TranSMART Roadmap

TMF TRANSMART WORKSHOP – BERLIN, AUGUST 5, 2016

Kees van Bochove, CEO & Founder, The Hyve

Architecture Working Group Chair, TranSMART Foundation

Outline

- Introduction
- The future of TranSMART, as seen from 4 TranSMART Projects
  - CTMM TraIT
  - IMI RADAR-CNS
  - IMI Translocation
  - TranSMART 17.1
- Conclusion / Discussion

1.

INTRODUCTION

Open Source

- Source code openly accessible and reusable **for everyone**
- Enables **pre-competitive collaboration**: both academics and industry can use and enhance it
- **Transparency**: verification (scientific as well as IT security) can be done by anyone, no 'black box'

redhat
android
mongoDB
Neo4j
Firefox
OpenOffice.org
The Open Source Office Suite
ubuntu
openstack
docker
cassandra
xbmc
MEDIA CENTER
Apache MESOS
Spark
VLC
media player

The Hyve

- Professional support for **open source software** for bioinformatics and translational research software, such as tranSMART, cBioPortal, i2b2, Galaxy, ADAM and OHDSI

Core values

Share

Reuse

Specialize

Office Locations

Utrecht, Netherlands
Cambridge, MA, United States

Services

Software development
Data science services
Consultancy
Hosting / SLAs

Mission

Enable pre-competitive collaboration
in life science R&D
by leveraging
open source software

Fast-growing

Started in 2012
35+ people by now

Interdisciplinary team

software engineers, data scientists, project managers & staff; expertise in bioinformatics, medical informatics, software engineering, biostatistics etc.

New offices at the Arthur van Schendellaan in Utrecht

Open Source in Precision Medicine

Clinical / Healthcare:

OpenClinica
Open Source for Clinical Research

SMART

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Imaging:

XNAT

Biobanking:

OBiBa molgenis

Study design:

Phenotype database

Datawarehousing:

OMOP

i2b2

tranSMART
v1.2

Workflow / NGS:

Galaxy

Arvados

Data visualisation:

CALEYDO

NDEx

cBioPortal
for Cancer Genomics

Scientific compute:

jupyter

OpenCPU

R

Spark

TranSMART Platform: Scientific Function

CLINICAL

GENETICS

SENSORS

IMAGING

DATA

BIOLOGY

MEDICINE

UNDER STANDING

TranSMART Open Source History

- February 2012: J&J releases tranSMART as open source on GitHub under GPL v3
- December 2012: CTMM TraIT project decides to use tranSMART as core infrastructure component
- January 2013: IMI eTRIKS starts, uses tranSMART as core infrastructure component
- February 2013: kickoff of tranSMART Foundation, U. Michigan publishes PostgreSQL port
- March 2014: IMI EMIF kickoff, tranSMART is used as data integration component

70

Amsterdam, June 2013: tranSMART Workshop

Attendees from 10 Pharma companies, 11 University Medical Centers and 12 IT companies

http://lanyrd.com/2013/transmart

130

Ann Arbor, Michigan, October 2014: Annual Meeting

http://lanyrd.com/2014/transmart

TranSMART wins all the prizes: Best Show Award, Best Practices Award, Best Poster Award

Bio IT World, Boston, April 2015

http://bit.ly/1R2N6uz

160

Amsterdam, October 2015: Annual Meeting

http://lanyrd.com/2015/transmart-foundation-annual-meeting/

TranSMART Annual Meeting 2016

mark your calendar!

October 25 – 27

UC San Diego – La Jolla, California

The future of tranSMART, as seen from:

- **CTMM TraIT**

  Translational Research data exchange infrastructure

  for all 8 University Medical Centers in The Netherlands

- **IMI RADAR-CNS**

  Using tranSMART to store mHealth and wearables data

- **IMI Translocation**

  Building a new user interface for ‘TranSMART 2.0’

- **TranSMART 17.1**

  Architectural improvements in the tranSMART backend

2.

