Knowledge Systems @ DFCI

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

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

cBioCenter @ DFCI
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 Sharing Initiatives

Data acquisition

Data Mining Platform

Drive Genomic Discoveries

Clinical Decision Support Platform

Improve Patient Care

Data utilization
Highlighted projects

Data utilization
- cBioPortal
- MatchMiner
- Insight engine

Data acquisition
- cBioOne
- AACR GENIE
- Intel CCC
• 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
Enterprise genomics: PROFILE

- 15,927 patients sequenced
- Targeted DNaseq
- 447 genes / regions targeted
- CLIA-certified
cBioPortal, Insight Engine, MatchMiner

DATA UTILIZATION

Ethan Cerami, Jianjiang Gao, Ugar Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bilal Arman Aksy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, Erik Larsson, Yevgeny Antipin, Boris Reva, Arthur P. Goldberg, Chris Sander, and Nikolaus Schultz

DOI: 10.1158/1940-6207.CD-12-0095 Published May 2012

This article has a correction. Please see:
cBioPortal for Cancer Genomics

A: Genomic Data + Clinical Data

OncoPrint: Compact Visualization of Genomic Events Across Tumors
- Gene A
- Gene B
- Gene C

Survival Analysis
Network Analysis
Correlation plots

B: cBioPortal Statistics
- Tumor Samples ~20K
- Visitors Per Month ~25K
- Citations 1605

Graph showing:
- Unique users per year
- Help requests and citations per year

http://cbioportal.org
**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**

```r
# 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

Summary of Virtual Cohort Characteristics

Further Filtering via Interactive Plots

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

1,101 cases

Save Share Compare

B: Cohort Comparison

COHORT A
- Primary

COHORT B
- Metastasis

EXAMPLE COMPARISON

FREQUENTLY MUTATED GENES

GENOMIC INSTABILITY

SURVIVAL

AGE

Mutation Frequency

Mutations

Copy-number changes

Surviving

Count

Time

Age
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

R notebook

Python notebook

Python notebook

Python notebook

R Studio

Jupyter server or...

Feather: fast, interoperable binary data frame storage
Remove barriers for data science

1) Create new notebook in language of choice

   jupyter

2) Pull in data via programmatic interface

