berlin Institute of Health

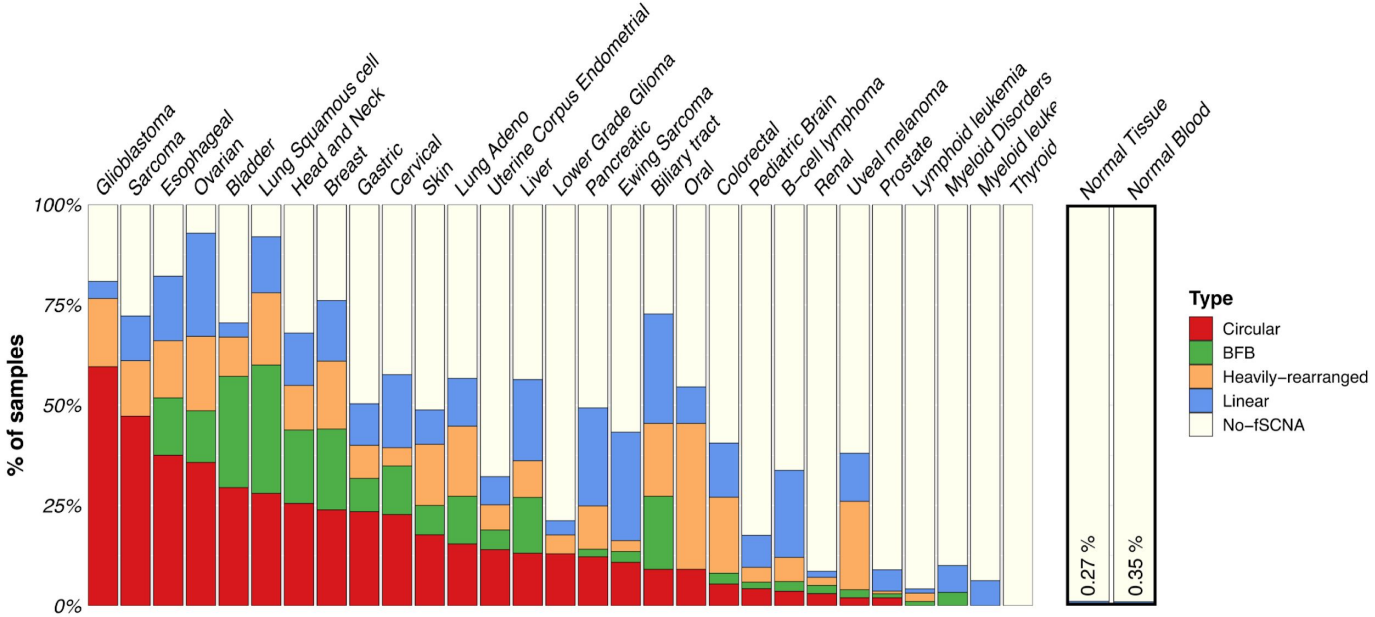
Nam-Phuong Nguyen

Jens Luebeck

Computer Science, UCSD

Kim et al, Nature Genetics, 2020

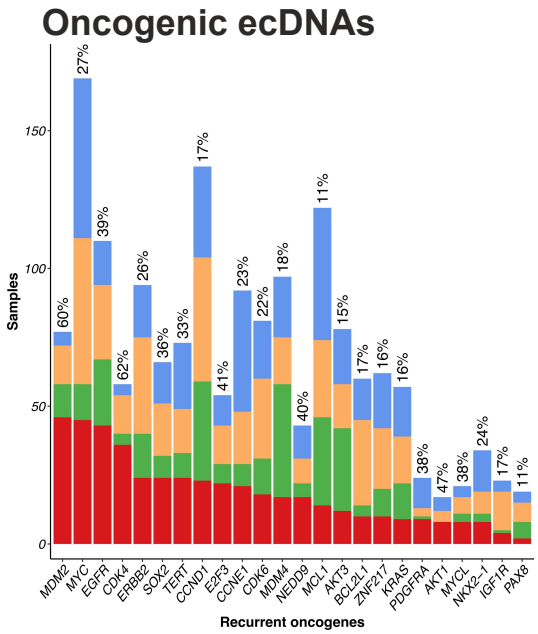
EcDNAs were found in 25 of 29 cancer types

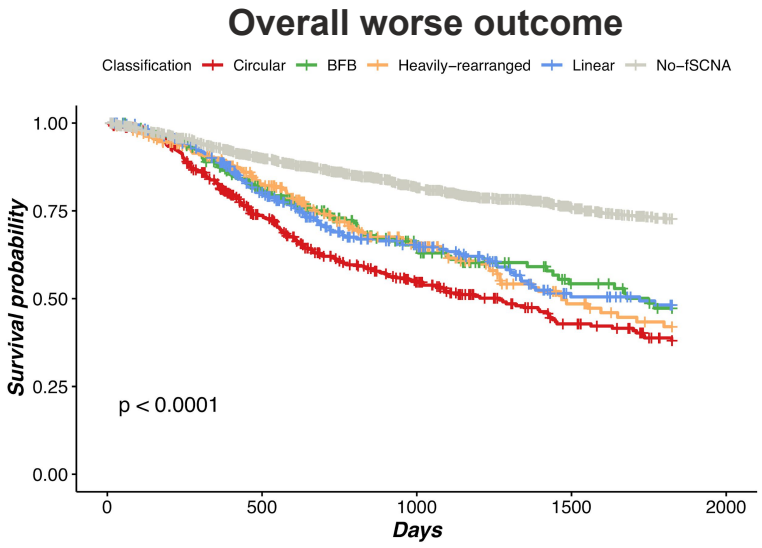


- *Higher frequencies in the most malignant forms of cancer, demonstrating that ecDNA plays a critical role in cancer.*
- *Almost none in normal*
- ***The previous ecDNA incidence rate (1.4%) is wrong.***

Kim et al, Nature Genetics, 2020

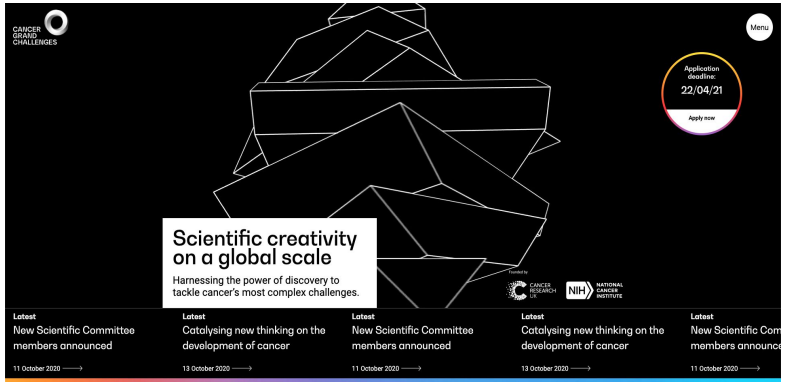
EcDNA tumors behave more aggressively, having an overall worse outcome.

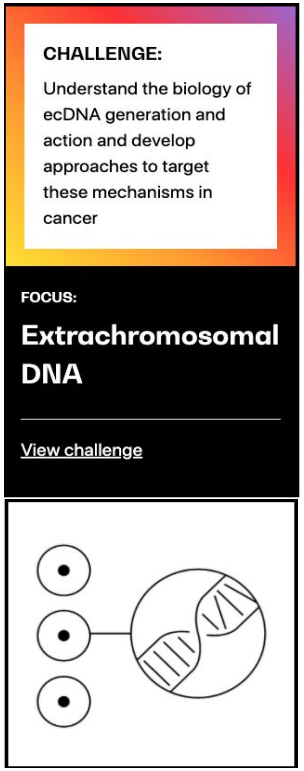




Kim et al, Nature Genetics, 2020

On Oct. 2020, ecDNA was selected as the most important problem in cancer research by the global research community





https://www.cancer.gov/grants-training/grants-funding/cancer-grand-challenges
https://cancergrandchallenges.org/

Summary

- **Extrachromosomal DNAs**
  - EcDNAs contribute to intratumoral heterogeneity.
  - EcDNA is operant in a large fraction of human cancers, contributing to the poor outcomes for patients.

- **Cloud computing**
  - Significant engineering needed to setup the resources on the cloud providers.
    - Fortunately, JAX has a cloud specialist.
  - Workflow manager with multiple systems are helpful to avoid reengineering of the workflow, rather than directly using the native executors like AWS Batch or GCP Pipelines API.

Acknowledgements

All patients providing valuable samples for research.

ICB-CGC

- **Sheila Reynolds**
- **David Pot**
- **William Longabaugh**

Henry Ford Hospital

- **Ana DeCarvalho**
- Tom Mikkelsen

UCSD

- **Nam Nguyen**
- **Vineet Bafna**
- **Paul Mischel**

Jackson lab

- **Roel Verhaak**
- **Sandeep Namburi**
- Jihe Liu
- Eun Hee Yi
- Kevin Johnson
- Floris Barthel
- Samirkumar Amin
- Kevin Anderson
- Amit Gujar
- Fred Varn

Funded by MD Anderson Odyssey Program, Google Cloud Platform

Lunch Break

We will resume at 1:30 pm ET.

Morning Session Key Messages

1. **Awesome, impactful, accelerated** science can *actually* happen by harnessing the multi-platform cloud setting!

2. Both "expert" users and "new" users are able to leverage the advantages of cloud platforms when supported.

3. Users still face "binaries" in decision making that limit their full potential for harnessing platforms/cloud:

    a. Costs/platforms→ On Prem vs. Cloud (and which cloud?), where and from whom do I have my credits, how do I support "other" data (see b.) -- help with cost optimization.

    b. Terra vs. SBG vs. ISB vs. "X" →

        i. What data do I have to move where since I not only am I accessing multiply hosted datasets, **but have some of my own data, own cohorts, or other existing studies that I need to intersect with the cloud-based cohorts (relates to the multiple cohort creation processes users will engage when navigating interop)**.