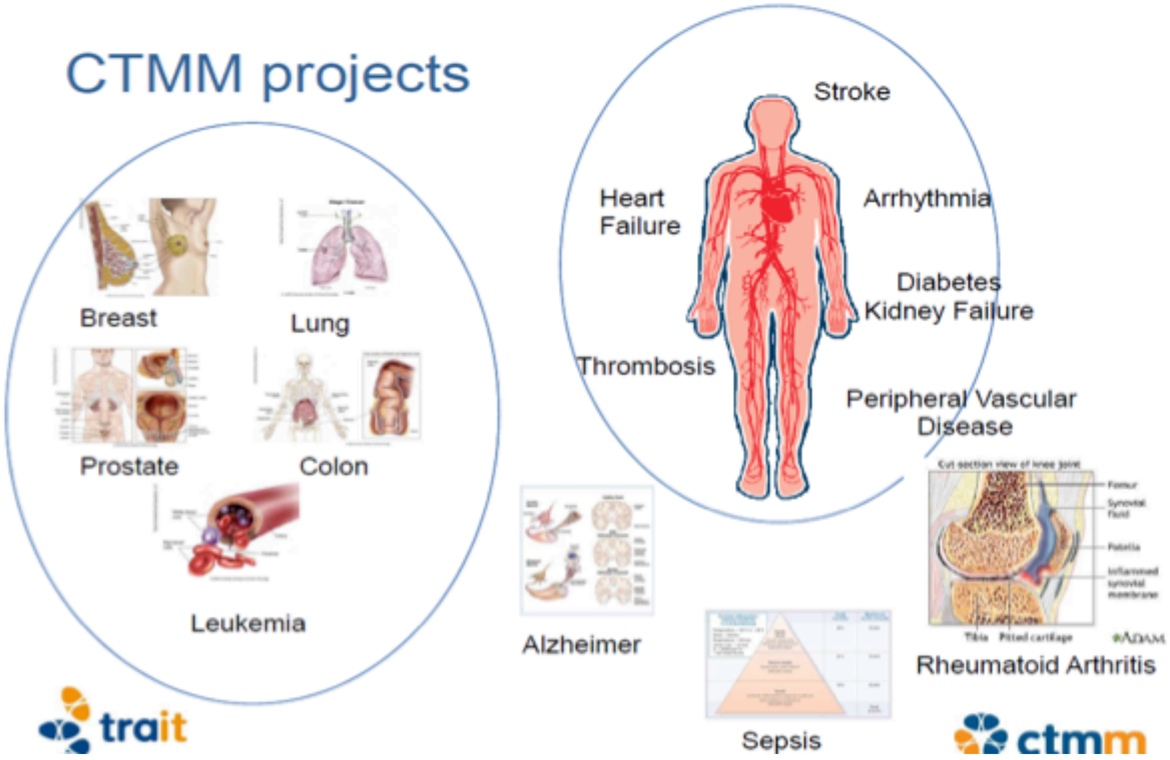
trait

a ctmm project

PI: Gerrit Meijer, Netherlands Cancer Institute

Center for Translational Molecular Medicine (CTMM)

- Public-private consortium
- Dedicated to the development of Molecular Diagnostics and Molecular Imaging technologies
- Focusing on the translational aspects of molecular medicine.
- 120 partners
  - universities, academic medical centers, medical technology enterprises and chemical and pharmaceutical companies.
- Budget 300 M€
- 22 projects / research consortia
- TraIT is the Translational Research IT project supporting these projects with a joint IT infrastructure