```python
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")
```
**MatchMiner**

**MatchMiner: Open Source Clinical Trial Matching Platform**

1. **Oncologist Mode**
   - Identify matching trials for specific patient.

2. **Clinical Trial Investigator Mode**
   - Identify matching patients for a specific clinical trial.

- **Genomic Alterations**
- **Clinical Data**
- **Structured Clinical Trial Information**
- **Clinical Trial Status Information**
Patient View – Trial Matches

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.

<table>
<thead>
<tr>
<th>Genomic match</th>
<th>Protocol #</th>
<th>Disease Center</th>
<th>Coordinating Center</th>
<th>DFCI Trial Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIK3CA p.H1047R Variant-Level Match Tier 4</td>
<td>TASELISIB+FULVESTRANT VS PLACEBO+FULVESTRANT FOR BRCA</td>
<td>DF/HCC Breast Cancer</td>
<td>Dana-Farber Cancer Institute</td>
<td>OPEN TO ACCRUAL</td>
</tr>
<tr>
<td>PIK3CA (4)</td>
<td>Protocol No: 15-153</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIK3CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wt PIK3CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invasive Breast Carcinoma</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adults</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Locally Advanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metastatic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recurrent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FULVESTRANT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLACEBO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TASELISIB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase III</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LETROZOLE +/- BYL719 OR BUPARLSIB FOR HER2-NEGATIVE BREAST CANCER</td>
<td>Protocol No: 14-202</td>
<td>DF/HCC Breast Cancer</td>
<td>Dana-Farber Cancer Institute</td>
<td>OPEN TO ACCRUAL</td>
</tr>
<tr>
<td>PIK3CA p.H1047R Gene-Level Match Tier 4</td>
<td>Principal Investigator: Krop, Ian E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIK3CA (4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wt PIK3CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adults</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Untreated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Localized</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Locally Advanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUPARLSIB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BYL719</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LETROZOLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIK3CA p.H1047R Gene-Level Match Tier 4</td>
<td>GDC-0032 + DOCEXTAXEL OR PACLITAXEL FOR BREAST</td>
<td>Protocol No: 13-123</td>
<td>Dana-Farber Cancer Institute</td>
<td>OPEN TO ACCRUAL</td>
</tr>
<tr>
<td>PIK3CA (4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wt PIK3CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adults</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Untreated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Localized</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Locally Advanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUPARLSIB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BYL719</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LETROZOLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Trial centric – Create filter

GENOMIC FILTER EDITOR

Edit genomic filter - ERBB2 Mut

GENOMICS  CLINICAL  GENERAL

Genomic attributes

Gene

ERBB2  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

Matches 96

Clinical 4985

Genomic 96

Accrual rate

2012 2013 2014 2015 2016 2017

Patient count

0 1 2 3 4 5 6 7 8 9

Simulated Data
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

Eligibility Criteria

Genomic Criteria

- **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 Criteria

- **age_numerical:** '>18'
- **oncotree_primary_diagnosis:** _SOLID_
MatchMiner usage stats

138 Trials, 675 register oncologists, 125 clinical trial investigators

### Daily Visits

![Graph showing daily visits per day from June 25 to July 24](image)

### Search Terms (July)

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR</td>
<td>6</td>
</tr>
<tr>
<td>16-494</td>
<td>4</td>
</tr>
<tr>
<td>16-711</td>
<td>3</td>
</tr>
<tr>
<td>demetri</td>
<td>3</td>
</tr>
<tr>
<td>(cdkn2a)</td>
<td>2</td>
</tr>
<tr>
<td>(cdkn2a,cdkn2b)</td>
<td>2</td>
</tr>
<tr>
<td>(cdkn2b)</td>
<td>2</td>
</tr>
<tr>
<td>(pik3ca)</td>
<td>2</td>
</tr>
<tr>
<td>(pik3ca,All solid tumors)</td>
<td>2</td>
</tr>
<tr>
<td>(pik3ca,All solid tumors,Open to Accrual)</td>
<td>2</td>
</tr>
</tbody>
</table>
Trial complexity of 138 curated trials

Number of genes specified:
- 1: 40.6%
- 2: 29.0%
- 3: 13.0%
- 4: 5.1%
- 5: 4.3%
- >=6: 8.0%

Number of tumor types specified:
- 1: 47.1%
- 2: 10.9%
- 3: 5.1%
- >=4: 6.5%
- All solid/liquid: 30.4%
cBioOne, Intel CCC, GENIE

DATA ACQUISITION
cBioPortal “as a service”
- for individual investigators
- for disease centers

1. Profile Data
2. Other Genomic Data
3. 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!

**DFCI**
**GRCC**
**JHU**
**MDA**
**MSK**
**NKI**
**UHN**
**VICC**

Clinical Sequencing

**Synapse**

- Data mapped to common ontology and harmonized
- Limited PHI removed
- Data governance, provenance, & versioning in a secure, HIPAA-compliant environment.

**cBioPortal**

- Institution-only access 6 months
- Consortium-only access 6 months

Genomic and clinical data linked

- Consortium/sponsor-only access 6 months to time of publication

**Quarterly data uploads**

**Clinical queries are posed based on registry content**

- Clinical data required to answer the question are manually abstracted
GENIE Landscape

B

Number of Samples

Cancer Type

Total 18966
MSK 7341
DFCI 6299
UHN 1296
JHU 1203
MDA 961
VICC 832
GRCC 529
NIKI 505
Landscape of Clinical Actionability

The graph shows the percentage of tumors that are actionable at different levels for various types of cancer:

- Level 3B: Highest level of evidence, with 70% of tumors actionable.
- Level 3A: Lower level of evidence, with 60% of tumors actionable.
- Level 2B: Further evidence required, with 50% of tumors actionable.
- Level 2A: Additional evidence needed, with 40% of tumors actionable.
- Level 1: Initial evidence, with 30% of tumors actionable.

The cancers listed include:
- Gastrointestinal Stromal Tumor
- Melanoma
- Glionoma
- Breast Cancer
- Bladder Cancer
- Endometrial Cancer
- Colorectal Cancer
- Gastroesophageal Cancer
- Renal Cell Cancer
- Skin Cancer
- Renal Cell Carcinoma
- Prostate Cancer
- Non-Hodgkin Lymphoma
- Mesothelioma

The graph indicates that certain types of cancer have a higher percentage of actionable tumors than others, with gastrointestinal stromal tumor and melanoma having the highest percentage of actionable tumors at 70% and 60%, respectively.
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

- ACLR GENIE: Direct sharing of data between 8 cancer centers, but data is withheld for ~1 year.
- Intel CCC: Data is not shared directly, but made available to joint analyses.
- Each person submits their net worth to a trusted secure broker.
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”

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

<table>
<thead>
<tr>
<th>Gene</th>
<th>DFCI/OHSU</th>
<th>DFCI/OICR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AKT3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRAF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDK4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDKN2A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDR2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGFR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGFR1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGFR3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNA11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNAQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KDR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP2K1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NRAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTRK2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTRK3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOTCH1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RET</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF3B1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TLK4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

OHSU

<table>
<thead>
<tr>
<th>Gene</th>
<th>DFCI/OHSU</th>
<th>DFCI/OICR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARID1A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBLB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDH1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDKN1B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPHA5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATA3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP2K4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP3K1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MYB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUNX1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF3B1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STAG1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TLR4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

OICR

<table>
<thead>
<tr>
<th>Gene</th>
<th>DFCI/OHSU</th>
<th>DFCI/OICR</th>
</tr>
</thead>
<tbody>
<tr>
<td>VHL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DFCI

<table>
<thead>
<tr>
<th>Gene</th>
<th>DFCI/OHSU</th>
<th>DFCI/OICR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTEN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RB1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP53</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

249 (66%)
26 (6.9%)
1 (0.3%)
15 (4%)
0 (0%)
76 (20.2%)
Full Gene Set Overlap

Present at all Centers
AKT1
ERBB2
KRAS
NF1
PIK3CA
PIK3R1
PTEN
RB1
TP53
Acknowledgements

Knowledge Systems

Nikolaus Schultz
JianJiong Gao
Benjamin Gross

MatchMiner

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

The Hyve

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

cBio @ MSKCC

DFCI / BWH

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