

# Knowledge Systems @ DFCI

Ethan Cerami, Ph.D.  
Director, Knowledge Systems Group



**James Lindsay, Ph.D.**  
**Associate Director, Knowledge Systems Group**

cBioCenter @ DFCI

# DFCI Knowledge Systems Group, cBioCenter

Applied genomics software and data science group.

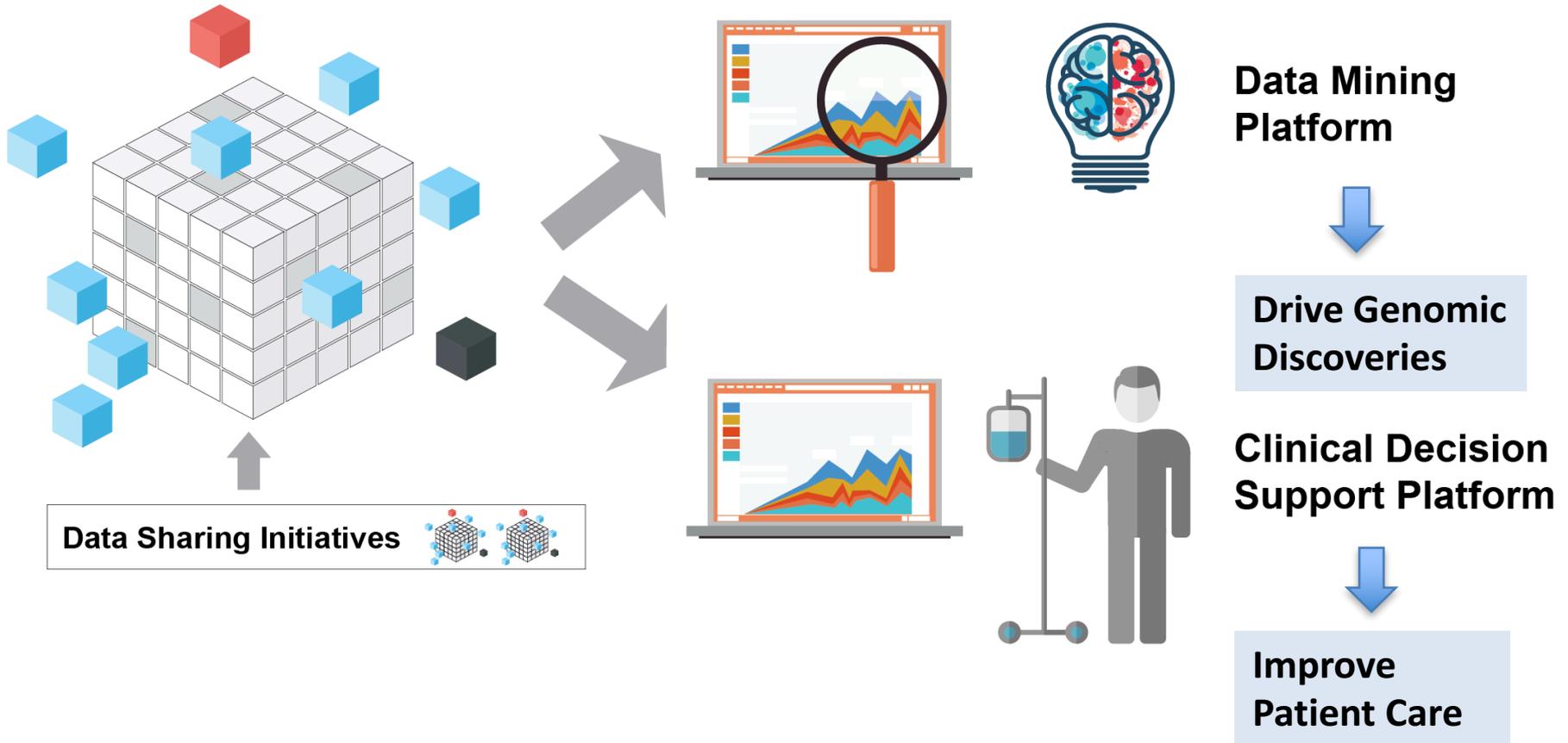
Building genomic software to enable:

- cancer genomics research
- clinical application of genomic data.



Part of the new cBioCenter @ DFCI (Head: Chris Sander)

# Enterprise Genomics and Data Science @ DFCI



Data acquisition

Data utilization

# Highlighted projects

---

## Data utilization

- cBioPortal
- MatchMiner
- Insight engine

## Data acquisition

- cBioOne
- AACR GENIE
- Intel CCC



DANA-FARBER  
CANCER INSTITUTE

# about

---

- Founded in 1947 by Sidney Farber
- > 4,500 employees
- > 450,000 patient visits / yr
- 900 clinical trials
  - > 500 treatment trials
- percent\_research <- 0.5
- percent\_clinical <- 0.5



Sidney Farber, MD, with a young patient

# Enterprise genomics: PROFILE

- 15,927 patients sequenced
- Targeted DNAseq
- 447 genes / regions targeted
- CLIA-certified

DANA-FARBER/BRIGHAM AND WOMEN'S



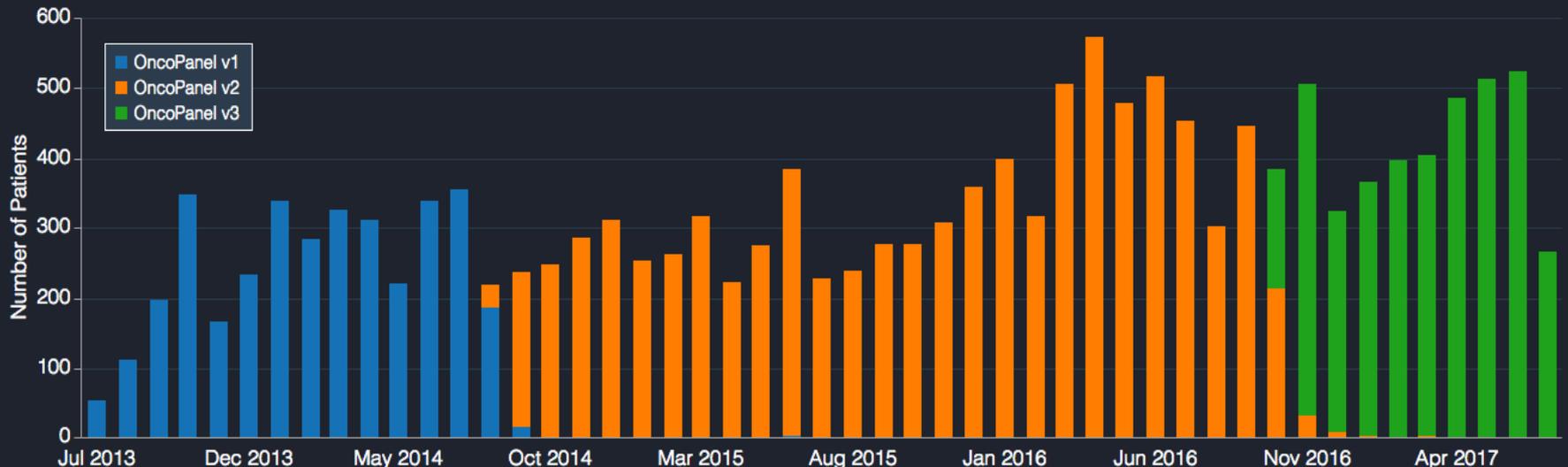
CANCER CENTER



Boston  
Children's  
Hospital

Until every child is well™

## Sequencing History

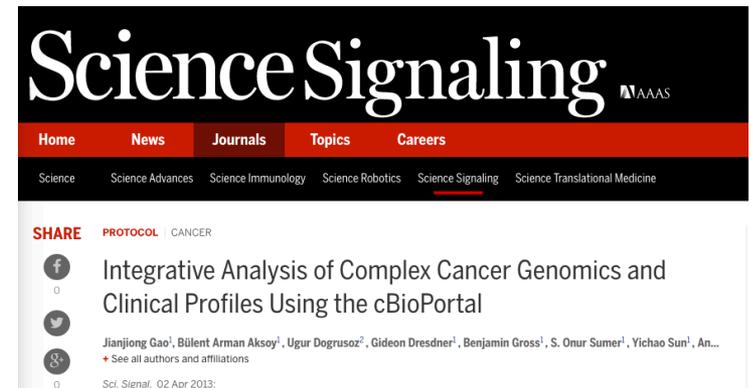


---

cBioPortal, Insight Engine, MatchMiner

# **DATA UTILIZATION**

# cBioPortal



**Science Signaling** AAAS

Home News Journals Topics Careers

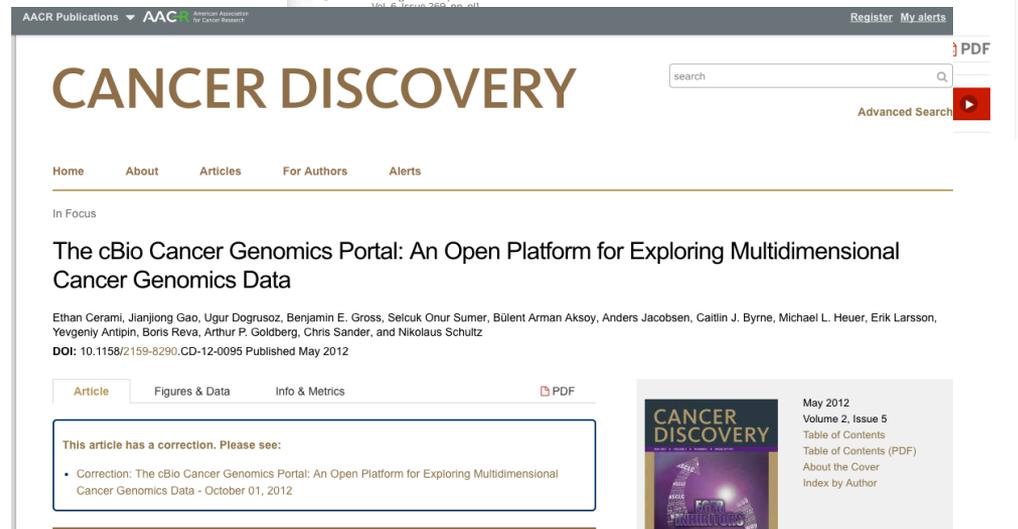
Science Science Advances Science Immunology Science Robotics Science Signaling Science Translational Medicine