TraIT Consortium

netherlands eScience center

ctmm

University Medical Center Utrecht

PSI Parelsnoer Instituut

VU university medical center

Hartstichting

KWF KANKER BESTRIJDING

Erasmus MC University Medical Center Rotterdam

PHILIPS

nbic

umcg

amc

MAASTRO

VU VRIJE UNIVERSITEIT AMSTERDAM

UMC St Radboud

ICIN Netherlands Heart Institute

GENALICE TECHNOLOGY FOR PEOPLE & SCIENCE

thehyve

SURF SARA

CSC

CORDYS The Enterprise Cloud Platform

Growing TraIT project team

Vancis

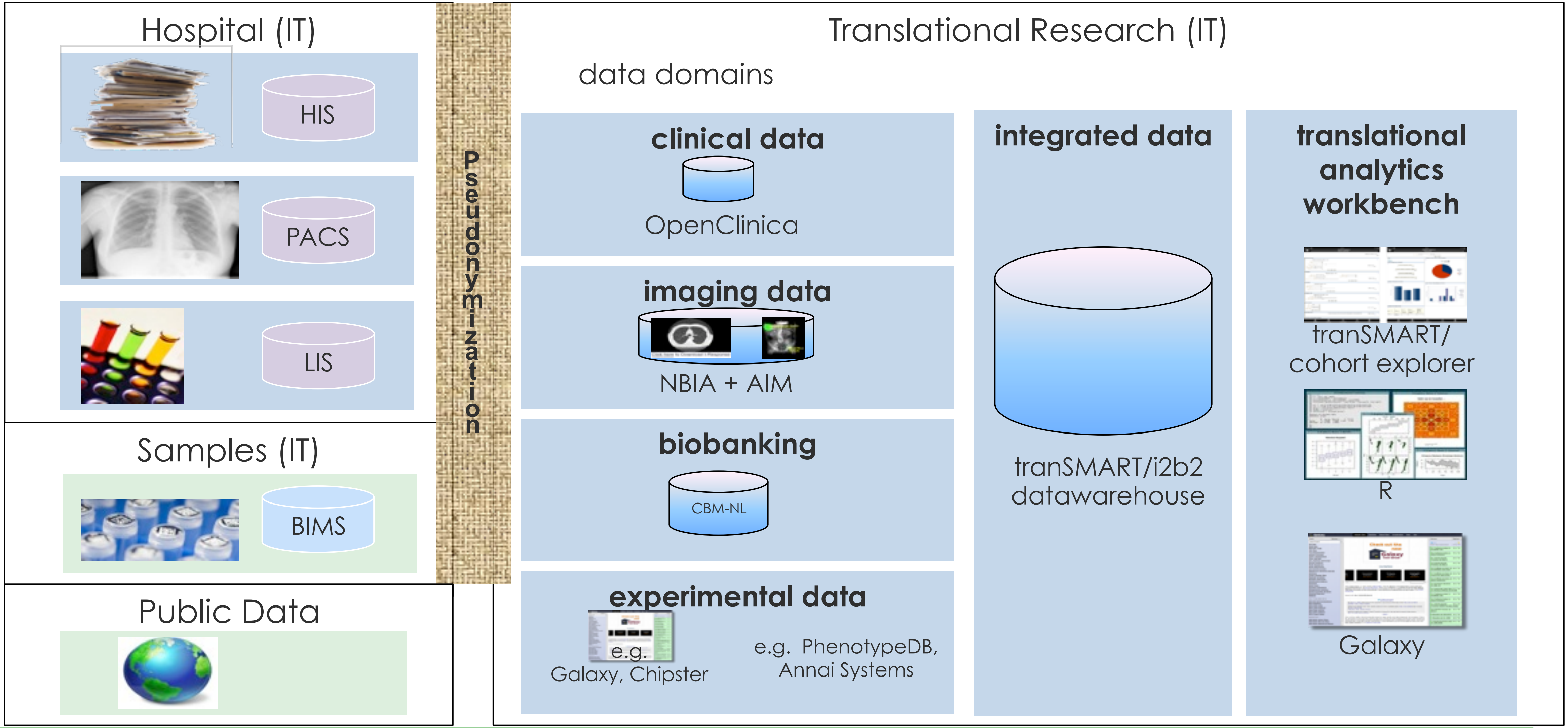
Maastricht UMC+ azM | Maastricht University

KEOSYS Medical Imaging

Roche Diagnostics

OpenClinica® Open Source for Clinical Research

TraIT data workflow



Hospital (IT)

HIS

PACS

LIS

Samples (IT)

BIMS

Public Data

Pseudonymization

Translational Research (IT)

data domains

clinical data

OpenClinica

imaging data

NBIA + AIM

biobanking

CBM-NL

experimental data

e.g.
Galaxy, Chipster

e.g. PhenotypeDB,
Annai Systems

integrated data

tranSMART/i2b2
datawarehouse

translational analytics workbench

tranSMART/
cohort explorer

R

Galaxy

cBioPortal for Cancer Genomics

current community a.o.