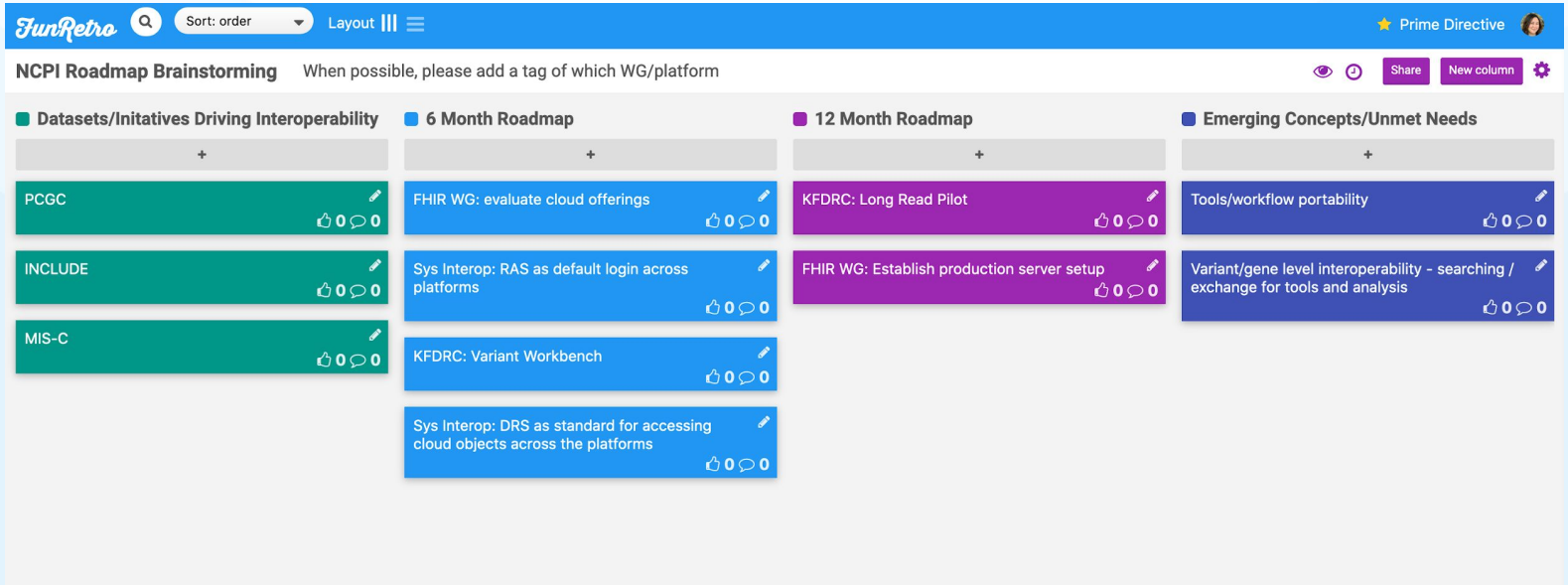
    c. CWL vs. WDL → where should I either invest in transforming my pipelines or are the "right" combinations of multiple pipelines available? Is there a way not to be "locked in" by this?

Intro: Capturing Roadmap Ideas

Utilizing Fun Retro

Can start putting ideas down during WG updates

Hour interactive session at the end of the day



Working Group Updates:
NIH Coordination

Valentina Di Francesco & Ken Wiley
NHGRI/AnVIL

Membership

NHGRI AnVIL
- Valentina Di Francesco (Co-Chair)
- Ken Wiley (Co-Chair)
- Natalie Kucher

NHLBI BioData Catalyst
- Jon Kaltman
- Alastair Thomson
- Chip Schwartz

CF GMKF
- Valerie Cotton
- James Coulombe
- Huiqing Li

NCI CRDC
- Tanja Davidsen
- Allen Dearry
- Vivian Ota-Wang
- Erika Kim
- Zhining Wang
- Ian Fore

NIH CFDE
- Lora Kutkat
- Haluk Resat
- Chris Kinsinger

Coordination WG's Responsibilities

- Serve as the NIH Governance body for NCPI
- Stewardship of the NCPI WGs activities
- Liaison with NIH ODSS and other parts of the NIH

NCPI Governance

- Ratified the NCPI Interoperability Principles proposed by the Community Governance WG
- Aiming to balance the NCPI's goals and priorities versus IC-specific platform goals and priorities
- Addressing specific issues that arise, such as those related to the NCPI's developers access to the resources for testing platforms' interoperability tools
- Forum for ICs reps interactions and information sharing

Stewardship

Launched five trans-NIH WGs

NCPI All Hands Workshops

- 1st kick-off workshop hosted in Oct 2019 by NHLBI/BDC at RENCI
- Internal “Train your Colleague” workshop organized by the NHGRI/AnVIL and the Training WG in March 2020 *(virtual)*
- 2nd workshop hosted in April 2020 by NHGRI/AnVIL *(virtual)*
- 3rd workshop hosted in Oct 2020 by CF/Kids First *(virtual)*

Liaison with NIH Constituents

- Align NCPI efforts with the goals of the NIH Strategic Plan for Data Science
  - Facilitate collaboration with the NIH RAS Project
  - Leverage of ODSS supplement funds
  - Leveraged the 2020 ODSS Data Scholar program
- Interaction with the NIH Data Access Policy groups
- Information dissemination across the NIH

Goals for Year 2

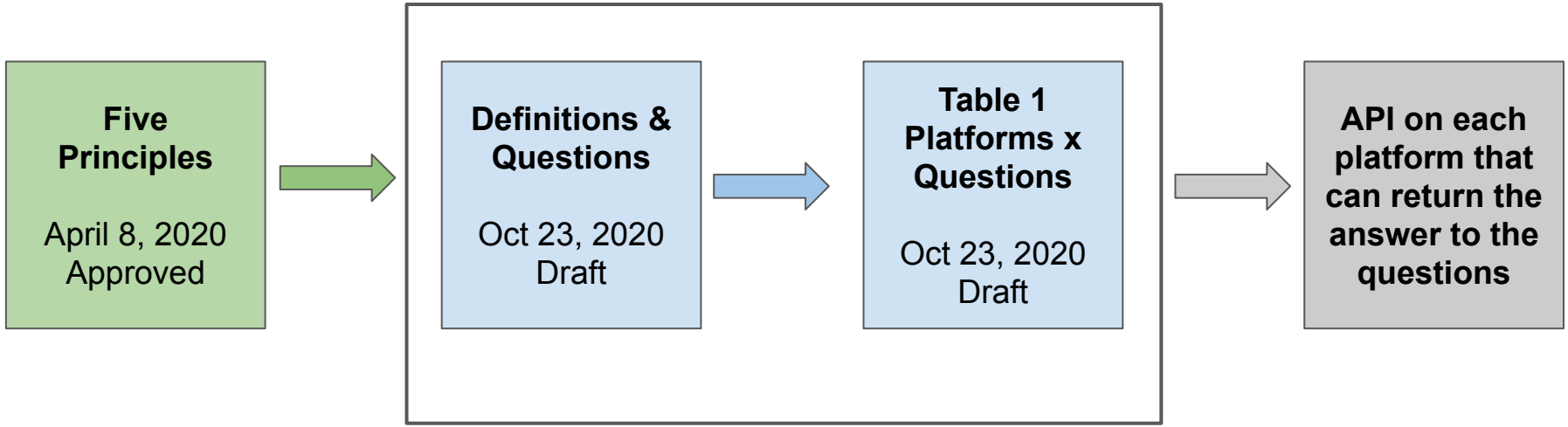
- Identify and agree upon next year's priorities and milestones
- Implement interoperability principles
- Host NCPI all hands workshops every 6 months
- Offer training opportunities for outside investigators
- Pursue additional funding support
- Continue collaboration with RAS
- Improve visibility across the NIH and share best practices for platforms interoperability across NIH
- Solidify collaboration with GA4GH work streams

Working Group Updates: Community / Governance

Bob Grossman, Professor, University of Chicago

Stan Ahalt, Director, RENCI

Community / Governance WG - Overview

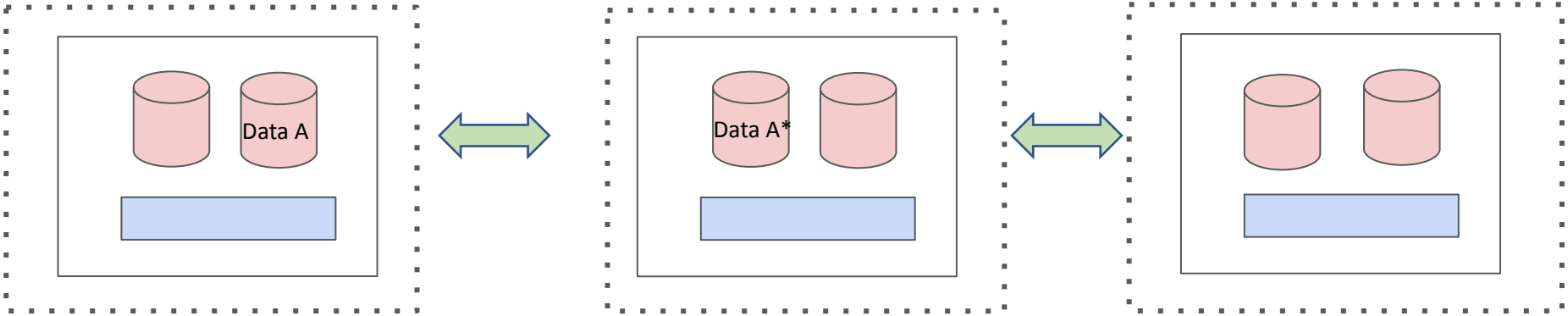


focus since the last meeting

White Papers

- Five Principles (Version C) for Interoperating Data Platforms was approved on April 8, 2020
- These principles were not precise enough to determine easily whether a platform was following them or not
  - Three of the questions are the most relevant to interop between cloud platforms
  - We have drafted a white paper that provides **definitions** and a series of **questions** that each platform can answer that provides enough specificity so that a platform's adherence can be determined
  - Towards Characterizing Cloud Platform Interoperability (October 23, 2020)
  - Short name - C2PI White Paper

We have some blockers ...



Platform A boundary

Platform B boundary

*copy or DRS identifiers

Platform C boundary

In general, platforms would like to access other platforms data, but are hesitant to let other platforms access their data.

What type of agreement ⟷ are required for a User in Platform B or C to access data that they are authorized to access?

Key Concepts

- **Trust** - if two platforms trust each other they should be able to exchange data

- **Authorized environment** -
  - New concept in our C2PI White Paper
  - Example, with dbGaP the organization's IT Director through the organization's SO authorizes an environment for data downloaded from dbGaP
  - Example, for a cloud platform, the Institute's CISO can authorize an environment, say by approving an ATO for FISMA Moderate environment
- **Authorized Environment Principle -** authorize environments and authorize users and trust the authorizations

Key Questions

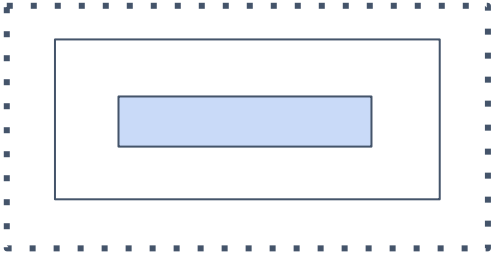
- What categories of data?
  - open, controlled access, sensitive – low, sensitive – medium, sensitive – high
- What are the requirements to authorize a user?
  - InCommon, ORCID, RAS, dbGaP, platform white list
- What are the requirements to authorize an environment?
- What are the requirements to trust another platform?