SHARE PROTOCOL CANCER

**Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal**

Jianjiong Gao<sup>1</sup>, Bülent Arman Aksoy<sup>1</sup>, Ugur Dogrusoz<sup>2</sup>, Gideon Dresdner<sup>1</sup>, Benjamin Gross<sup>1</sup>, S. Onur Sumer<sup>1</sup>, Yichao Sun<sup>1</sup>, An...  
+ See all authors and affiliations

Sci. Signal. 02 Apr 2013; 10(23):ra111



AACR Publications AACR American Association for Cancer Research

Register My alerts

**CANCER DISCOVERY** search PDF

Advanced Search

Home About Articles For Authors Alerts

In Focus

**The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data**

Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P. Goldberg, Chris Sander, and Nikolaus Schultz

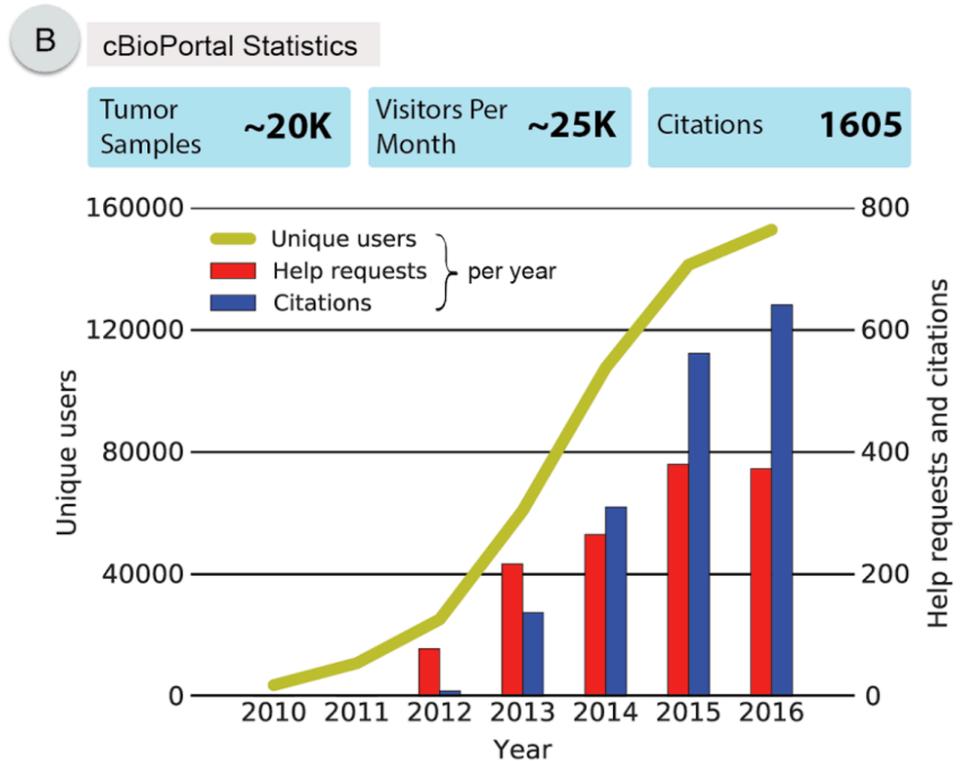
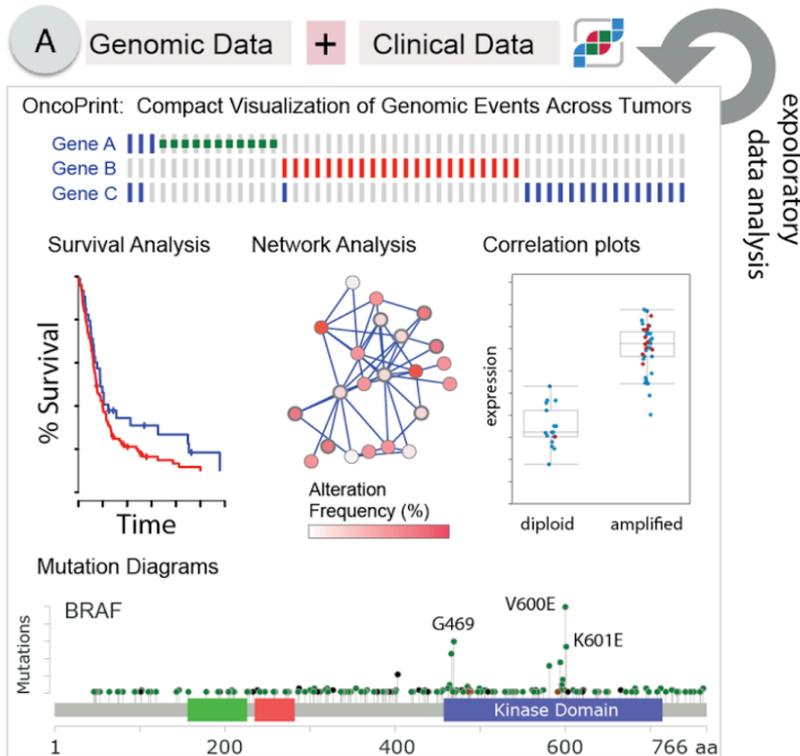
DOI: 10.1158/2159-8290.CD-12-0095 Published May 2012

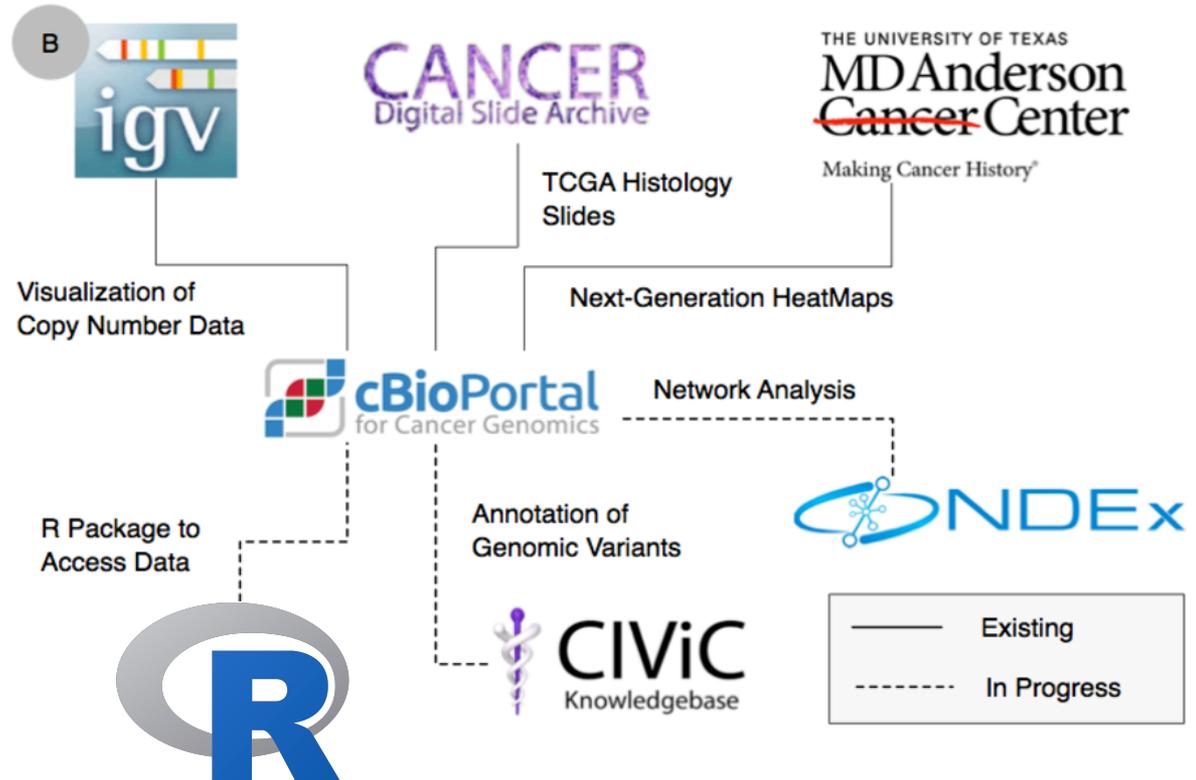
Article Figures & Data Info & Metrics PDF

This article has a correction. Please see:

- Correction: The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data - October 01, 2012

May 2012  
Volume 2, Issue 5  
Table of Contents  
Table of Contents (PDF)  
About the Cover  
Index by Author





# R package

## Installation

1. The CDGS-R package currently **only works with R Version 2.12 or higher**.
2. Then install the cgds-R package from within R: `install.packages('cgdsr')`

