Building a sustainable, open source community

July 22, 2015
Outline of today’s presentation

- Nikolaus Schultz (Memorial Sloan Kettering Cancer Center, New York)
  - cBioPortal background
  - License change
  - New group structure
  - Roadmap (features & data)

- Ethan Cerami (Dana-Farber Cancer Institute, Boston)
  - DFCI experience
  - Roadmap ctd. (collaborative software development, communication)

- Stuart Watt (Princess Margaret Cancer Centre, Toronto)
  - Princess Margaret experience

- Kees van Bochove (The Hyve, The Netherlands)
  - The Hyve: Support for cBioPortal

- Nikolaus Schultz
  - Possible funding options
  - Discussion
cBioPortal: Data to Knowledge

Intuitive interface, quick response time, reduction of complexity
Gene alteration events per sample: Which genes are altered in each sample?

<table>
<thead>
<tr>
<th>Data type</th>
<th>Alteration event calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutations</td>
<td>Non-synonymous somatic mutations</td>
</tr>
<tr>
<td>Copy number changes</td>
<td>Homozygous deletion or amplification</td>
</tr>
<tr>
<td>Methylation</td>
<td>Epigenetic silencing</td>
</tr>
<tr>
<td>mRNA and/or DNA</td>
<td>Gene fusions</td>
</tr>
<tr>
<td>mRNA expression changes</td>
<td>Over- or under-expression</td>
</tr>
</tbody>
</table>

Alteration types and thresholds can be customized for each gene.
cBioPortal usage and interest

>5,000 unique users per week, doubling every year
cBioPortal use at MSKCC

Integration of MSKCC research sequencing, clinical sequencing (in-house and Foundation Medicine), with public data sets (TCGA, ICGC, manually curated data).

MSK clinical data is loaded from institutional databases

MSK-internal access control.
cBioPortal usage and interest

>5,000 unique users per week, doubling every year

cbioportal.org

Numerous academic installations of cBioPortal:
Dana-Farber, Princess Margaret, CHOP, Weill Cornell, Fred Hutchinson, UCSC, Columbia, NYU, NY Genome Center, British Columbia, University of Michigan, SickKids, Vanderbilt, Emory, UNC, University of Pittsburgh, CRUK, EMBL, Charite Berlin, institutions in Japan, China, ...

Interest by several people to modify or customize the code, and to contribute new features

Interest by pharmaceutical companies and others to use cBioPortal

- For internal data analysis (large pharma)
- In customer-facing applications (smaller service companies)
Switch to open source

cBioPortal source code is available via GitHub: https://github.com/cBioPortal/cbioportal

AGPL license v3 (Affero GPL):
   A GPL variant, main difference is that redistribution over a network triggers the copyleft requirements

Impact on cancer research, patient treatment, drug development through:
   ● More robust and flexible software
   ● Accelerated development of new features
   ● Wider user base, collaborative culture
Core cBioPortal Development group

Memorial Sloan Kettering Cancer Center
   Nikolaus Schultz, Chris Sander, Benjamin Gross, JJ Gao

Dana Farber Cancer Institute
   Ethan Cerami

Princess Margaret Cancer Centre
   Trevor Pugh, Stuart Watt

Re-uniting two cBioPortal founders
Coordination of architectural decisions, feature development, merges, etc.
cBioPortal roadmap overview

Analysis and Visualization Features
- Enhancements to the patient view
- Improved cohort analysis
- Query interface improvements

Software Architecture Improvements
- Performance and scalability improvements
- Customization options, logins
- Improved APIs

New Data Types and Pipelines
- Support for new genomic data types
- Data import pipelines
- Manual data curation

Community Engagement
- Users
- Local installations
- Developers
Planned features: Improved study view
Planned features: Improved study view
Planned features: Improved study view
Planned features: Improved study view
Planned features: Improved study view
Planned features: Improved patient view
Planned features: Enrichment analysis