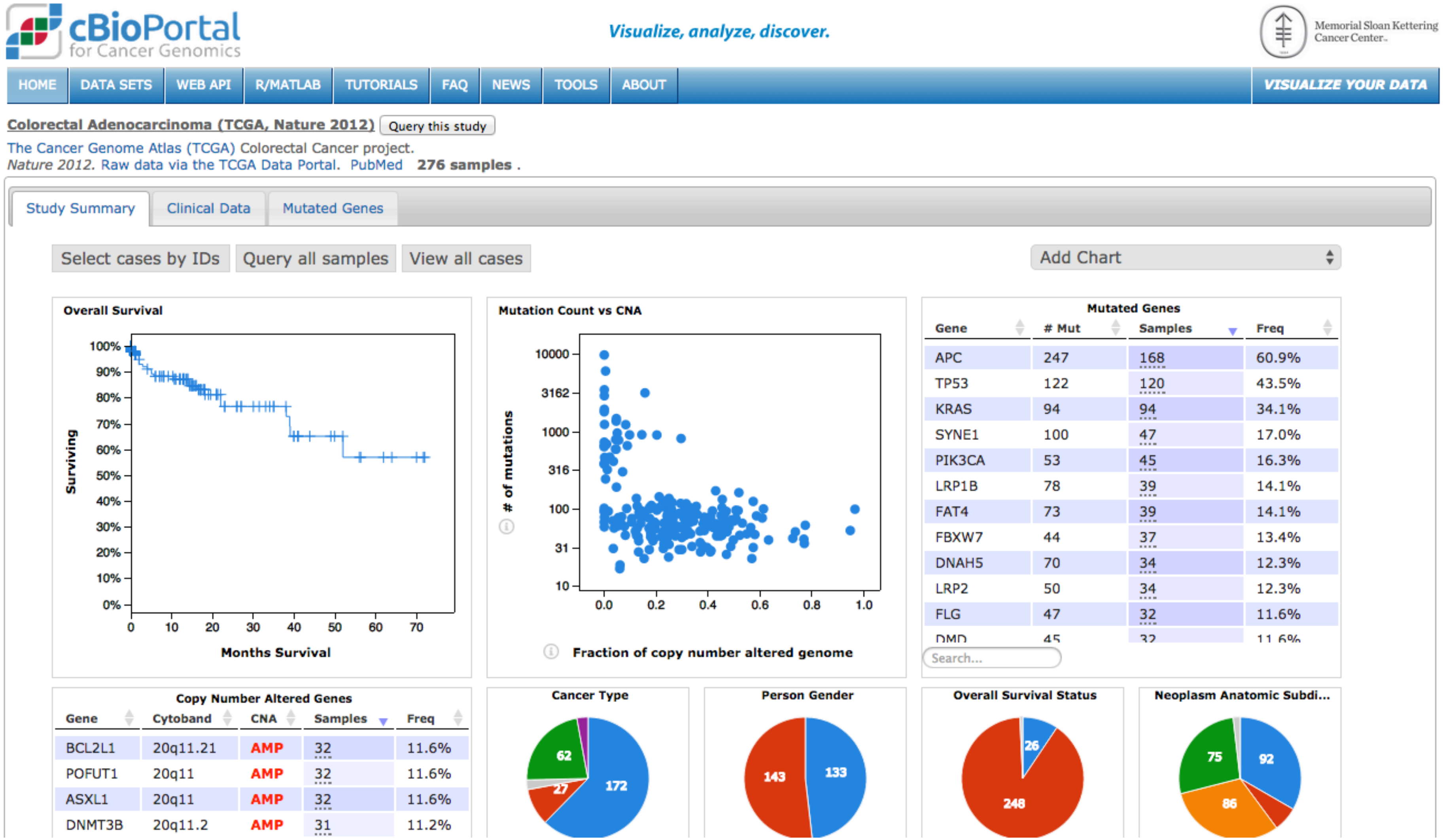
Memorial Sloan Kettering Cancer Center

DANA-FARBER CANCER INSTITUTE

The Princess Margaret Cancer Centre
University Health Network

NETHERLANDS CANCER INSTITUTE
ANTONI VAN LEEUWENHOEK

the hyve



Colorectal Adenocarcinoma (TCGA, Nature 2012)