## Example usage

```
# Create CGDS object
mycgds = CGDS("http://www.cbioportal.org/public-portal/")

test(mycgds)

# Get list of cancer studies at server
getCancerStudies(mycgds)

# Get available case lists (collection of samples) for a given cancer study
mycancerstudy = getCancerStudies(mycgds)[2,1]
mycaselist = getCaseLists(mycgds,mycancerstudy)[1,1]

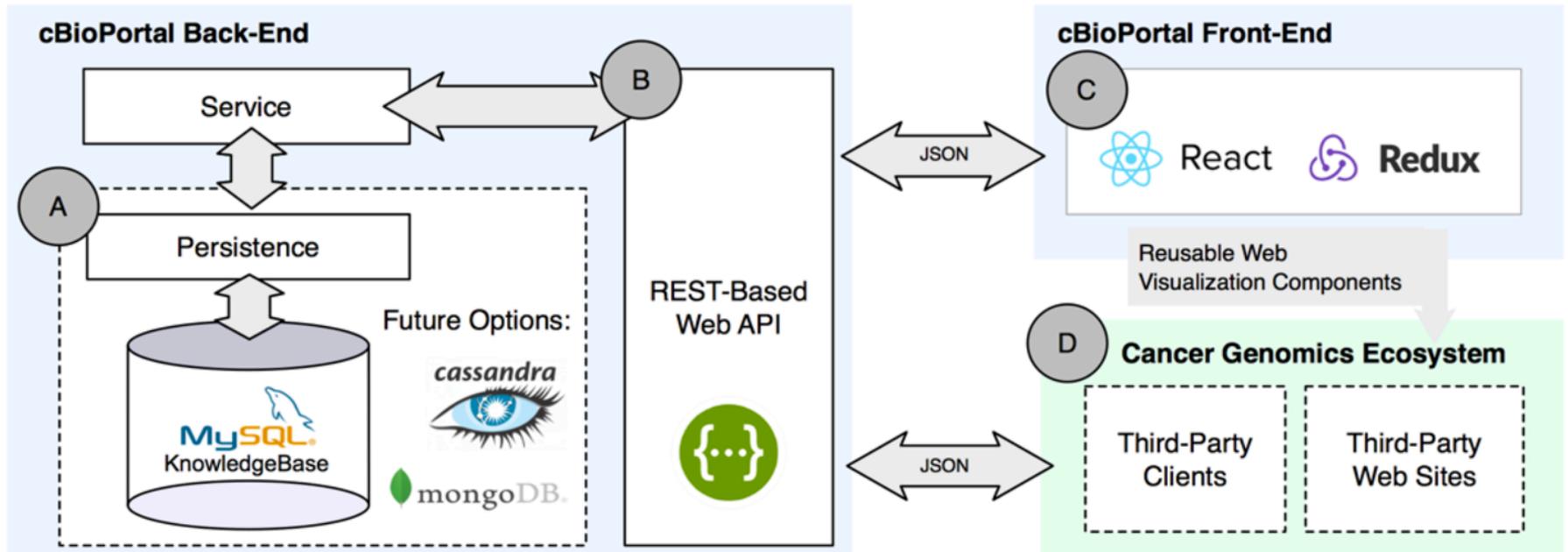
# Get available genetic profiles
mygeneticprofile = getGeneticProfiles(mycgds,mycancerstudy)[4,1]

# Get data slices for a specified list of genes, genetic profile and case list
getProfileData(mycgds,c('BRCA1','BRCA2'),mygeneticprofile,mycaselist)

# Get clinical data for the case list
myclinicaldata = getClinicalData(mycgds,mycaselist)

# documentation
help('cgdsr')
help('CGDS')
```

# New Architecture



# Virtual Cohorts

## A Virtual Cohort Builder

### MY COHORTS

ER+ Breast (TCGA + ICGC)