Example: Mutations enriched in EGFR/KRAS-wildtype lung cancer samples (TCGA)
Planned features: Enrichment analysis

Example: Mutations enriched in EGFR/KRAS-wildtype lung cancer samples (TCGA)
Planned features: Cross-cancer expression

Example: Expression of SMARCA2 across TCGA tumor types
Planned features

**Patient view:**
- Variant annotation and tiering
- Clinical trial integration
- New simplified interface for clinicians
- Multi-sample view and timelines
- Support for xenografts
- Dynamic copy-number view
- Integration of pathology images

**Query interface improvements:**
- Building of case sets
- Study view improvements
  - Improved rendering speed
  - Selection of genes to start queries

**Cohort analysis:**
- Enrichment analysis
- Cross-cancer RNA expression
- Cross-cancer query within mixed studies
- Improved plotting of clinical vs genomic data
- True cross-cancer queries (merged OncoPrints)

**OncoPrint enhancements:**
- Improved rendering speed
- Better clinical annotation tracks
- More coloring options for mutation type
- Not sequenced vs no mutation
- mRNA expression heatmaps
- Merging of tracks

**Mutation-specific improvements:**
- Support for multiple gene isoforms
- Support for silent mutations
**Planned new data**

**New data types:**
- Richer support for gene fusions
- Support for normal mRNA expression
- Support for normal methylation levels
- Epigenetic silencing calls
- Mass-spec data
- microRNA expression
- Different types of z-scores

**New data:**
- ICGC data import pipeline
- More curated data sets from the literature
- Normal expression and methylation levels from TCGA
- Epigenetic silencing calls from TCGA
- Normal expression levels from GTEx
- RNA fusions for TCGA samples via PanCanFus
DFCI Experience

Ethan Cerami
Precision Medicine

Knowledge Systems @ DFCI

Data Mining Platform
-- Genomic Interpretation
-- Integrative analysis
-- Clinical Trial Design
-- Integrated w/ Clinical Data
-- Predictive Modeling

Clinical Decision Support Platform
-- Genomic Interpretation
-- Clinical Interpretation
-- Knowledge capture
-- Tumor Evolution / Resistance
-- Clinical Trial Eligibility

Integrate w/ global network of cancer center data sets (AACR GENIE)
A vision for personalized cancer medicine - Profile

Cohort cancer population

Every cancer patient is offered screening

Informed consent + Test order → Genomic test performed on tumor specimen

Analysis and interpretation of data → Report with Tiering and interpretation delivered to oncologist

Central database

Targeted Therapy Treatment

Clinical Trials Options

Tier 1
Matched to expected targeted therapy

Tier 2
Matched to unexpected targeted therapy

Diagnostic implications leading to change in therapy

Rare cancer-actionable alteration leading to change in therapy

More sensitive test leading to change in therapy

Matched to molecularly-defined clinical trial
Welcome to the cBioPortal @ DFCI

The cBioPortal is a web-based platform for exploring and analyzing large scale cancer genomic data sets. The site currently contains a subset of Profile data. Description of the Profile Data Freeze.

Additional data sets will be loaded over the next several months.

Select Cancer Study:
- All study selected
  - Pan-Cancer
  - Profile Paper - Data Freeze

Select Genomic Profiles:
- Mutations
- Copy Number

Select Patient/Case Set: Lung (863)

Enter Gene Set: Advanced: Oncor Query Language (OQL)

User-defined List

Enter HUGO Gene Symbols or Gene Aliases

Submit
DFCI Portal @ DFCI

- Getting data to researchers and clinicians with lowest barrier to access.
- Data transformation.
- Integrating with institutional authentication procedures: SAML
  - roles and users, fine grained authorization.
- Prioritized Features
  - cBioSkins
  - Pan-Cancer Queries
  - Build a Case
  - Integrating Tiering information
RoadMap Continued
Collaborative Software Development

● All code and issues are now on GitHub
● Improved Documentation
● Stabilizing all Unit Tests
  o automation via https://travis-ci.org/
● Finalizing
  o branch process
  o contributor agreement
  o review of pull requests
RoadMap Continued
Community outreach / communication

Users:
- **cBioPortal.org website**. Approximately 1,000 unique users visiting cBioPortal.org every day - this site will be used to announce new features, data, and the general availability of the software.