For a particular category of data

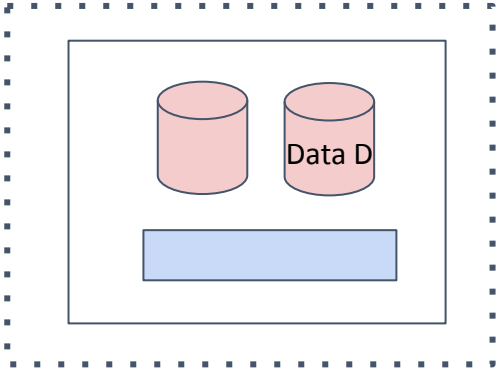
- Meta-principle: an authorized user can access data in authorized environment (for an appropriate category of data).

	dbGaP Model	GDC Model	CRDC	BDC	AnVIL	KF
Status	reviewed	reviewed	under review	reviewed	reviewed	reviewed
User Auth	dbGaP	dbGaP	dbGaP	dbGaP & white list	dbGaP, white list, DUOS	dbGaP & white list
Environment Authorization	Signing Official who has the legal authority to attest to the organization's CIO's data security assessment "dbGaP Model"	Signing Official who has the legal authority to attest to the organization's CIO's data security assessment "dbGaP Model"	SBG, Terra & ISB are authorized environments; need to get list of other authorized environments	Institute CISO	Broad CISO approves ISAs for connecting to AnVIL; and, AnVIL uses dbGaP model for data that is downloaded	Research organization's IT Director
Data access (aka "egress") by another cloud platform	Any platforms authorized by researcher's organization (via dbGaP) "dbGaP Model"	Any platforms authorized by researcher's organization (via dbGaP) "dbGaP Model"	to be determined	Data cannot leave BDC Platform.	Restricted to platforms with an ISA with AnVIL	Any platforms authorized by researcher's org. (via dbGaP)
Data Egress - "download"	Any platforms authorized by researcher's organization (via dbGaP) "dbGaP Model"	Any platforms authorized by researcher's organization . (via dbGaP) "dbGaP Model"	Any platforms authorized by researcher's org. (via dbGaP)	Data cannot leave BDC Platform.	dbGaP model for downloaded data	Any platforms authorized by researcher's org. (via dbGaP)
API	archive can be downloaded, but no API to data	All data is available via an API	Data objects available via API; CCDH and CDA will provide access to clinical data	API within BDC for data objects and harmonized data (in the future APIs for multiple data models); PicSURE API for clinical/Phen.	API within AnVIL for data objects and harmonized data (in the future APIs for multiple data models)	All data is available via Gen3/portal APIs. Gen3 for genomic data. FHIR API for clin/phen Q1 2021.
Trust relationships	NA	open to any auth. env.	need to determine	need to determine	need to determine	need to determine

Questions About Workspaces



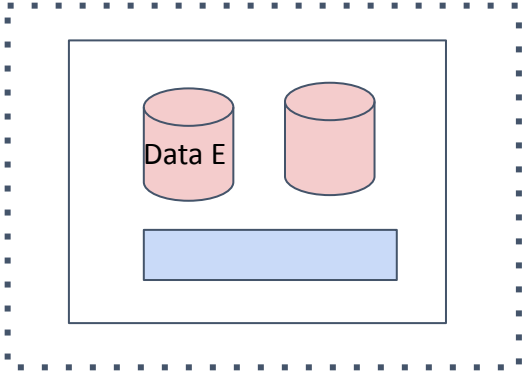
Workspace W (authorized environment)



Platform A boundary

A. Environments that are authorized by the user's organizational Signing Official (SO) through the organization's CISO through a dbGaP application.
B. Environments that are authorized by an Institute's CISO or another authorized CISO.
C. Environments that operated under FISMA Moderate.
D. Environments that are operated under FedRAMP.

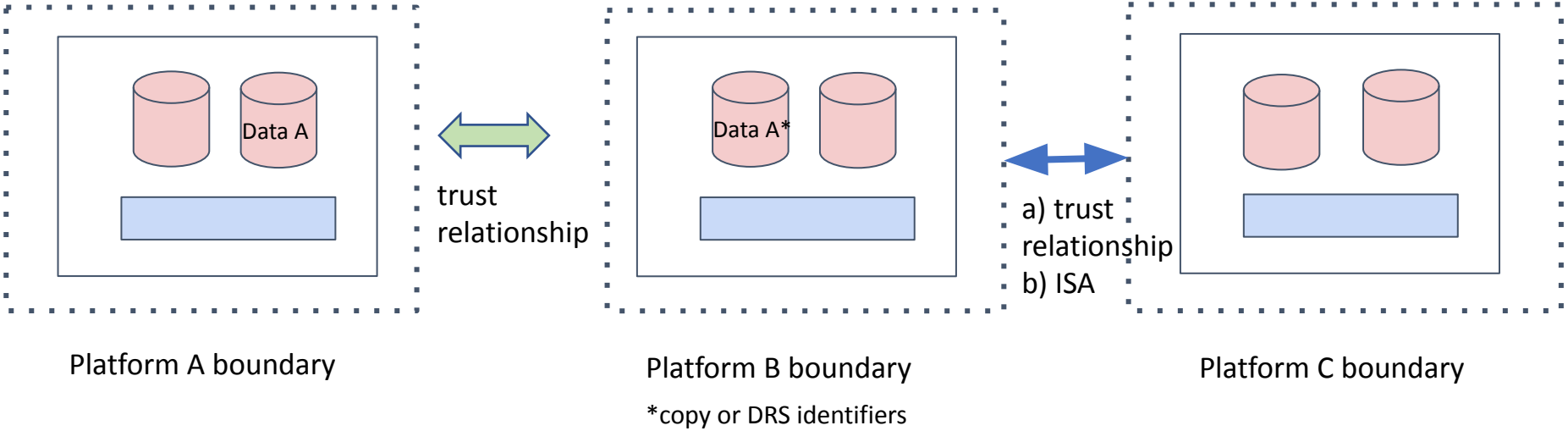
Note that as a special case Workspace W may within the security boundary of Platform A, Platform C or both.



Platform C boundary

Questions: Can an authorized user in Workspace C access Data D from Platform A and data E from Platform B if Workspace W is an authorized environment of Type A? Of Type B? Of Type C? Of Type D?

Questions About Data Access Between Cloud Platforms



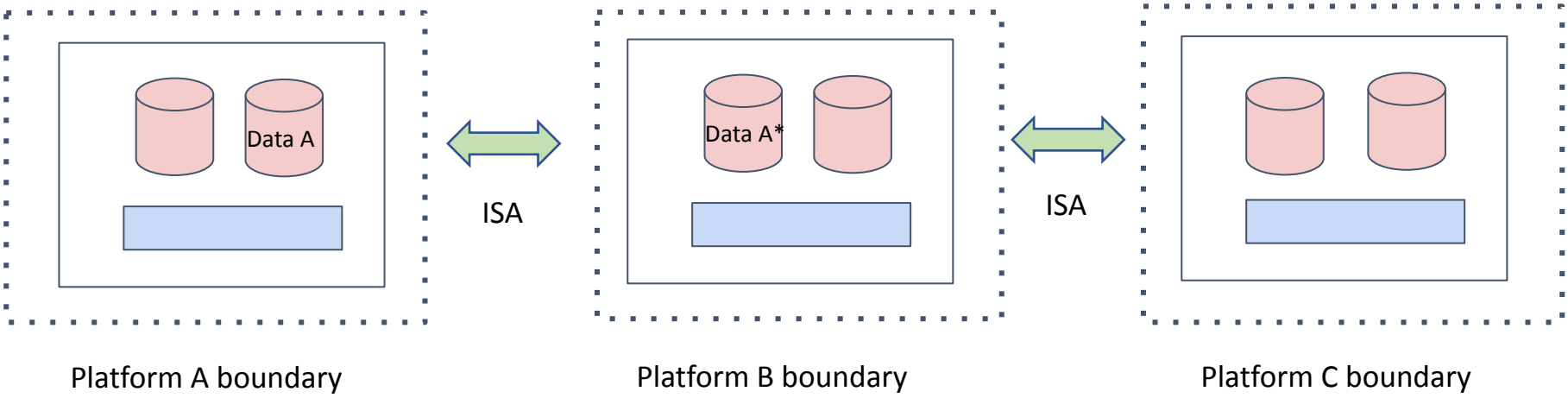
Question: Can an authorized user in platform C access Data A from Platform B?
Question: Can an auth. user in platform C access Data A from Platform B, if Platforms A and C have a trust relationship?
Question: Can an auth. user in platform C access Data A from Platform B, if Platforms B and C have a trust relationship?
Question: Can an authorized user in an authorized workspace in Platform C analyze Data A from Platform B?
Question: Can an authorized user in platform C access Data A from Platform B, if Platforms B and C have a trust relationship and platforms A & C have a trust relationship?

Road Map

1. Complete the C2PI White Paper, including Table 1
2. Define an API so that cloud platforms can self-attest how they answer the C2PI Questions
3. Work towards approving a policy for the commons in NCPI that an authorized user can access data in authorized environment (for an appropriate category of data).
4. Work towards getting some of the NCPI platforms to trust each other

Backup Slides

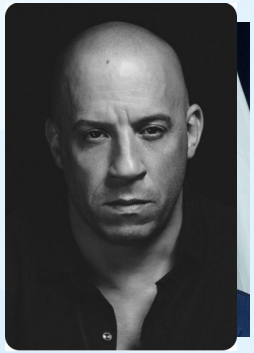
Questions About ISAs



Question: Can an authorized user in platform C access Data A from Platform B?
Question: Can an authorized user in platform C access Data A from Platform B, if Platforms B and C have a trust relationship?
Question: Can an authorized user in an authorized workspace in Platform C analyze Data A from Platform B?

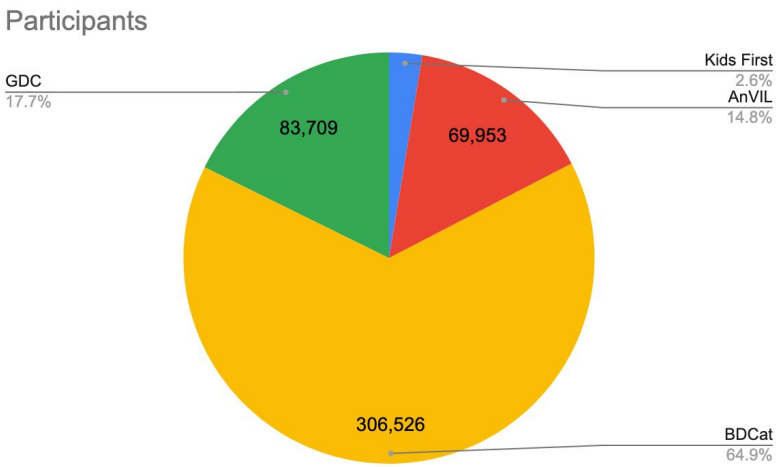
Working Group Updates: Systems Interoperation

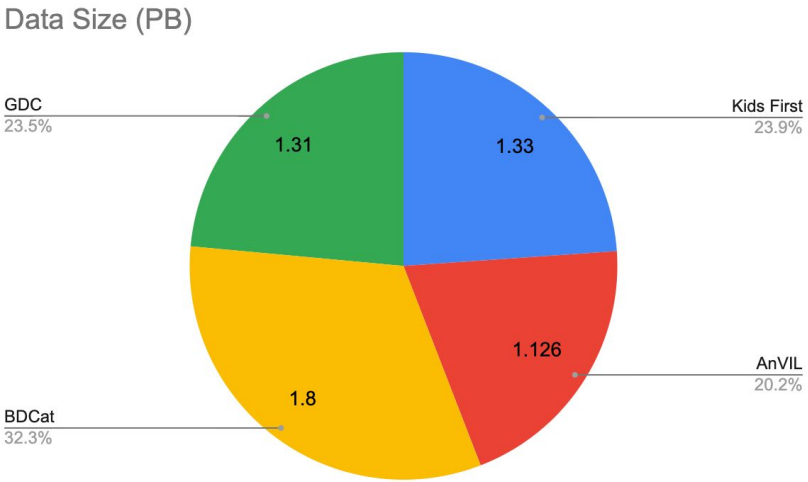
Brian O'Connor Broad
& Jack DiGiovanna Seven Bridges



Systems Interoperation WG - Motivation

Researchers want to access data across ICs/stacks.