### CANCER TYPE

- Adrenal
- Bladder
- Breast
- Head and Neck
- Lung

### CANCER STUDY

Faceted Filtering to Create Virtual Cohort

Summary of Virtual Cohort Characteristics

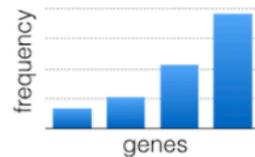
1,101 cases

Save

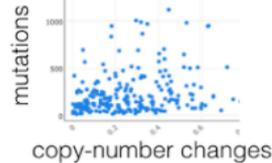
Share

Compare

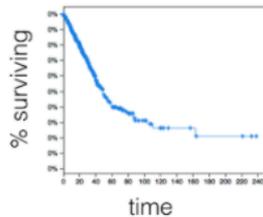
### FREQUENTLY MUTATED GENES



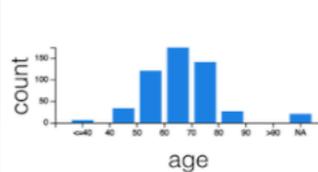
### GENOMIC INSTABILITY



### SURVIVAL



### AGE



Further Filtering via Interactive Plots

## B Cohort Comparison

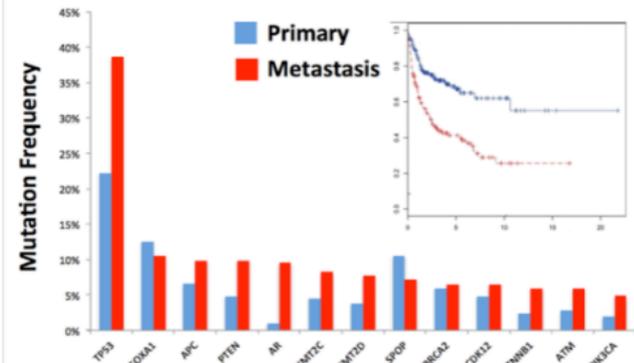
COHORT A

Primary

COHORT B

Metastasis

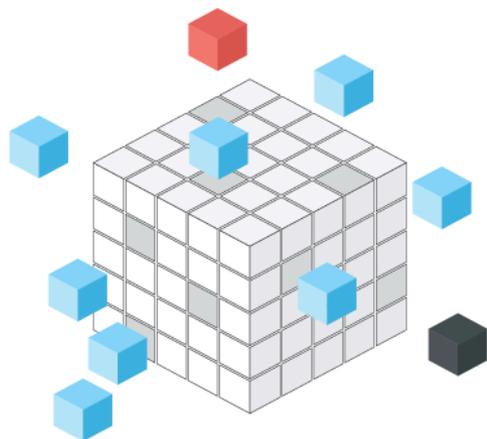
### EXAMPLE COMPARISON





# DFCI Insight Engine

<http://insight.dfci.harvard.edu>



Profile + RHP Data

De-identified  
clinical data

GENIE Data

Weekly Updates



Reproducible  
Pipelines +  
Jupyter  
Notebooks

Community Driven

Reports

Mutation Hotspots

Clinical Actionability

Trial Enrollment

Germline Analysis

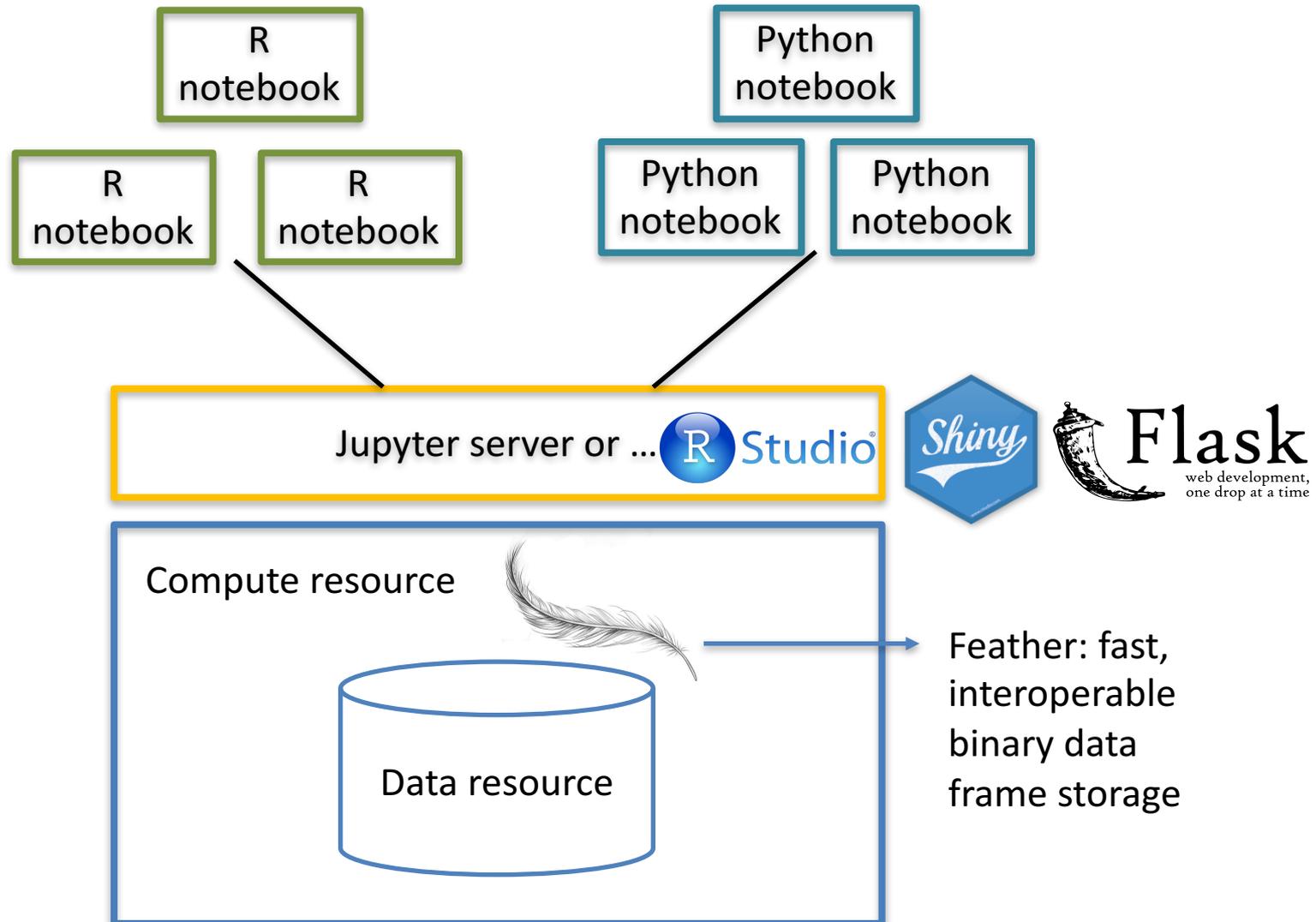
Patient Similarity

Network Analysis

Mutational Load

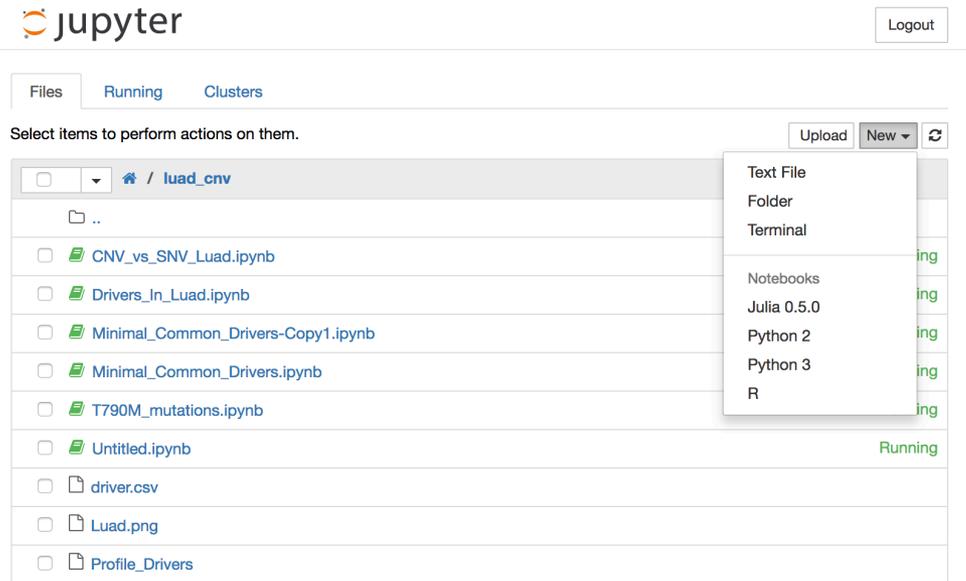
Future Profile Papers

# Compute, data and code in same place



# Remove barriers for data science

1) Create new notebook in language of choice

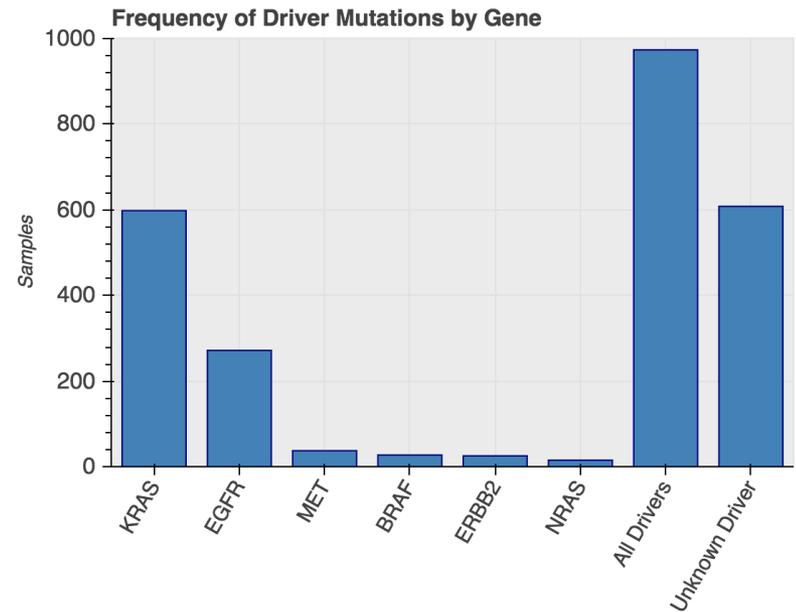


The screenshot shows the JupyterLab interface. At the top, there is a 'Logout' button. Below it, there are tabs for 'Files', 'Running', and 'Clusters'. The main area is a file browser for the directory '/ luad\_cnv'. A 'New' dropdown menu is open, showing options: Text File, Folder, Terminal, Notebooks, Julia 0.5.0, Python 2, Python 3, and R. The file browser lists several files: '..', 'CNV\_vs\_SNV\_Luad.ipynb', 'Drivers\_In\_Luad.ipynb', 'Minimal\_Common\_Drivers-Copy1.ipynb', 'Minimal\_Common\_Drivers.ipynb', 'T790M\_mutations.ipynb', 'Untitled.ipynb', 'driver.csv', 'Luad.png', and 'Profile\_Drivers'.

2) Pull in data via programmatic interface

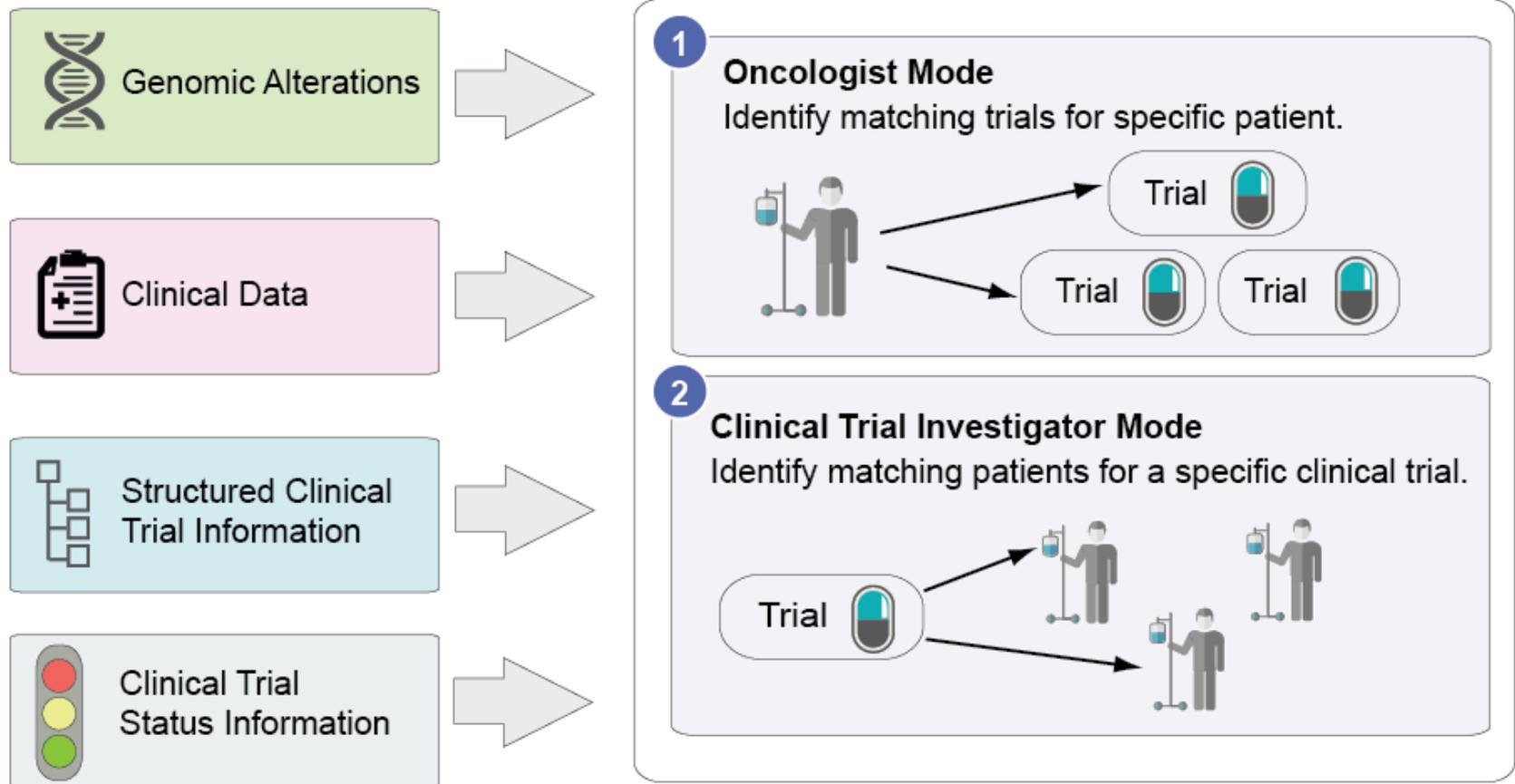
```
genomic_df = data.get_genomic_df()
clinical_df = data.get_clinical_df()
cna_df = data.get_cna_df()
cna_df=cna_df.set_index('SAMPLE_ID')
merged_df = genomic_df.join(clinical_df, on="Tumor_Sample_Barcode")
```

3) ... science?



# MatchMiner

## MatchMiner: Open Source Clinical Trial Matching Platform



# Patient View – Trial Matches

BETA



SIGN OUT

REPORT BUG

/ MATCHMINER / DASHBOARD / PATIENT RECORD

PATIENT

TRIAL MATCHES

ONCOPANEL

Smith[Fake], Raquel

MRN MB-0657  
Date of Birth 27 Jan 1958  
Primary Cancer Type Breast Invasive Ductal Carcinoma  
Biopsy Site Breast  
Biopsy Site Type Unspecified  
Sample ID BRCA-METABRIC-S1-MB-0657  
Report Date 11 Feb 2013

### Clinical trial matches



MatchMiner has identified 5 potential precision medicine clinical trial matches based on genomic profiling results obtained from OncoPanel.

All trials within MatchMiner have been expertly curated to capture genomic and basic clinical eligibility details, and matches have been computed based on the patient's genomic profile, tumor type, age, and sex.

As these criteria represent only a subset of all trial eligibility criteria, additional investigation and screening should be conducted to determine final eligibility.