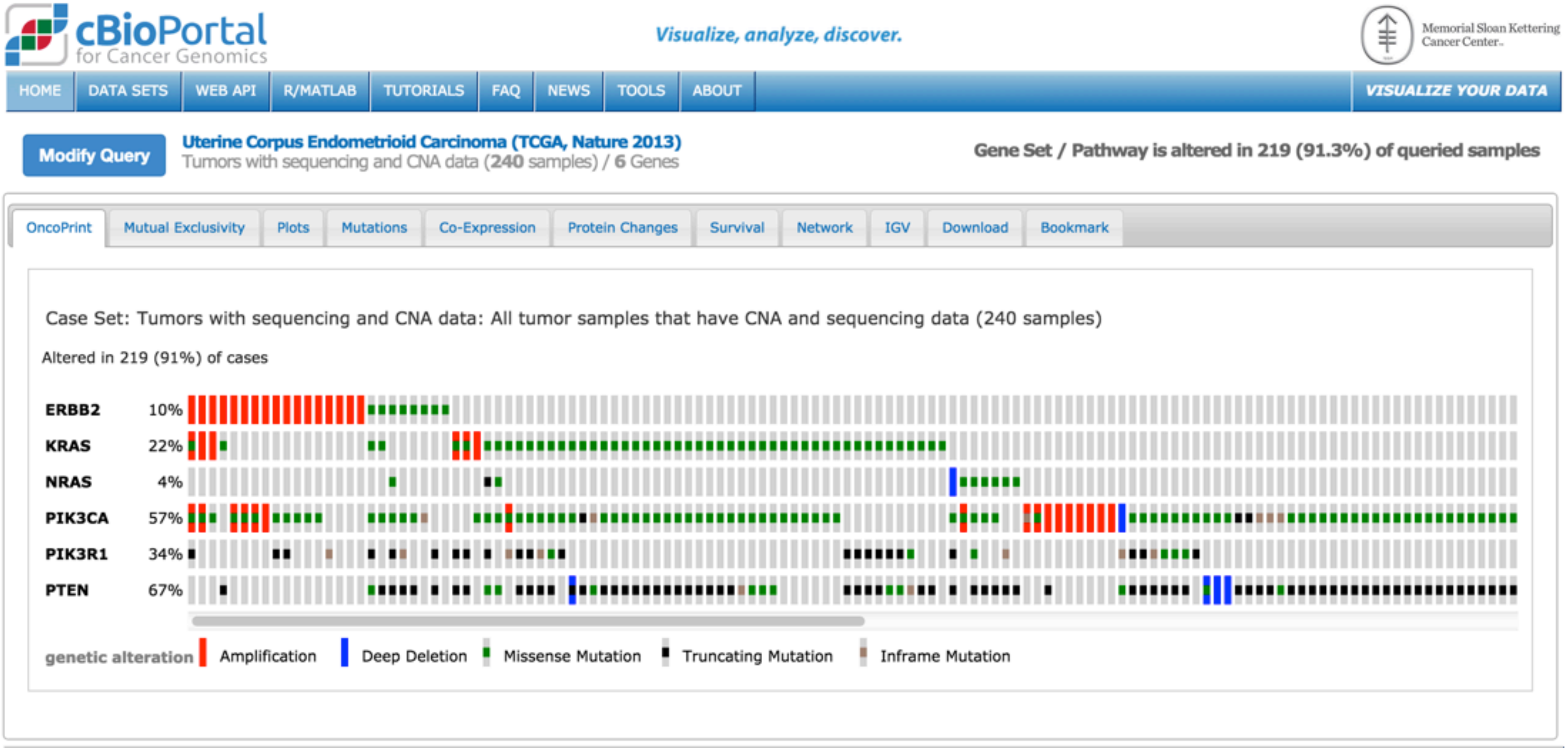
The Cancer Genome Atlas (TCGA) Colorectal Cancer project.
Nature 2012. Raw data via the TCGA Data Portal. PubMed **276 samples** .