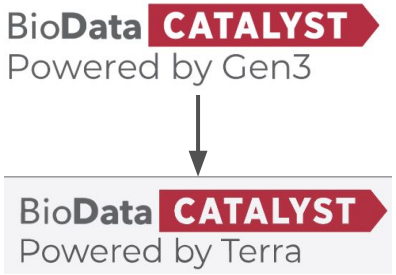


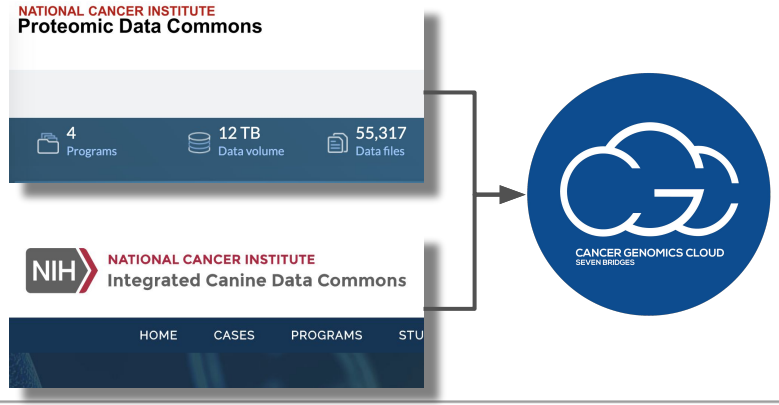


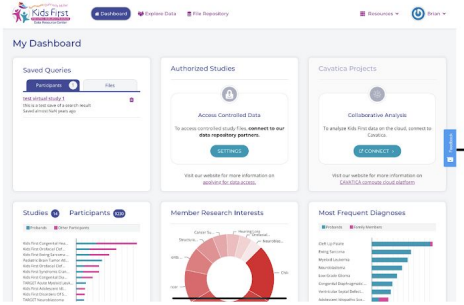
<u>Aggregation of data across these IC stacks is huge ~6PB</u>

Systems Interoperation WG - Motivation

*Data portals connect (**intra-IC**) with analysis systems (workspaces)*







Systems Interoperation WG - Mission

The group's *Charter* establishes the group's mission, members/teams, high-level scientific and technical goals, and timeline.

The group will spearhead **technical improvements** to cloud "stacks" created by the Common Fund, NCI, NHGRI, and NHLBI that enable improved interoperability. We will **demonstrate progress** in **realistic researcher use cases** every **6 months**.

Please join if you are interested.

Systems Interoperation WG - Use Cases

Immediately looked for scientific “driver projects”

Our WG quickly identified 8 interesting researcher use cases that required interoperability both within and between ICs:

- CRDC + AnVIL (n=2);
- BioData Catalyst + Kids First (n=3)
- AnVIL + Kids First (n=1)
- BioData Catalyst + Kids First + AnVIL (n=2)

Systems Interoperation WG - Tech Challenges

Standardized Handoff Mechanism ✅

Standardized Data Access Methods ✅

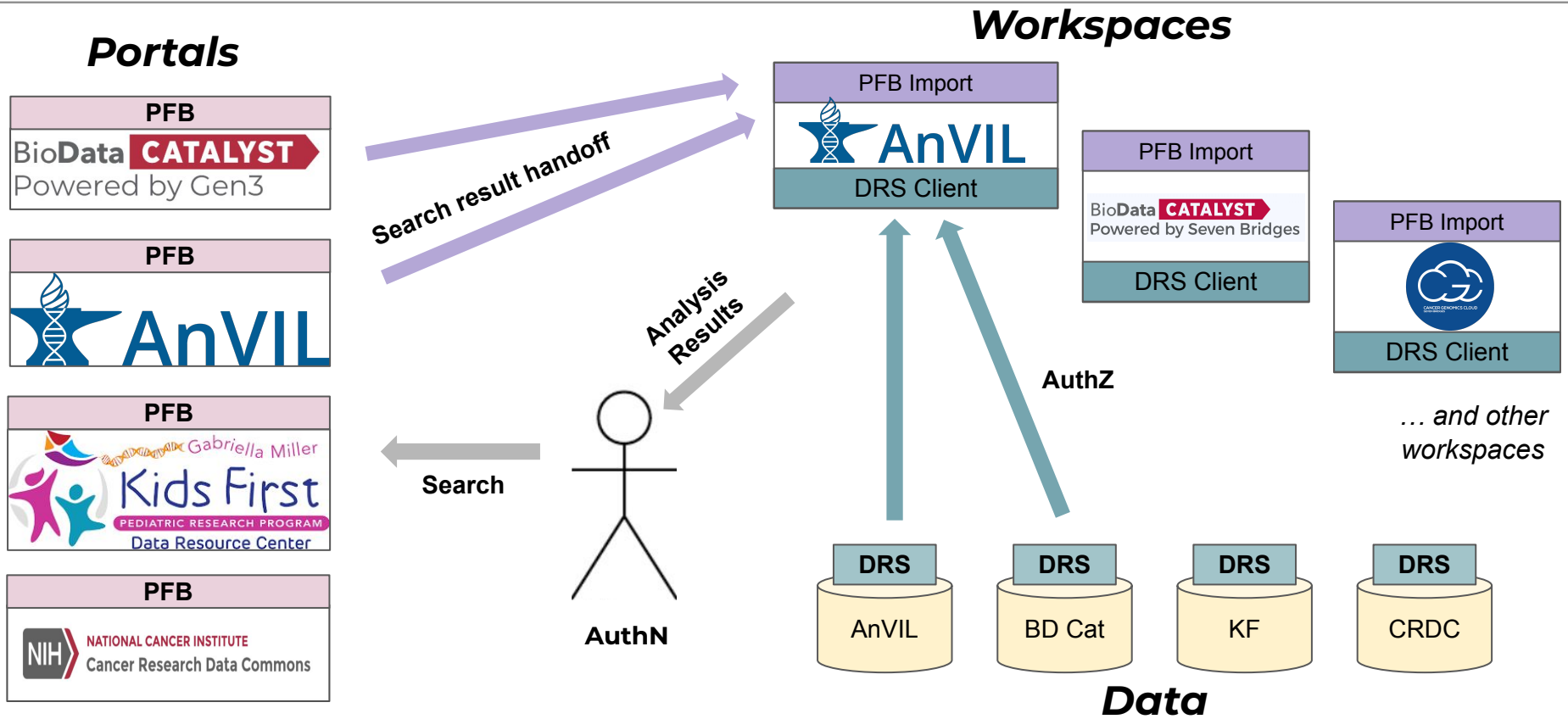
Avoiding Egress and Data Locality Costs ✅-ish

Unified Authentication/Authorization - *more progress than expected*

Common Metadata Model Between Systems - *progress on "light" solution*

Coordinated Project Work Plans and Technical Timelines ✅-ish

Systems Interoperation WG - Technical 1st Year Vision



Systems Interoperation Timeline - April 2020

Given the technical gaps, what might a timeline be for filling these?

	Q2			Q3			Q4
Complete	April	May	June	July	August	September	October

PFB Output
BD Cat
[Gen3->Terra]

Terra Workspace
PFB Support

SB Workspace
DRS 1.0 support

DRS 1.x
Server
[KFDRC]

DRS 1.x
Server
[BD Cat]

DRS 1.x
Server
[CRDC]

DRS 1.x
Server
[AnVIL]

Terra Workspace
DRS 1.1 support

DRS 1.1 client specification
release by GA4GH

Gen3
Workspace
PFB Support

PFB Output
AnVIL
[Gen3->Terra]

RAS Phase
1 testing

Gen3 Workspace
DRS 1.1 support
(from external servers)

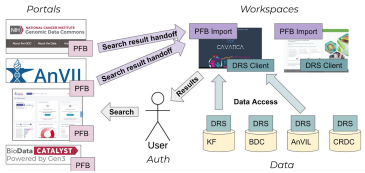
SB Workspace
PFB support

PFB Output
BD Cat
[Gen3->SBG]

FHIR WG
"Project Forge"
Completion

KFDRC DRS
Handoff

PFB Output CRDC



Technical Safe Harbor July 15th? Aug 15th?
A series of Safe Harbors through summer?

Opportunity to test use cases with the tech components built by Summer
Quantify improvements to the user experience… look for gaps…