### Results

Genomic match ↑	Protocol #	Disease Center	Coordinating Center	DFCI Trial Status
PIK3CA p.H1047R Variant-Level Match Tier 4	<b>TASELISIB+FULVESTRANT VS PLACEBO+FULVESTRANT FOR BRCA</b> <i>Protocol No:</i> 15-153 <i>Principal Investigator:</i> Krop, Ian E	DF/HCC Breast Cancer	Dana-Farber Cancer Institute	<b>OPEN TO ACCRUAL</b>
<p>PIK3CA wt PIK3CA Invasive Breast Carcinoma Adults Locally Advanced Metastatic Recurrent FULVESTRANT PLACEBO TASELISIB Phase III ...</p>				
PIK3CA p.H1047R Gene-Level Match Tier 4	<b>LETROZOLE +/- BYL719 OR BUPARSILIB FOR HER2-NEGATIVE BREAST CANCER</b> <i>Protocol No:</i> 14-202 <i>Principal Investigator:</i> Mayer, Erica L	DF/HCC Breast Cancer	Dana-Farber Cancer Institute	<b>OPEN TO ACCRUAL</b>
<p>PIK3CA wt PIK3CA Breast Adults Untreated Localized Locally Advanced BUPARLISIB BYL719 LETROZOLE Phase II ...</p>				
PIK3CA p.H1047R Gene-Level Match Tier 4	<b>GDC-0032 + DOCETAXEL OR PACLITAXEL FOR BREAST</b> <i>Protocol No:</i> 13-123 <i>Principal Investigator:</i> Krop, Ian E	Simulated Data Cancer	Institute	<b>OPEN TO ACCRUAL</b>

FILTERS

Gene (Mutant)

PIK3CA (4)

Gene (Wildtype)

WT KRAS (1)

WT PIK3CA (3)

WT TP53 (1)

Tumor Type

All Solid Tumors (2)

Breast (2)

Colorectal Adenocarcinoma (1)

Invasive Breast Carcinoma (2)

# Trial centric – Create filter

GENOMIC FILTER EDITOR FILTERS

Edit genomic filter - ERBB2 Mut

> GENOMICS   CLINICAL   GENERAL

### Genomic attributes

↔ Gene

ERBB2 X ABL1, EGFR etc.

🔄 Protein Change

Protein change

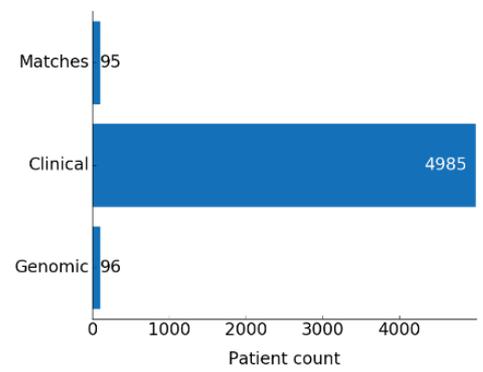
≡ Exon Number

Transcript exon

✕ Genomic Alteration Type

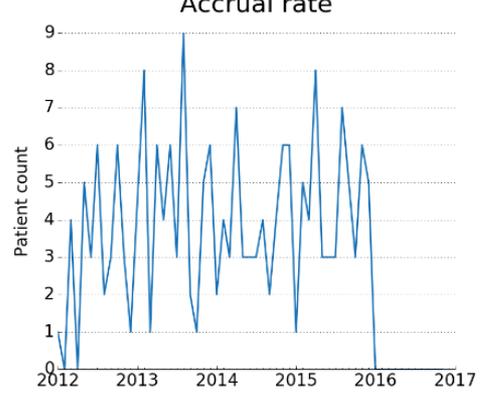
- Mutation
- High level amplification
- Gain
- Homozygous deletion
- Structural Rearrangement

### Patient Matches



Category	Patient count
Genomic	96
Clinical	4985
Matches	95

### Accrual rate



2012 2013 2014 2015 2016 2017

Simulated Data

< PREVIOUS   UPDATE   NEXT >

# Clinical Trial Markup Language

## Clinical Trial Details

```
nct_id: NCT02097225
nct_purpose: This phase I trial studies the side effects ...
phase: I
protocol_id: 6534
protocol_no: 14-186
protocol_target_accrual: 32
protocol_type: Treatment
short_title: AT13387 W/ DABRAFENIB + TRAMETINIB IN BRAF-MUTANT MELANOMA
status: Open to Accrual
```

### treatment\_list:

#### step:

##### - arm:

##### match:

##### - and:

##### - and:

##### - or:

##### - genomic:

```
hugo_symbol: BRAF
protein_change: p.V600E
variant_category: Mutation
```

##### - genomic:

```
hugo_symbol: BRAF
protein_change: p.V600K
variant_category: Mutation
```

##### - genomic:

```
hugo_symbol: KRAS
wildtype: true
```

##### - genomic:

```
hugo_symbol: NRAS
wildtype: true
```

##### - clinical:

```
age_numerical: '>=18'
oncotree_primary_diagnosis: _SOLID_
```

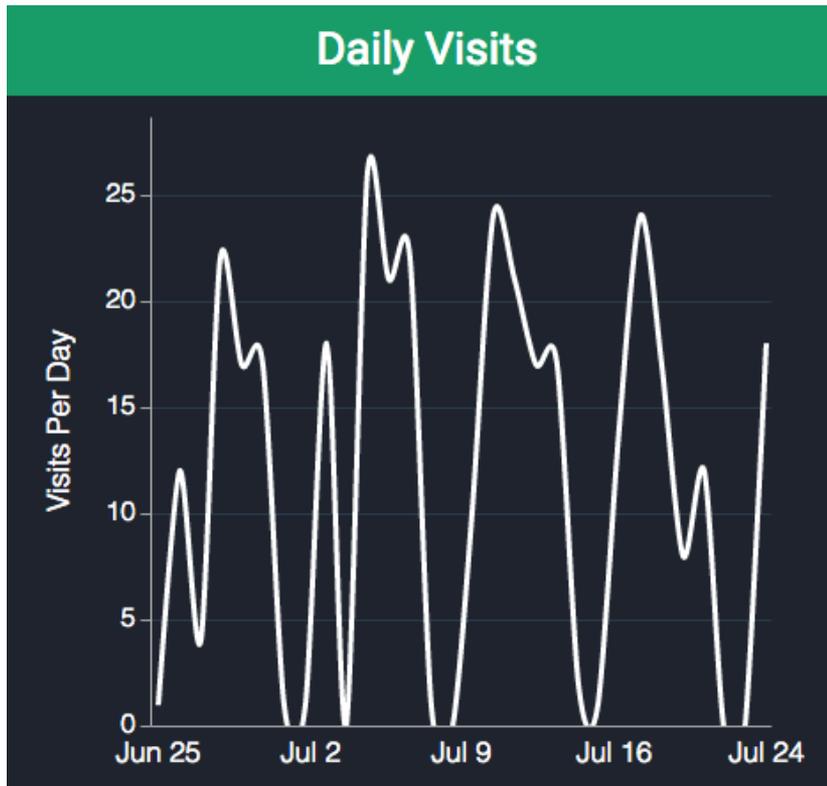
## Eligibility Criteria

## Genomic Criteria

## Clinical Criteria

# MatchMiner usage stats

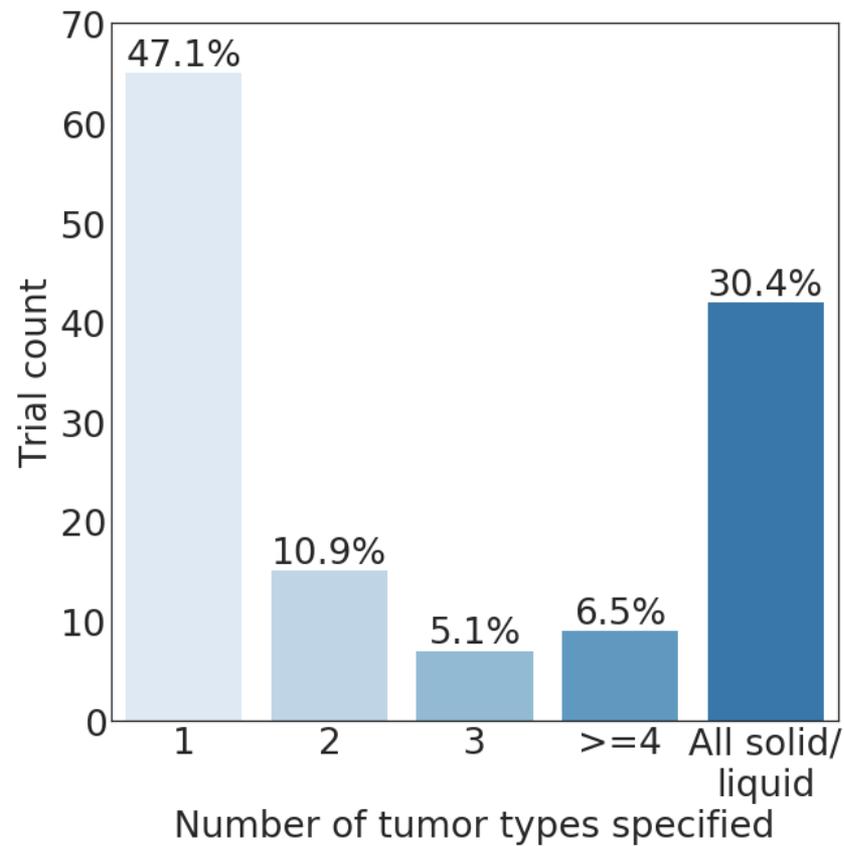
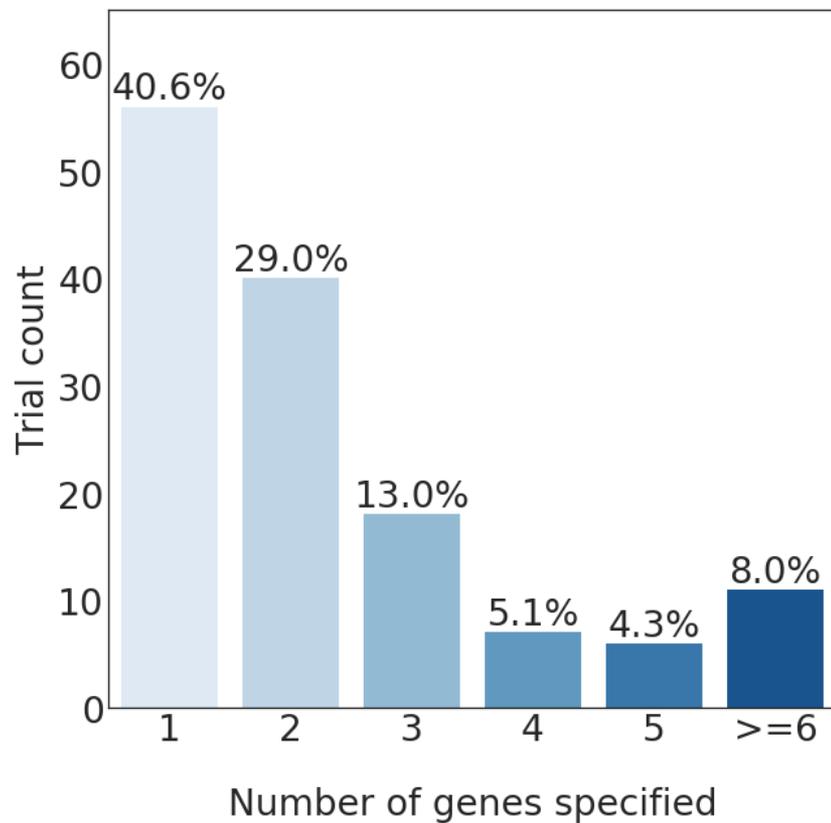
138 Trials, 675 register oncologists, 125 clinical trial investigators