Mutated Genes

Gene	# Mut	Samples	Freq
APC	247	168	60.9%
TP53	122	120	43.5%
KRAS	94	94	34.1%
SYNE1	100	47	17.0%
PIK3CA	53	45	16.3%
LRP1B	78	39	14.1%
FAT4	73	39	14.1%
FBXW7	44	37	13.4%
DNAH5	70	34	12.3%
LRP2	50	34	12.3%
FLG	47	32	11.6%
DMD	45	32	11.6%

Copy Number Altered Genes

Gene	Cytoband	CNA	Samples	Freq
BCL2L1	20q11.21	AMP	32	11.6%
POFUT1	20q11	AMP	32	11.6%
ASXL1	20q11	AMP	32	11.6%
DNMT3B	20q11.2	AMP	31	11.2%



Uterine Corpus Endometrioid Carcinoma (TCGA, Nature 2013)
Tumors with sequencing and CNA data (240 samples) / 6 Genes

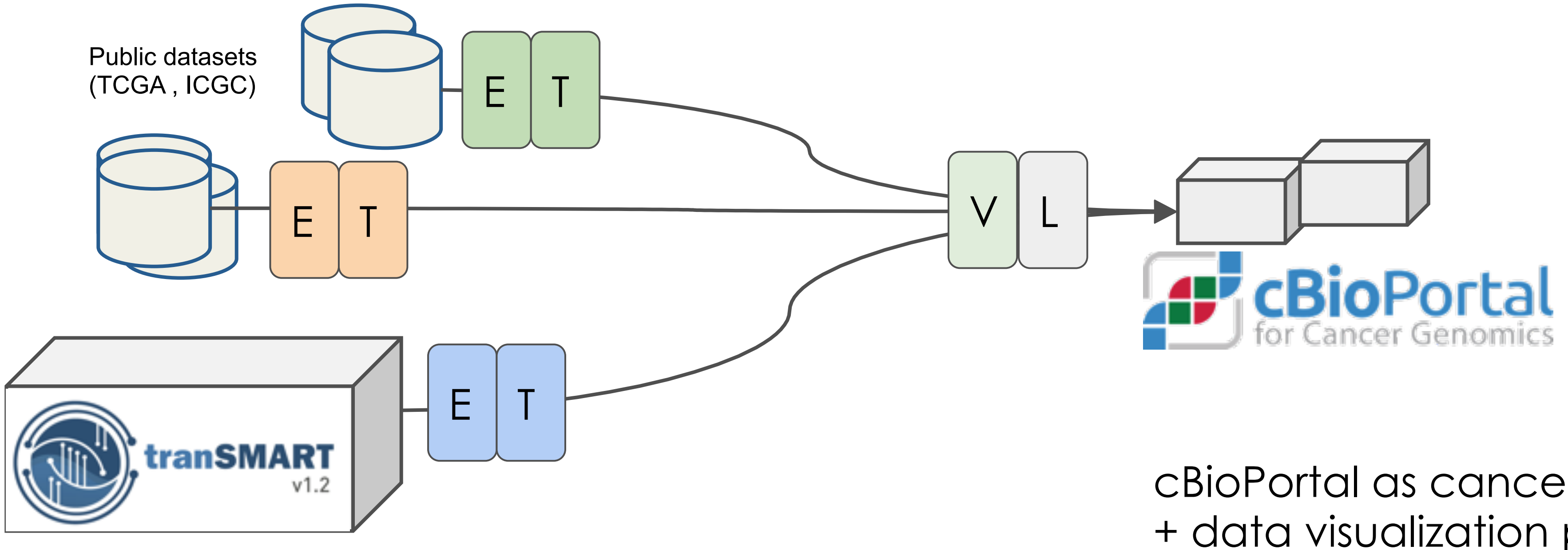
Gene Set / Pathway is altered in 219 (91.3%) of queried samples

Case Set: Tumors with sequencing and CNA data: All tumor samples that have CNA and sequencing data (240 samples)

Altered in 219 (91%) of cases

Visualization of events across genes and data types

cBioPortal for Cancer Genomics integration



TranSMART as enterprise datawarehouse

cBioPortal as cancer studies data mart
+ data visualization portal

3.

imi
Innovative Medicines Initiative

RADAR-CNS

PI: Matthew Hotopf, King's College London

Challenges in managing chronic diseases

- Health assessments via physician visits are *time limited* and *subjective*, often not representative for everyday life of the patient
- The disease state can change a lot in between visits, and important events are not visible to the physician