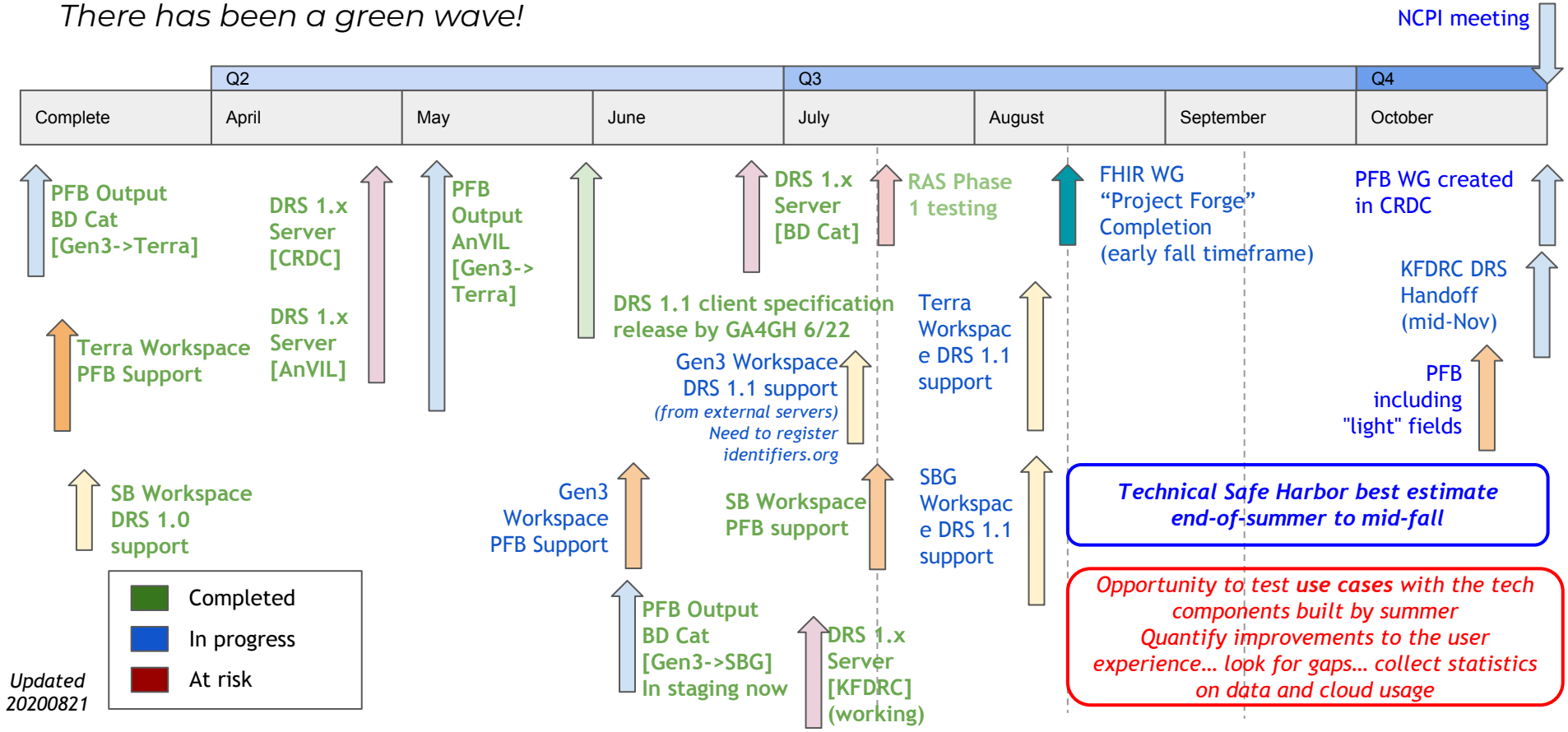
Safe Harbor 1 ***Safe Harbor 2*** ***Future Safe Harbors…***

Completed
In progress
At risk

Update
20200416

Systems Interoperation Timeline - Q3 2020

There has been a green wave!



Systems Interoperation WG - 2020 Accomplishments

*Collectively, we have achieved improved interoperability in 2020 across multiple systems through **PFB**, **GA4GH DRS**, and **GA4GH Passports**.*

2020 Results

- **Search Result Handoff:** PFB

  2 portals
  ~417K subjects accessible

- **Data Access:** DRS 1.1

  4 DRS Servers
  ~6PB of data

- **Auth**: RAS for AuthN

  RAS

Supported Platforms

- The **NHGRI AnVIL** and **NHGRI BioData Catalyst** portals both support handoff of search results to **workspaces** (Terra, Gen3, SBG)
- We have data accessible on **AnVIL, BDCat, CRDC,** and **Kids First** via **DRS 1.1** support
- **GA4GH Passports** are in use by **RAS** and support visas from dbGaP made accessible by Gen3.

NCPI Systems Interoperability Demo

NCPI 2020 Fall Workshop
2020-10-30

Jack DiGiovanna (SB)
& Brian O'Connor (Broad)

Systems Interoperation & Global Efforts

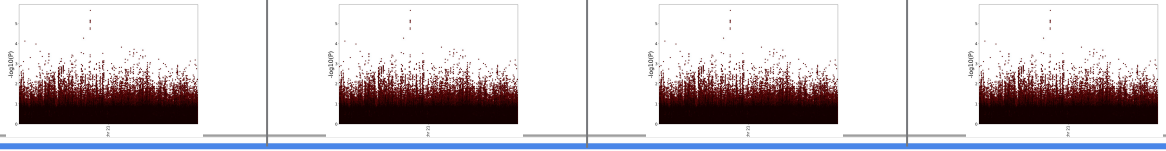
GA4GH also recently demonstrated systems using API standards to interoperate

Because NCPI Systems Interoperation uses many GA4GH APIs, we were able to participate in a global interop demo!

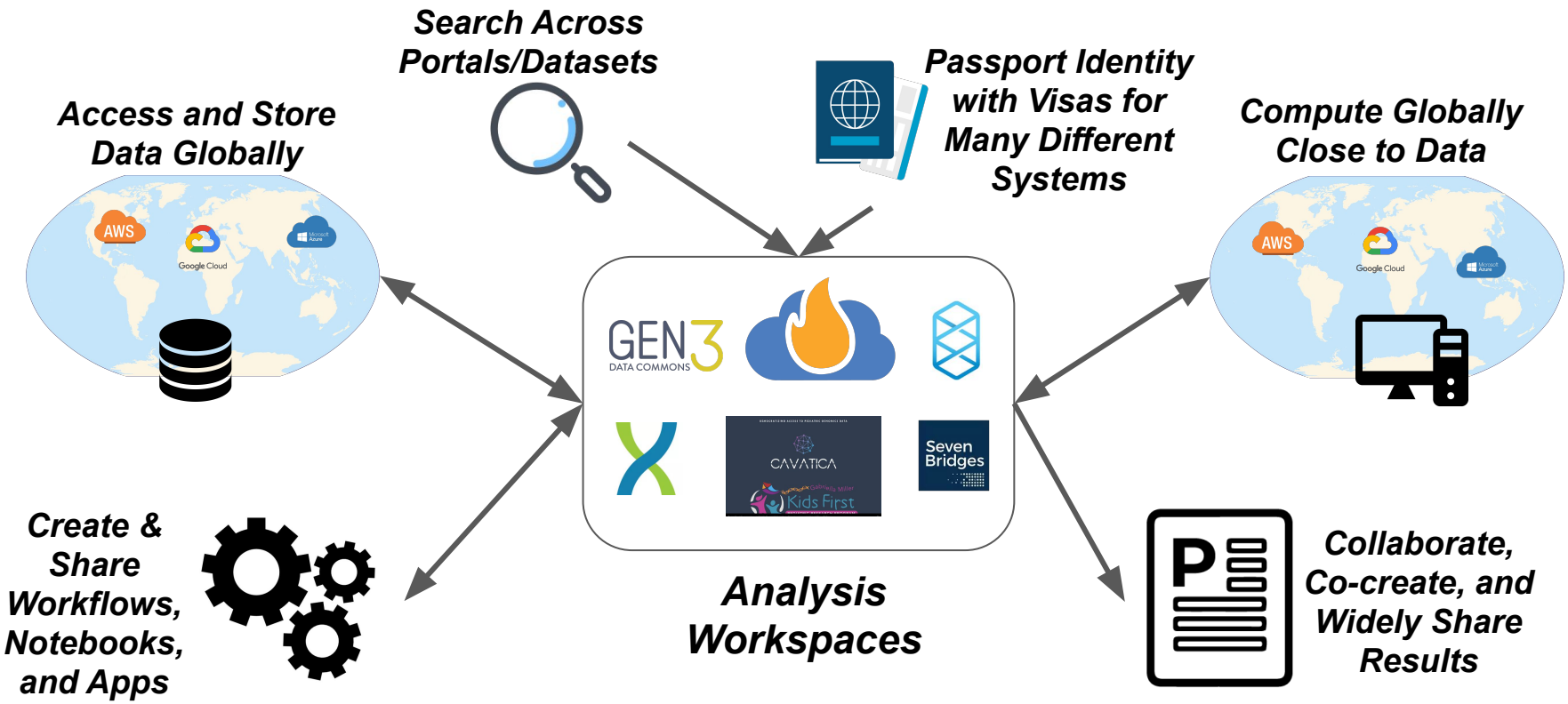
Biomedical Platform UI

1. DNASTACK
2. Terra
3. elixir AAI
4. SevenBridges

Analysis Results



Systems Interoperation Long Term Vision



Systems Interoperation WG - Second Year Goals

Technical Challenges (next 6 months+):

- **Production*:** How do we transition our work to more production systems?
- **Auth*:** How to leverage RAS & passports for authorization going forward?
- **Search/Discovery***: How to find data across portals e.g. FHIR, CDA, etc?
- **Common Metadata Models*:** How portals and resources can structure metadata consistently?
- **Workflows, Data Locality and Egress*:** How to compute in place automatically, across clouds, avoiding egress?
- ***And more... roadmapping later today***

Policy Challenges (next 6 months+):

- **Policy:** Complex, heterogeneous, & evolving landscape, remains a blocker
- **Adoption:** Engagement and outreach to drive adoption of these standards and drive new scientific analyses.
- **Tool Availability/Portability**: Leveraging different workspaces for different parts of analysis, finding the equivalent tool for your workflow language
- **Reproducibility / Knowledge Life cycle**: Strategies for expiring docker images, target support timeframe for a tool

** key potential areas for future collaboration*

Working Group Updates:
FHIR

Allison Heath, PhD
Director of Technology @ D3b, CHOP
Robert Carroll, PhD
Assistant Professor, VUMC

Overview

- First Seven Months of the WG
  - Project Forge
  - Development Infrastructure
- Demo
  - Data Dashboard
  - Exploration and filtering of data
  - Linkage to Monarch APIs
- Roadmap
  - Expansion of data covered
  - Tool support for data
  - Deploy limited production implementation

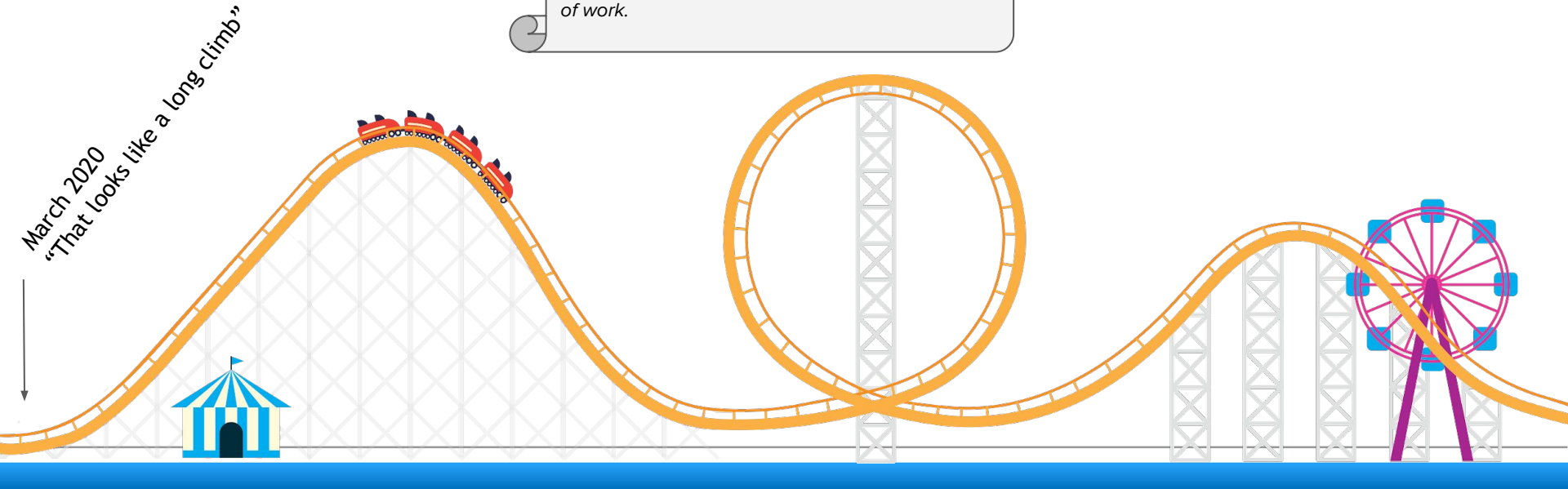
Seven Months Ago...