- **Google Group**. We have an active message board in Google Groups, which users can turn to for any questions about cBioPortal features and data on the public site.
  
  [https://groups.google.com/forum/#!forum/cbioportal](https://groups.google.com/forum/#!forum/cbioportal)

- **Twitter**. The cBioPortal Twitter handle @cbioportal currently has over 300 followers.

- **Mailing list**. We have a mailing list with over 300 subscribers.

Developers / installers:

- **cBioPortal GitHub page**. We will use this page to communicate with cBioPortal software developers.

- **cBioPortal Slack channel**. We have recently set up a cBioPortal Slack channel for core developers.
Princess Margaret Experience

Stuart Watt
Trevor Pugh
Princess Margaret Cancer Centre

- Largest cancer centre in Canada
- Fifth largest in the world
- 3,000 staff, 400,000 patient visits annually
- Central part of the Toronto biotech hub
  - University of Toronto teaching hospital
  - Close associations with SickKids, OICR
PMCC Cancer Genomics Programme

- Translational research programme
  - Aim for comprehensive molecular characterization as standard practice by 2018
  - Joint action by research and clinical groups
  - CAP/CLIA-certified diagnostics lab (AMDL)

- Connecting research and the clinic by:
  - Designing, developing, validating and implementing new tests
  - E.g., 555-gene sequencing panel for tumor profiling
cBioPortal at PMCC

• It’s the closest to where we want to be
  o Most functionality needed is already there
  o Integration of clinical and genomic data
  o Already used by some clinicians
  o Significant uptake from research groups

• Some challenges we face
  o Deployment processes
  o Data import
  o Access control
Current plans: cBioPortal at PMCC

- Improve verifiability of the platform
- Integration into standard clinical pipeline
- Integration into institutional systems
  - Clinical records
  - Authentication
  - Access control
The Hyve

Kees van Bochove
Summary

- cBioPortal now available under open source license
- Features and data facilitate cancer research, target discovery, and personalized patient treatment
- Used at many research institutions, sometimes already in “production” mode
- Coordinated software development across three groups
- Detailed roadmap of planned features, open to feedback about other features
- Commercial support by The Hyve
How can we fund this going forward?

- **Institutional funds?** Often used for custom internal work
- **Grants?** Less available than they used to be
- **Pre-competitive funding** through pharma?
Pre-competitive funding options

- Create independent foundation
- Join existing foundation (e.g., tranSMART)
- Create a smaller coalition of funders
- Funding via contracts for specific features

How can the contributions be used? Who allocates them? “Steering” committee vs “advisory” committee?
Small and nimble vs larger and …?
How valuable is access to curated data?
Proposed cBioPortal funding model

Small consortium of 4-5 pharma with seat on advisory committee (plus the three of us) @ $100k / year

- Advisory role
- Access to all data, ability to nominate data sets for curation
- Money is allocated for development within our groups
- Lower data-only tier available @ $30k / year

2-3 year commitment, determine longer term sustainability plan during year 1
Discussion & Survey

We need your feedback on our ideas
What are your ideas?
What else is needed to make this work?

Please complete our survey at:
Thank you

Carmen Nitsche
Richard Holland

Julia Rudolph, MSKCC

cBioPortal developers

present:
JianJiong Gao
Benjamin Gross
Yichao Sun
Hongxin Zhang
Fred Criscuolo
Dong Li
Adam Abeshouse
Ritika Kundra
Annice Chen
Onur Sumer

past:
Caitlin Byrne
Gideon Dresdner
Arthur Goldberg
Michael Heuer
Anders Jacobsen
Erik Larsson