But… it's 2016: it's now possible to **objectively**, **remotely**, **and continuously** measure aspects of patient **physiology**, **behavior and symptoms**

RADAR-CNS: Focus areas

from diagnose & treat → predict & pre-empt

- Epilepsy
  - Monitoring and predicting epileptic seizures
- Multiple Sclerosis
  - Monitoring exacerbations and disease state
- Depression
  - Monitoring for possible relapses, plan timely interventions
  - Predict bipolar state transitions

Continuous Patient Assessment

Physiology

ECG
HR/HRV
Respiration
Skin temp
Activity/Sleep
O2 sat

Behavior

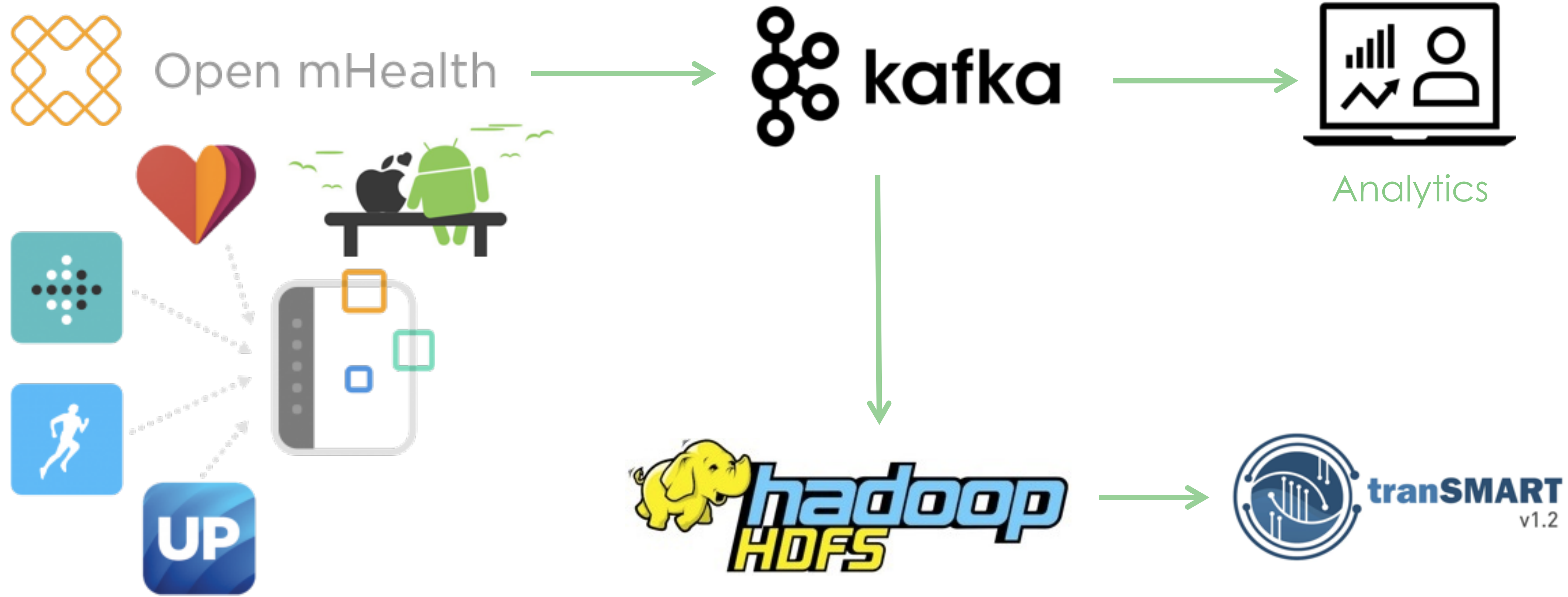
GPS
Talk patterns
Text patterns
Activity/Sleep

Symptoms

IVR

Smartphone
Symptom
assessment

Preliminary Technology Stack



Open mHealth

kafka

Analytics

hadoop
HDFS

tranSMART
v1.2

4.