2020-03-16: Initial Kickoff Meeting

Minutes (Not Verbatim)

AH- Leverage FHIR to enable interoperability across stacks.

RC- Goal is to break silos down between different resources. It's a new idea to use FHIR for this type of work.



Getting Started

2020-04-03: Collaborative Kickoff Project ("Project Forge")

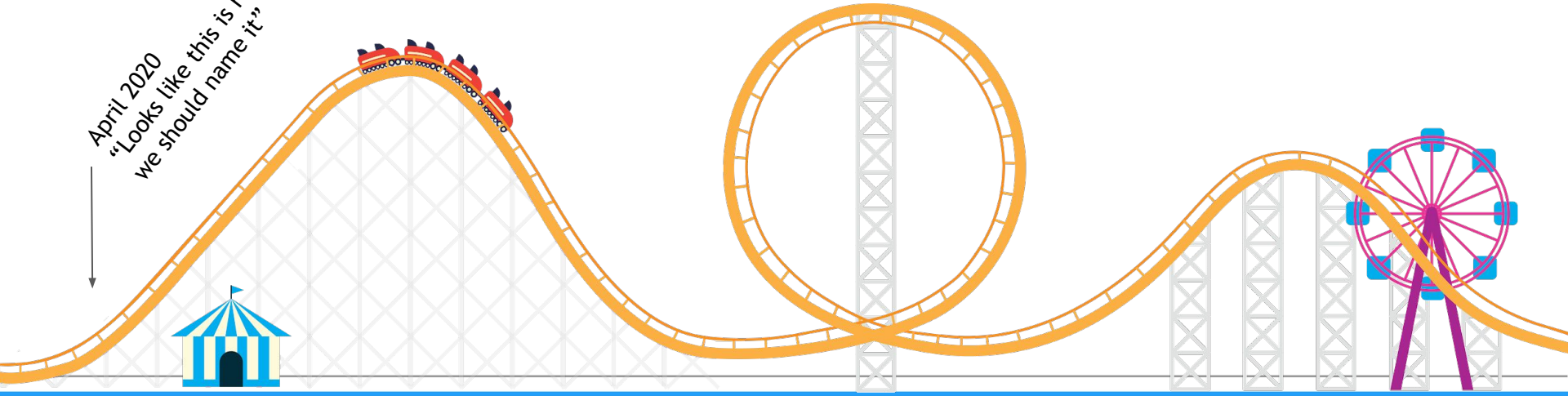
🔥 FHIR 101 - A Practical Guide

Hello there!

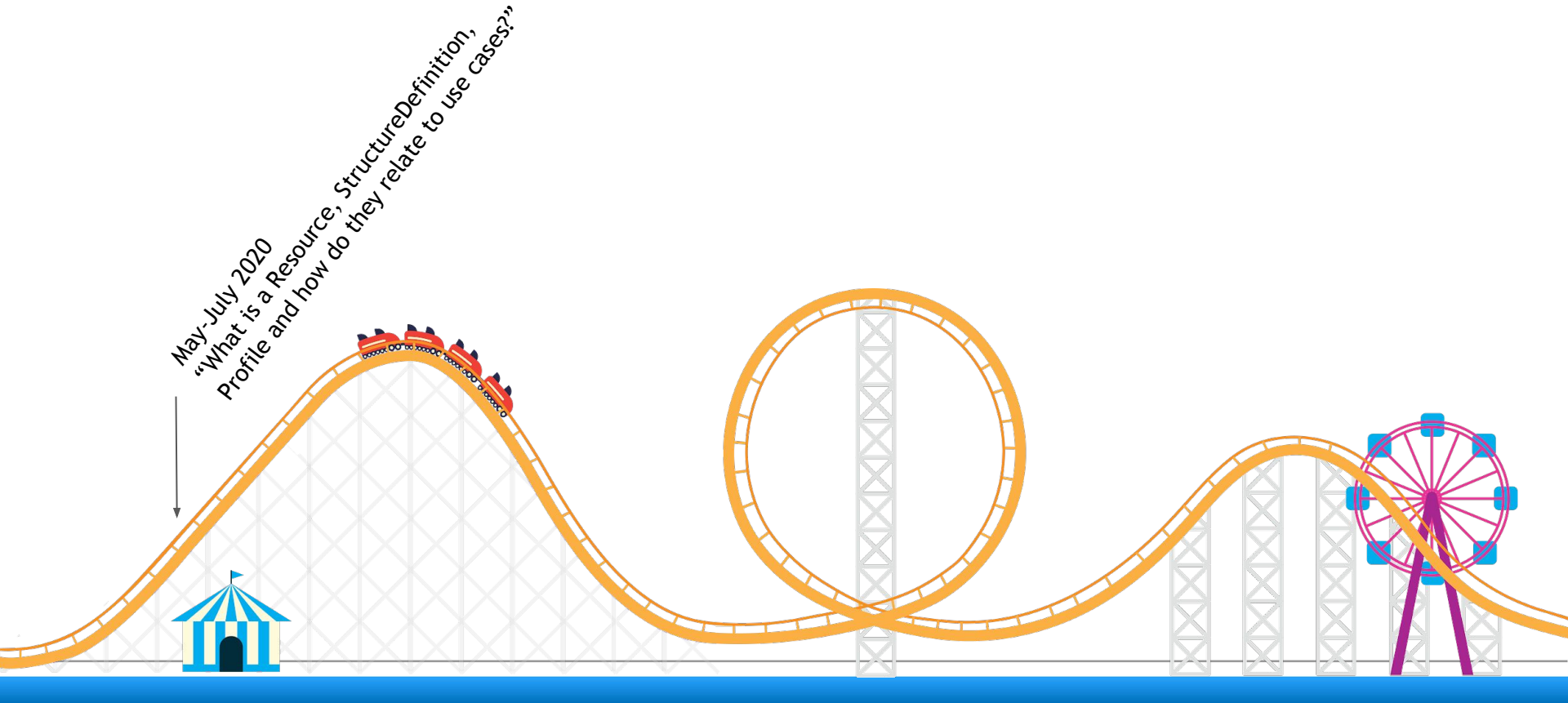
If any of the following statements applies or questions below apply to you, then this guide is for you :)

- 🤷‍♀️ I've heard the term FHIR. But is it a server? A database? A data model? Rapid oxidation?
- 🤯 I understand what FHIR is but have no idea how to begin implementing it!
- 🙁 I tried reading through HL7 FHIR, but I am still confused
- 🤔 How does FHIR modeling compare to relational database design?

April 2020
"Looks like this is really happening,
we should name it"