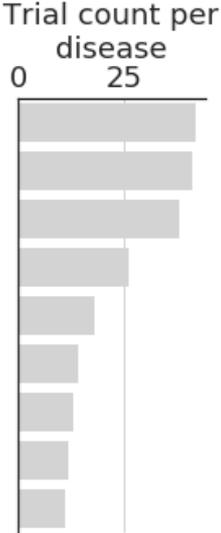
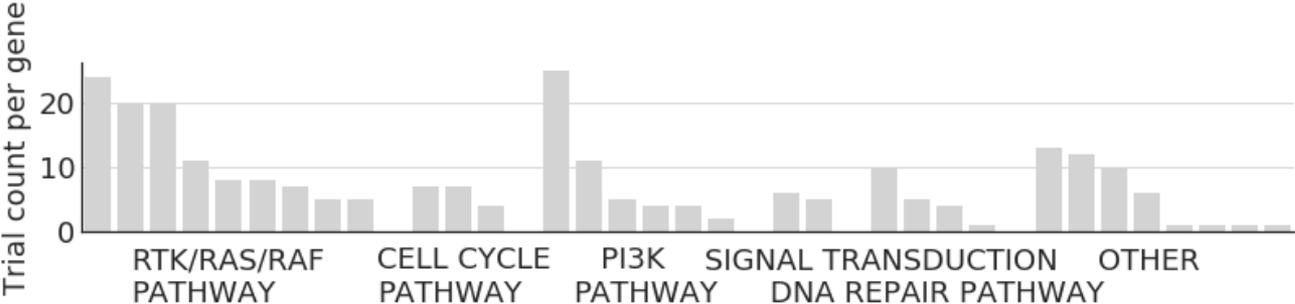
### Search Terms (July)

Term	Count
EGFR	6
16-494	4
16-711	3
demetri	3
(cdkn2a)	2
(cdkn2a,cdkn2b)	2
(cdkn2b)	2
(pik3ca)	2
(pik3ca,All solid tumors)	2
(pik3ca,All solid tumors,Open to Accrual)	2

# Trial complexity of 138 curated trials



# Genomic criteria overview



Solid Tumors  
 Non-Small Cell Lung Cancer  
 Leukemia  
 Breast Carcinoma  
 Colorectal Cancer  
 Glioma  
 Non-Hodgkin Lymphoma  
 Melanoma  
 Other



Criteria counts

0 8 16 24 32

N/KRAS  
 ALK  
 BRAF  
 MISC  
 MET  
 NTRK1/2/3  
 ROS1  
 FLT3  
 KIT

MYC  
 CELL CYCLE MISC  
 CCND1/2/3

EGFR  
 PIK3CA/CB  
 PTEN  
 ERBB2  
 AKT1/3  
 TSC1/2

FGFR1/2/3/4  
 PDGFRA/B

BRCA1/2  
 ABL1  
 ATM  
 POLE

IDH1/2  
 REGULATION/TRANSCRIPTION  
 TP53 PATHWAY  
 SIGNALING PATHWAYS  
 NOTCH PATHWAY  
 SPPLICISOME  
 ESR1  
 ARID1A