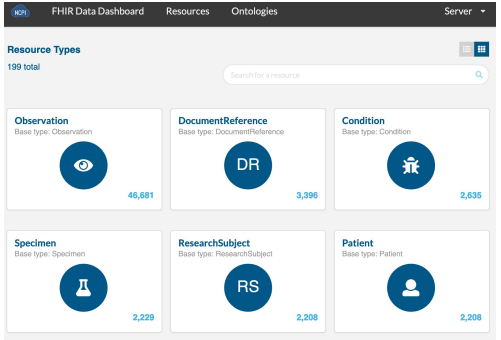


What is FHIR? Initial data: PCGC and CMG

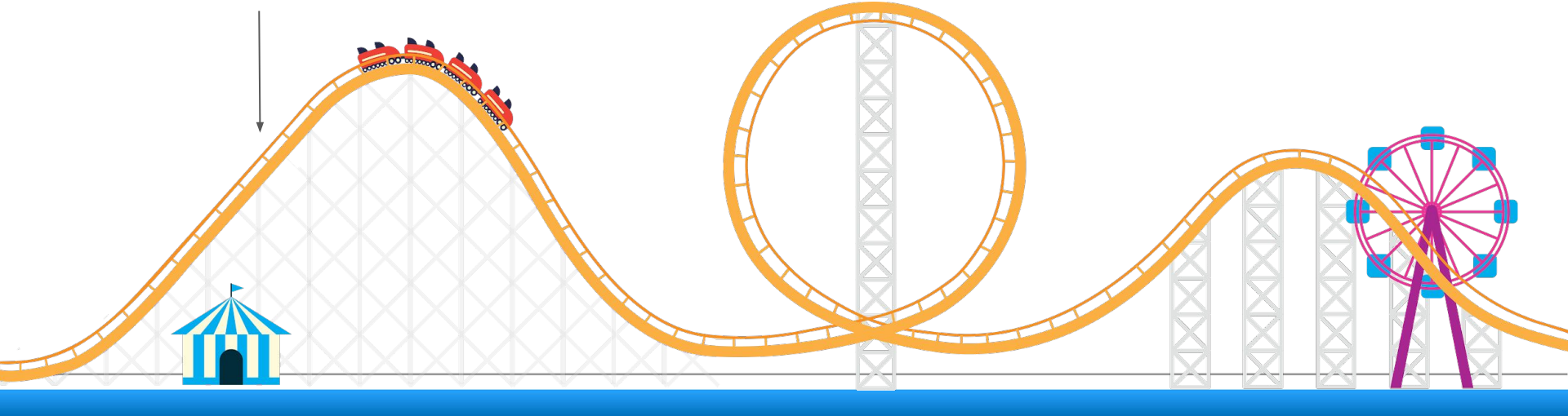


Setup Development Infrastructure

2020-07-14: ncpi-api-fhir-service-dev.kidsfirstdrc.org



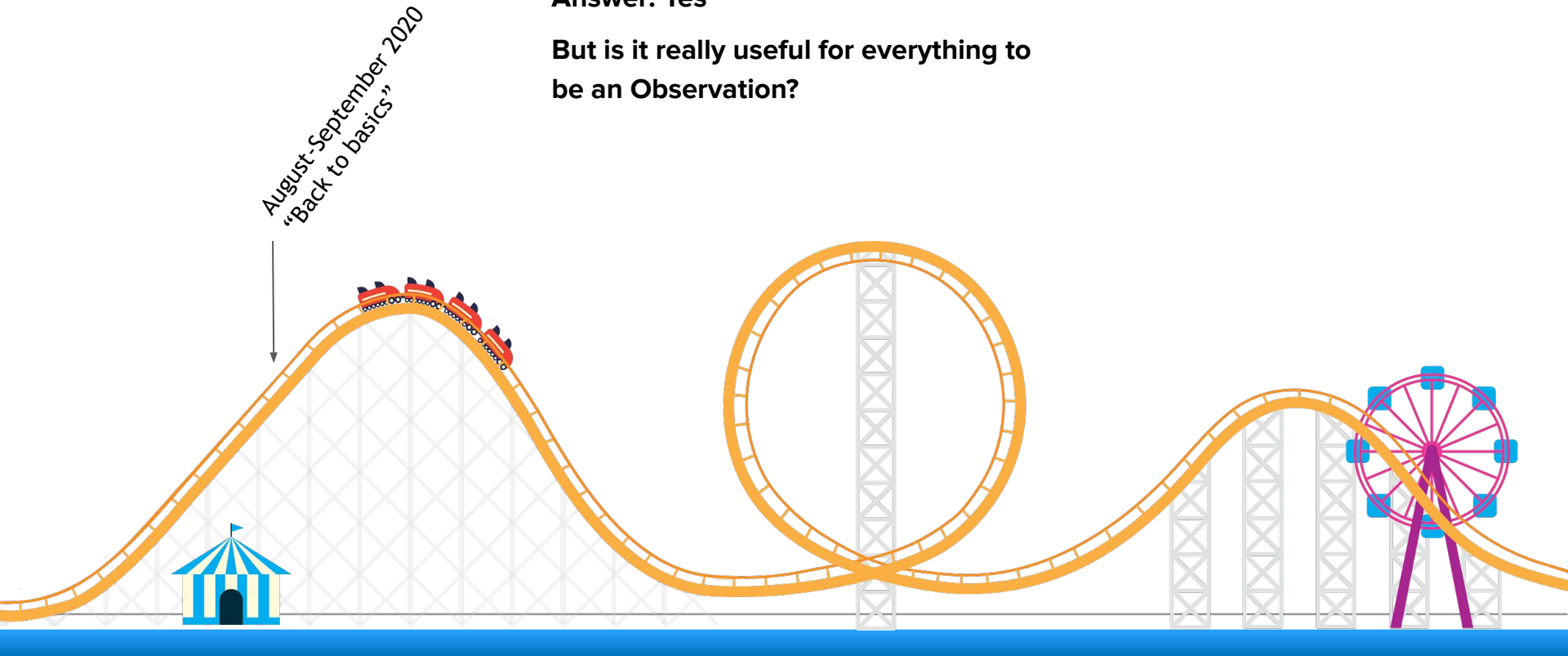
July 2020
"I think I see the top"



Can PCGC and CMG data be loaded in base FHIR?

Answer: Yes

But is it really useful for everything to be an Observation?



Initial Set of Profiles

October 2020
"Where some Observations are worth more notice"

- **Referencing DRS Objects** Model: New request — 4
  #46 by allisonheath was closed yesterday
- **Update obsolete info in Contributing section on README** bug — 1 — 1
  #44 by znatty22 was closed 5 days ago
- **Profile Disease** Model: New request — 1 — 8
  #36 by torstees was closed 3 days ago
- **Profile Human Phenotype** AnVIL Kids First DRC Model: Ready for development — 1 — 2
  #34 by torstees was closed 6 days ago
- **Profile Family Relationship** Model: New request — 1
  #33 by torstees was closed 20 days ago
- **Profile Specimen to include `DocumentReference`** Model: New request — 2
  #31 by bwalsh was closed 16 days ago
- **Profile ResearchSubject to include DocumentReferences** Model: New request — 2
  #30 by bwalsh was closed 16 days ago
- **NCPI Family Relationship** AnVIL Kids First DRC Model: Ready for development — 1 — 11
  #21 by torstees was closed 8 days ago

Demos!

- “Project Forge” Implementation Guide
- React App for browsing FHIR data
- Dash App for phenotype distribution exploration
- Shiny App for Monarch API gene search

No Really - What *is* FHIR?

FHIR is a *framework* for clinical data interoperability.

We use frameworks all the time when building platforms.
Why?

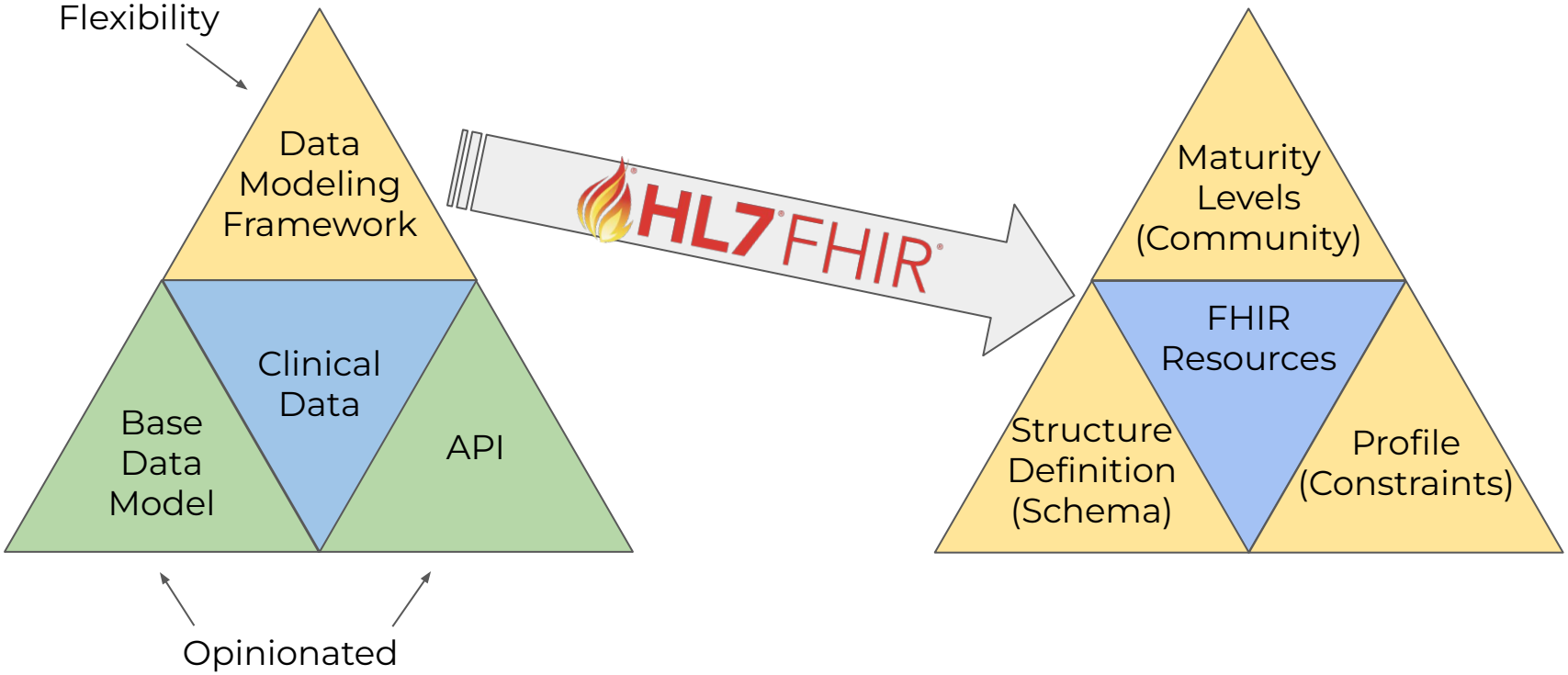
Gatsby

django

kubernetes

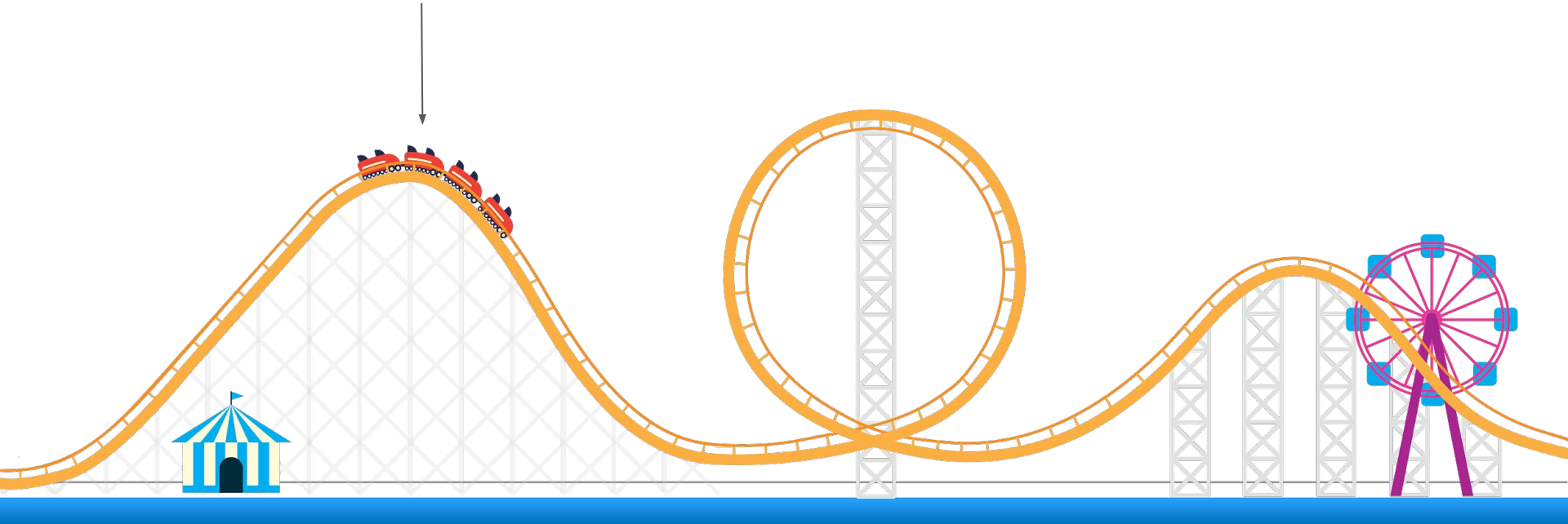
Good frameworks are opinionated where it matters to prevent effort in (re)solving recurring problems, but flexible where needed for creating solutions for new problems.