cBioOne, Intel CCC, GENIE

# **DATA ACQUISITION**



## cBioPortal “as a service”

- for individual investigators
- for disease centers



Profile Data

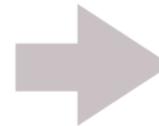
Other  
Genomic Data

Clinical Data



Secure Live Link File Server

Validate,  
Merge +  
De-Identify



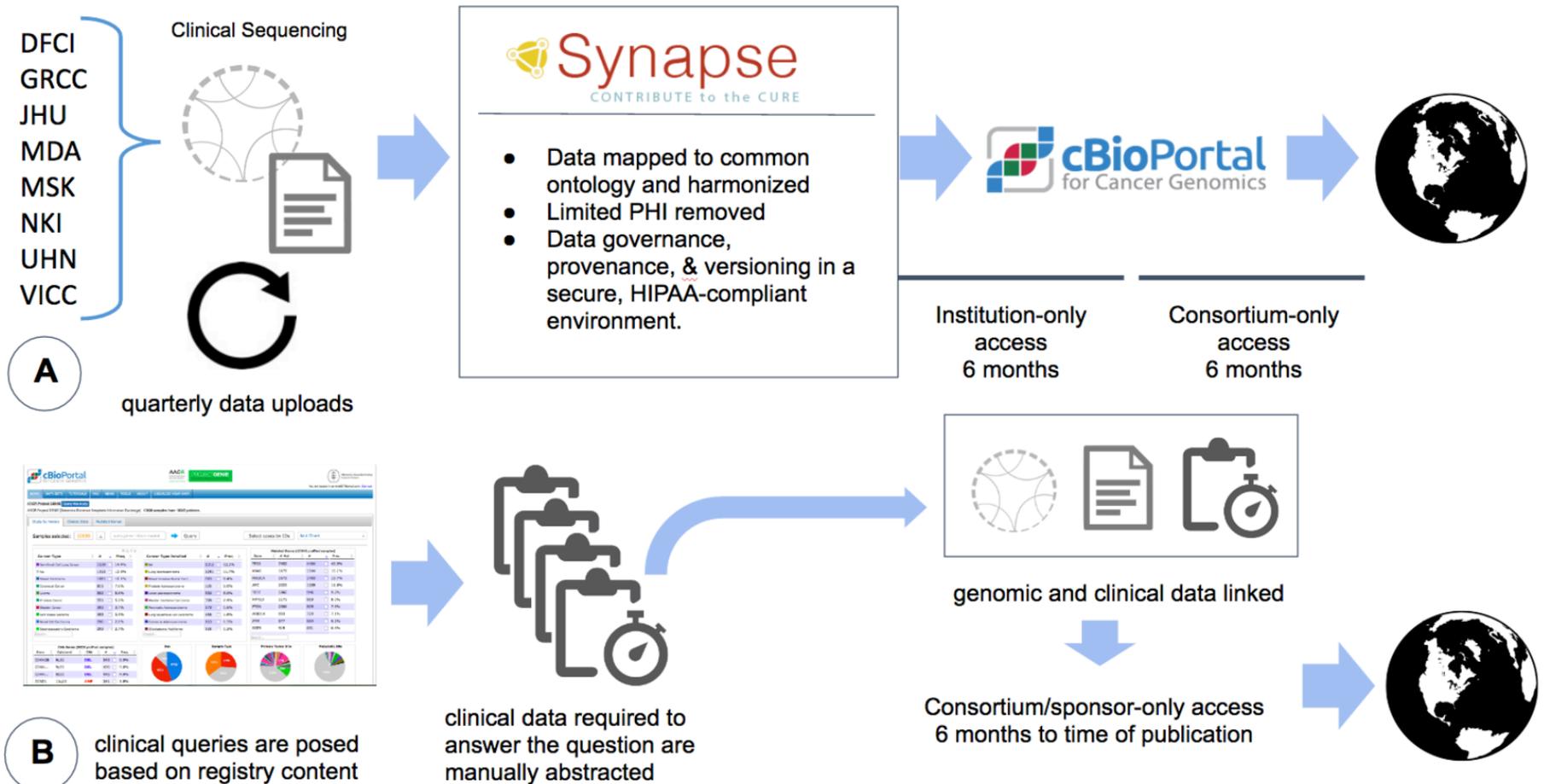
Private to Group

Shared @ DFCI



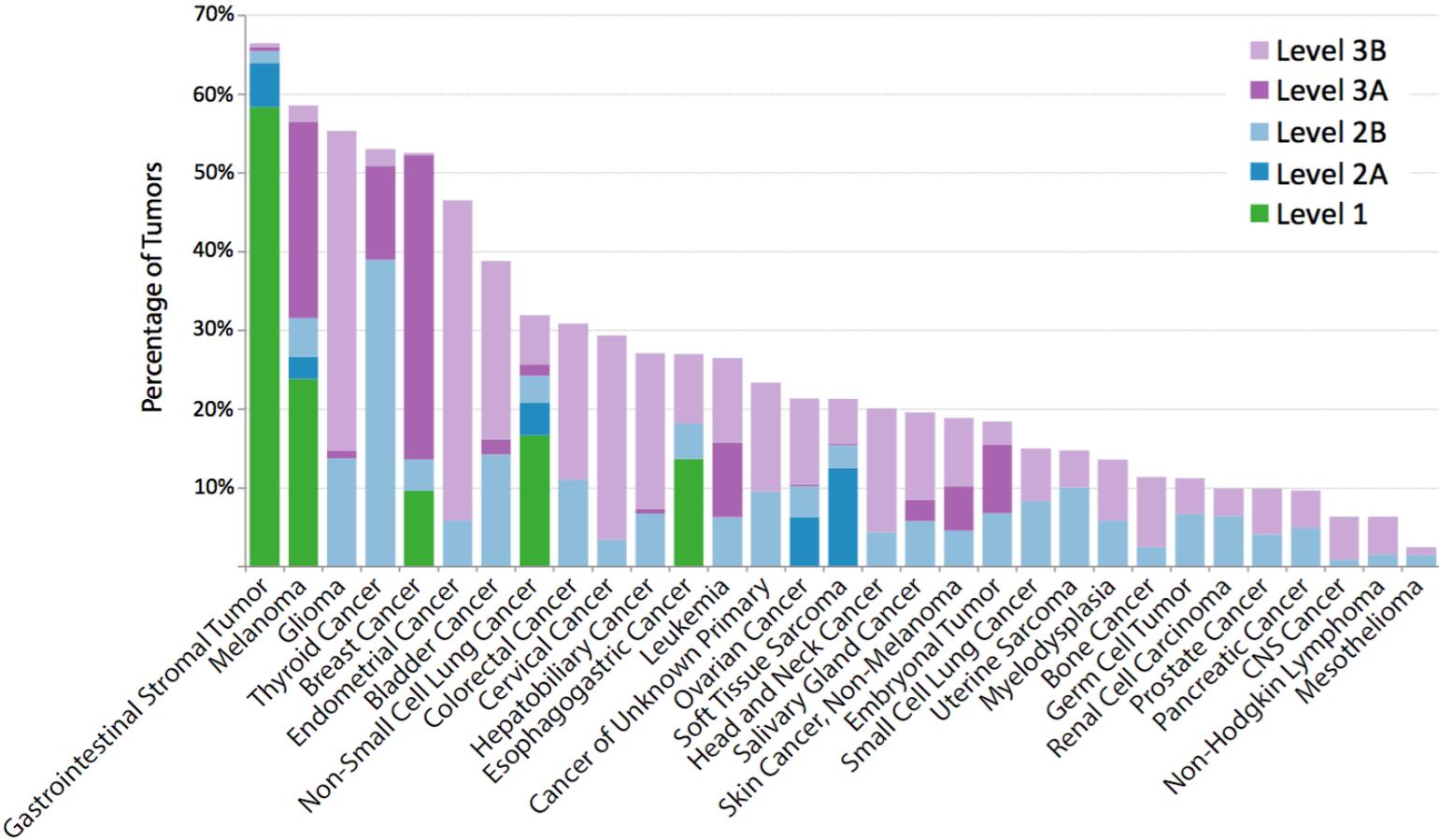
**7 PIs / Groups Now Active**

# AACR GENIE: First Data Release!

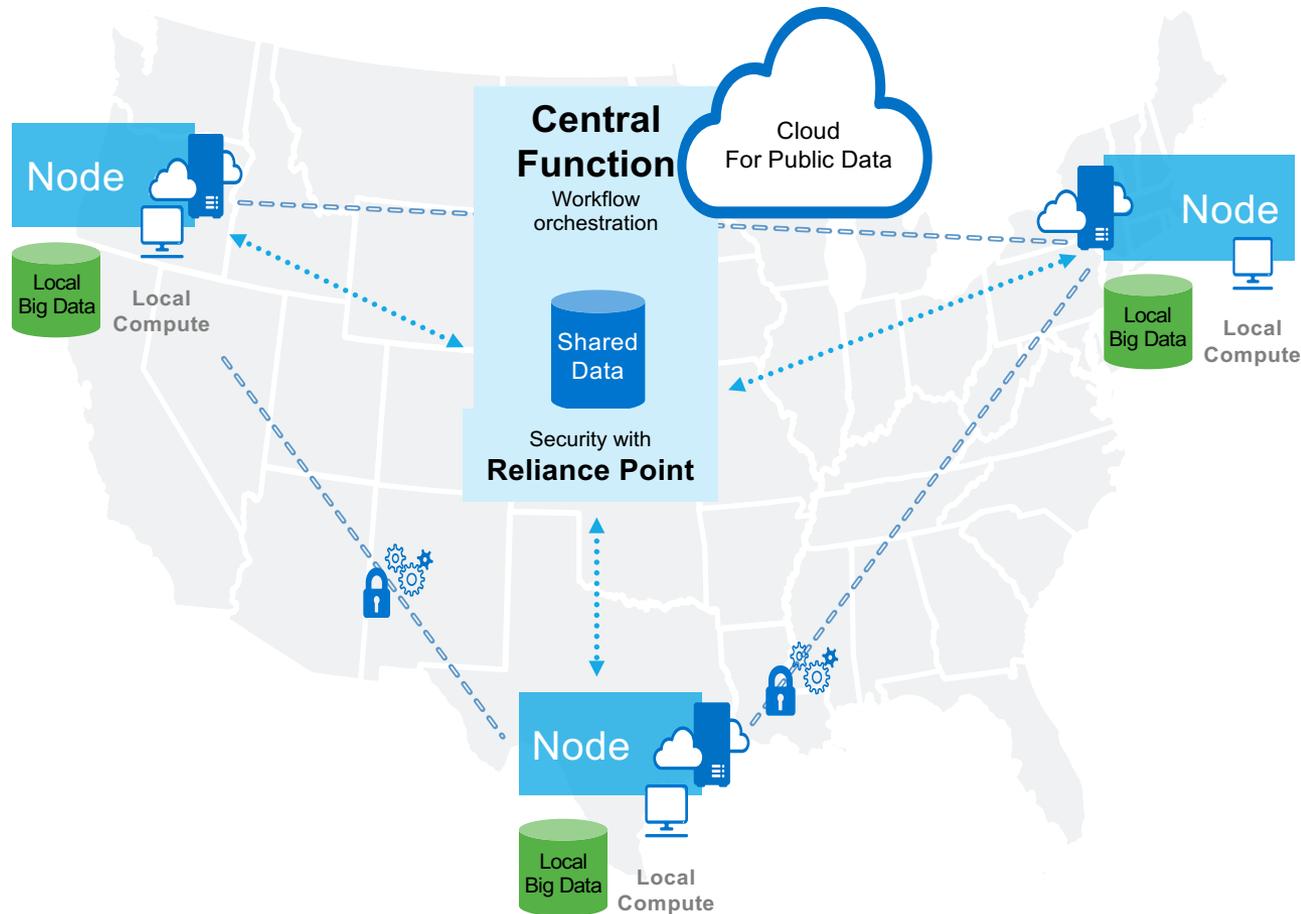




# Landscape of Clinical Actionability



# Intel Collaborative Cancer Cloud

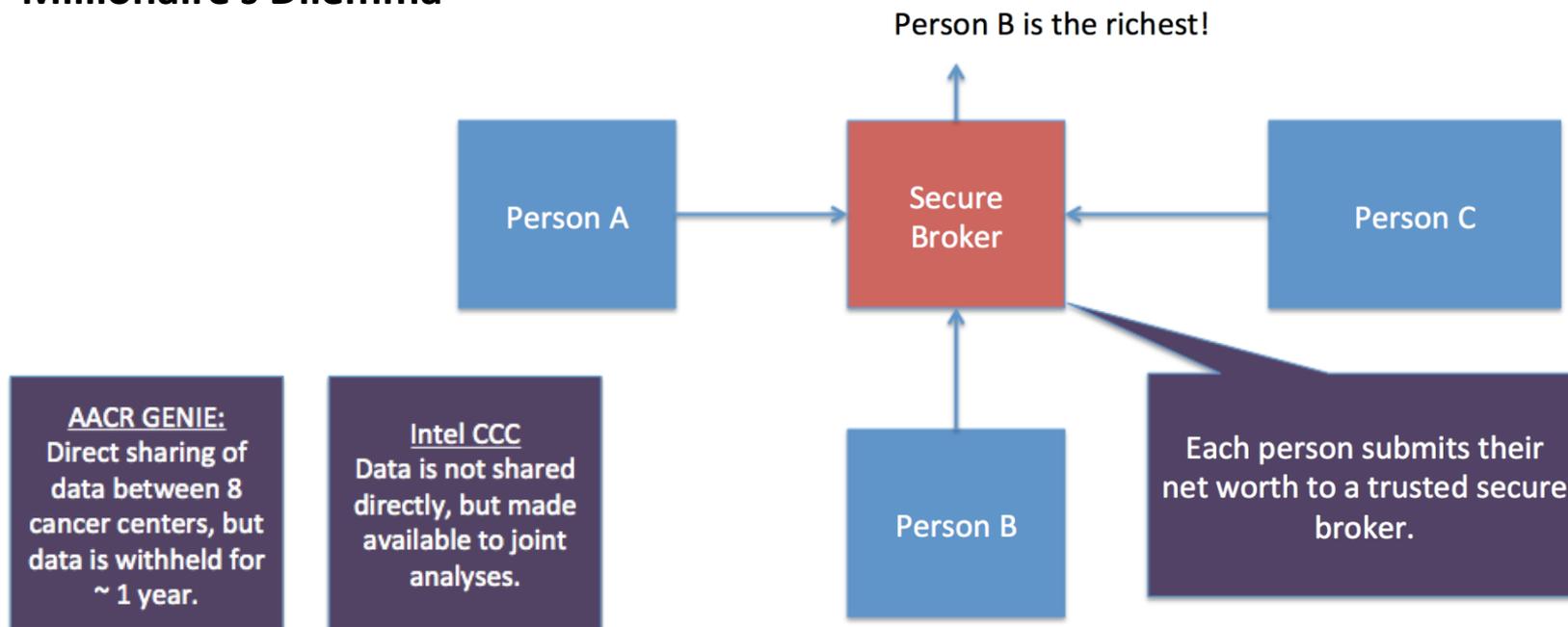


CCC is an open platform, enabling community best practice precision medicine analytics



# Unique Aspect of Intel CCC: Secure joint computation

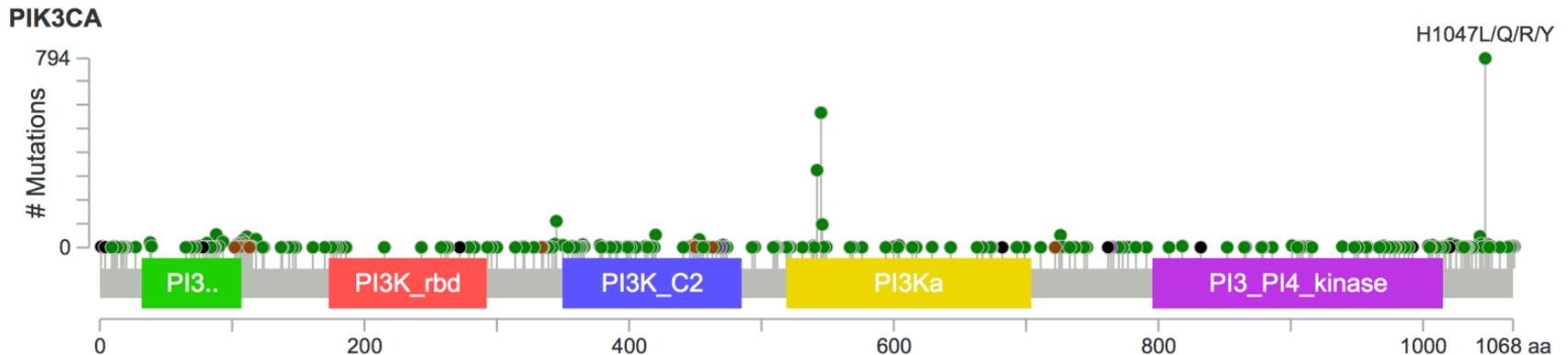
## Millionaire's Dilemma



# Motivation: Mutation Hotspots

- **Mutation Hotspot**

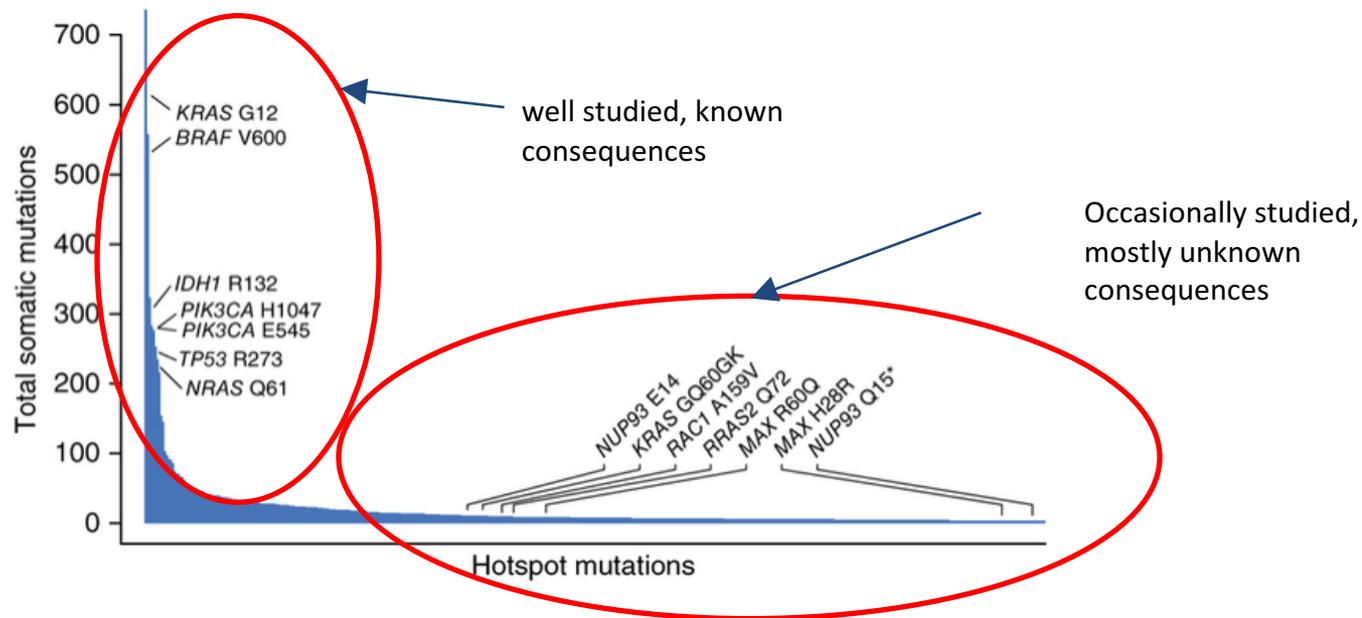
- a specific amino acid position that is mutated more frequently than expected by chance.
- likely indicative of oncogenic activity.



Example Mutation Hotspots in PIK3CA (Pan-Cancer)

# Motivation: Long Tail of Mutation Hotspots

*"85% of all hotspots identified were mutated in less than 5% of tumors"*



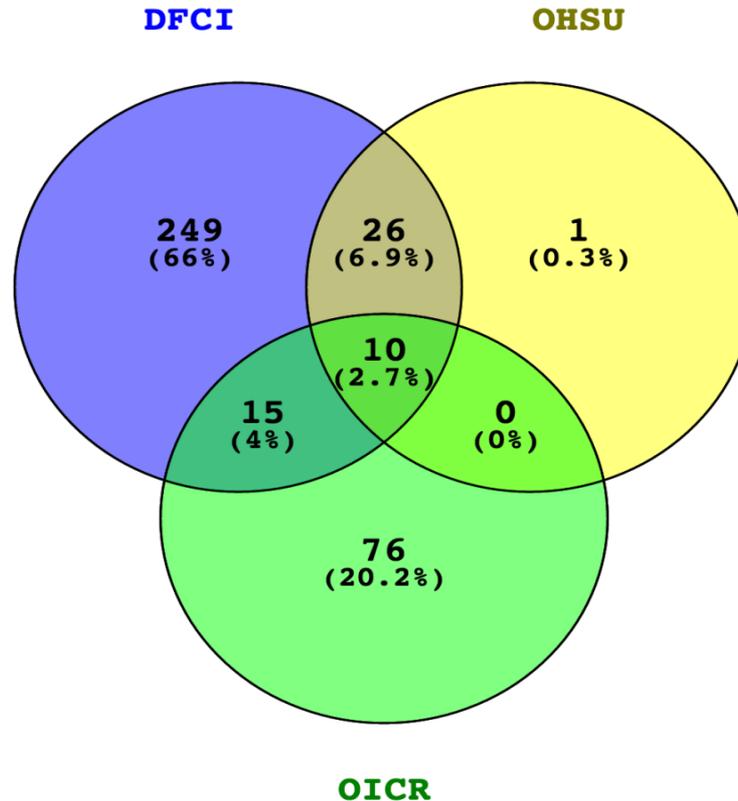
*"Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity". Chang et al. 2016*

Can be identified within a single cancer type or via pan-cancer analysis.

# Gene Overlap of Private Data Sets

## All Three Centers

AKT1  
 ERBB2  
 KRAS  
 MET  
 NF1  
 PIK3CA  
 PIK3R1  
 PTEN  
 RB1  
 TP53



## DFCI/OHSU

AKT2  
 AKT3  
 ALK  
 BRAF  
 CDK4  
 CDKN2A  
 DDR2  
 EGFR  
 FGFR1  
 FGFR3  
 GNA11  
 GNAQ  
 GNAS  
 HRAS  
 KDR  
 KIT  
 MAP2K1  
 NRAS  
 NTRK2  
 NTRK3  
 NOTCH1  
 RET  
 STK11  
 TSC1  
 TSC2  
 VHL

## DFCI/OICR

ARID1A  
 BRCA1  
 BRCA2  
 CBLB  
 CDH1  
 CDKN1B  
 EPHA5  
 GATA3  
 MAP2K4  
 MAP3K1  
 MYB  
 RUNX1  
 SF3B1  
 STAG1  
 TLR4



# Acknowledgements

---

## Knowledge Systems



## cBio @ MSKCC

Nikolaus Schultz  
JianJiong Gao  
Benjamin Gross

## The Hyve

Sjoerd van Hagen  
Pieter Lukasse  
Sander de Ridder  
Fedde Schaeffer  
Bernd van der Veen

## MatchMiner

Bruce Johnson  
Drew Memmott  
Geoffrey Shapiro  
George Demetri  
Khanh Do  
Steve DuBois  
Erica Woulf  
Adem Albayrak  
Susan Barry

## DFCI / BWH

Barrett Rollins  
Laura MacConaill  
Jane Song  
Matt Ducar  
Priyanka Shivdasani  
Lynette Sholl  
Neal Lindeman  
Stacy Gray  
Eliezer Van Allen