FHIR: Framework Within a Framework



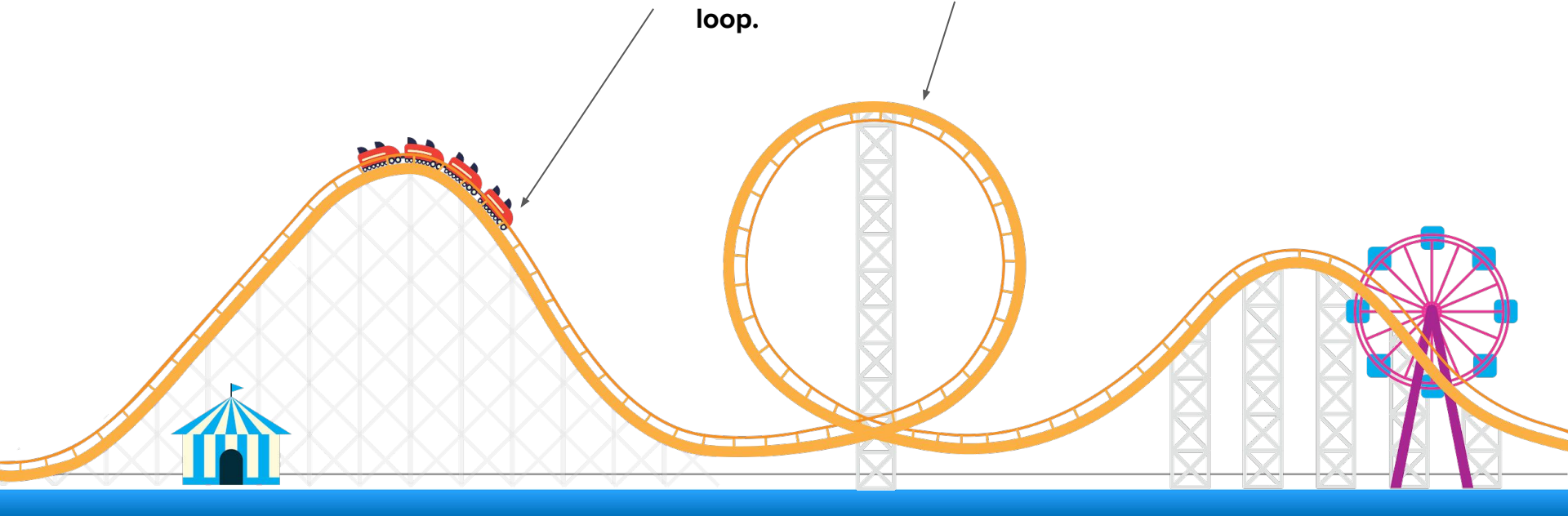
Demos Leveraged FHIR for Rapid App Development

In retrospect - the climb was understanding the FHIR framework, it's opinions, advantages, disadvantages of using it across NCPI



Demos Leveraged FHIR for Rapid App Development

The prospective is picking up momentum in building tools/libraries/apps “on FHIR” to better empower clinical data, while remaining aware there are sure to be things that throw us for a loop.

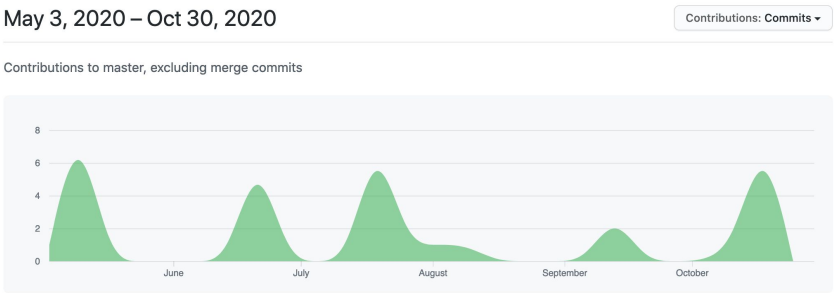


Roadmap: Data, Tools and Engagement

- Document current best practices and move from Project Forge Model to NCPI FHIR Model
  - Cancer research
  - Clinical genomics
  - <Your use case here>
- Identify key unmet needs and use cases for new tools that leverage FHIR as a framework for clinical data
  - Intake
  - Management
  - Availability
  - Interoperability
- Making these data, APIs and tools available to empower researchers is a key objective
- Community Engagement

Thank You!

Attendees of the FHIR WG Calls across all of the platforms and dbGaP!



Special thanks for the demos today:

AnVIL

- Brian Walsh
- Kristin Wuichet
- Eric Torstenson
- Katie Banasiewicz

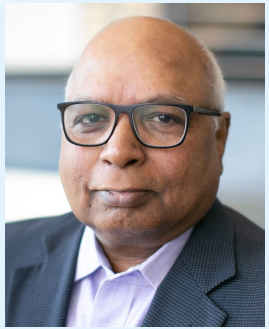
Kids First DRC

- Meen Chul Kim
- Nick Van Kuren
- Shahim Essaid
- Natasha Singh
- Avi Kelman
- Alex Lubneuski

Working Group Updates: Outreach and Training

Anton Nekrutenko
Professor, Penn State University
PD, galaxyproject.org

Ashok Krishnamurthy
RENCI
UNC, Chapel Hill



Outreach and Training WG

Goals of the working group

- Enable “cross-pollination” between the four NCPI projects by organizing regular NCPI Workshops
- Development and maintenance of the NCPI Portal
- Providing a catalogue of datasets available through each platform via NCPI Global Data Dashboard
- Providing a starting location for accessing training and outreach materials being developed and maintained by each platform as well as commonly used resources such as FHIR

Generic FHIR tutorial
(based on Kids First DRC example; http://bit.ly/fhir_nb)

FHIR_tutorial.ipynb

File Edit View Insert Runtime Tools Help Cannot save changes

Share

Table of contents

+ Code + Text Copy to Drive RAM Disk Editing

Base URL

- All HTTP requests will be sent to the HAPI public test server's FHIR version R4 base URL: http://test.fhir.org/r3 and dbGap GHIR server https://dbgap-api.ncbi.nlm.nih.gov/fhir/x1/

[2]
```
# Be sure to execute this cell so that we can use the client later
import fhirclient
from pprint import pprint
from fhirclient import client
from client import FHIRClient # FHIR client for python https://docs.smarthealthit.org/client-py/classfhirclient_1_1client_1
import urllib.request
import json
import fhirclient.models.patient as p ## import the patient datatype as p
import requests # Allows us to make GET/POST requests

settings = {
    'app_id': 'my_web_app',
    'api_base': 'http://test.fhir.org/r4'
}
smart = client.FHIRClient(settings=settings)

settingsdpgab = {
    'app_id': 'my_app',
    'api_base': 'https://dbgap-api.ncbi.nlm.nih.gov/fhir/x1'
}
dbgap = client.FHIRClient(settings=settingsdpgab)
```

Use Case 1 : Query Patient informations

You can get the Json file of any resource by using the function request_json([path of the resource])

[3]
```
smart.server.request_json('Patient')
```

```
          'postalCode': '3999',
          'state': 'Vic',
          'text': '534 Erewhon St PeasantVille, Rainbow, Vic  3999',
          'type': 'both',
          'use': 'home'}],
        'birthDate': '1974-12-25',
        'contact': [{'address': {'city': 'PleasantVille',
           'district': 'Rainbow',
           'line': ['534 Erewhon St'],
           'period': {'start': '1974-12-25'},
```

Delphine LaRiviere | special thanks to Allison Heath and Kristin Wuichet

NCPI Global Data Dashboard
(a bird's eye view of all data)

NCPI

Overview Datasets AnVIL

Search Summary

Platform	Studies	Subjects
AnVIL	21	59,325
BioData Catalyst	95	421,497
Kids First Data Resource Center	4	3,523
Cancer Research Data Commons	16	86,749
	136	571,094

Search Results

Platform	dbGap Id	Title	Diseases	Data Types	Consent Codes	Subjects
AnVIL	phs001272.v1.p1	Broad Institute Center for Mendelian Genomics	Genetic Diseases, Inborn; Bardet-Biedl Syndrome...	Genotype, SNP/CNV Genotypes (NGS)	HMB-MDS, GRU, DS-KRD-RD, DS-NIC-EMP-LENF	1,031
AnVIL	phs001913.v1.p1	CCDG - Cardiovascular: eMERGE - Northwestern Cohort	Cardiovascular Diseases	--	GRU-IRB	277
AnVIL	phs001502.v1.p1	CCDG-Cardiovascular: University of Pennsylvania Cohort	Cardiovascular Diseases	Genotype, Legacy Genotypes, SNP Genotypes (NGS)	HMB-IRB-PUB	1,373
AnVIL	phs001259.v1.p1	CCDG CVD: VIRGO - Variation in Recover-Role of Gender on Outcomes of Young Acute Myocardial Infarction (AMI) Patients	Myocardial Infarction; Inferior Wall Myocardial...	Genotype, SNP Genotypes (NGS)	DS-CARD-MDS-GSO	2,149
AnVIL	phs001894.v1.p1	CCDG-Neuropsychiatric: Autism- Genetics of Human Developmental Brain Disorders	Autism Spectrum Disorder	--	DS-EAC-PUB-GSO	724
AnVIL	phs001676.v1.p1	CCDG- Neuropsychiatric: Autism - Simons Simplex Collection (SSC)	Autism Spectrum Disorder	--	DS-AONDD-IRB	9,201
AnVIL	phs001740.v1.p1	CCDG- Neuropsychiatric: Autism- Study of Autism Genetics Exploration (SAGE)	Autism Spectrum Disorder	Genotype, SNP/CNV Genotypes (NGS)	DS-ASD-RD-IRB	580
AnVIL	phs001741.v1.p1	CCDG- Neuropsychiatric: Autism- The Autism Simplex Collection	Autism Spectrum	Genotype, SNP/CNV	DS-ASD-IRB	905

David Rogers / Kevin Osborne | special thanks to Garrett Rupp (UChicago) and Michael Feolo (NCBI/dbGaP)

NCPI Global Data Dashboard (magic and challenges)

- dbGaP (only) entries from all platforms
- Derived from a static spreadsheet at this time
- Uses dbGaP FTP/XML interface and dbGaP FHIR API for additional info
- dbGaP FHIR team is modifying APIs and is pleasure to work with
- Planning to use GA4GH Discovery API in the future

David Rogers | special thanks to Garrett Rupp (UChicago) and Michael Feolo (NCBI/dbGaP)

A Unified tutorial dashboard
(a landing page for all NCPI tutorials)

Core

These are the core, foundational topics for learning how to use Galaxy.

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to Galaxy						
A short introduction to Galaxy						
From peaks to genes						
Galaxy 101						
Galaxy 101 for everyone						
Introduction to Genomics and Galaxy						
NGS data logistics						
Options for using Galaxy						

A Unified tutorial dashboard
(a landing page for all NCPI tutorials)

Tutorial	**AnVIL**	**CRDC**	**KF**	**BDC**
Calling variants	✅		✅	
Cleaning variant calls	✅		✅	
Interpreting variants	✅	✅	✅	✅

A mockup of the training dashboard (will be housed at the NCPI portal)

Quick Break

We will resume at 3:10 pm ET.

Group Discussion
Drafting a Road Map

Allison Heath
Children's Hospital of Philadelphia
Brian O'Connor
Broad Institute

Other Template Slides

Feel Free to Copy/Paste as Needed

This Is Where the Title or Headline Goes.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

This Is Where the Title or Headline Goes.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

This Is Where the Title or Headline Goes.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

This Is Where the Title or Headline Goes.

Compared Subject #1

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Compared Subject #2

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

This Is Where the Title or Headline Goes.

Compared Subject #1

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Compared Subject #2

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.