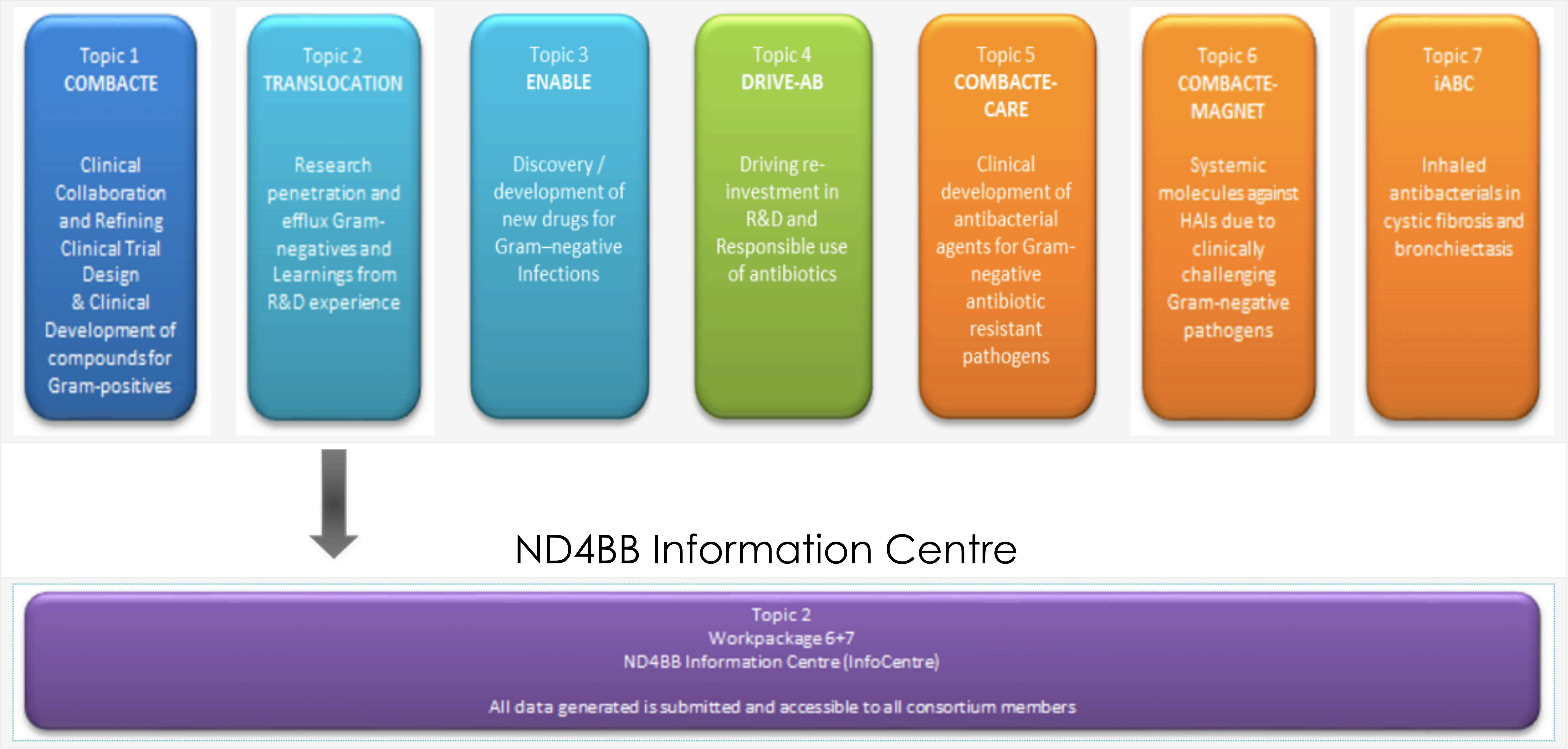
imi

Innovative Medicines Initiative

ND4BB TRANSLOCATION

PI: Mathias Winterhalter, Jacobs University Bremen

Translocation is part of New Drugs for Bad Bugs (ND4BB) programme



Topic 1
COMBACTE
Clinical Collaboration and Refining Clinical Trial Design & Clinical Development of compounds for Gram-positives

Topic 2
TRANSLOCATION
Research penetration and efflux Gram-negatives and Learnings from R&D experience

Topic 3
ENABLE
Discovery / development of new drugs for Gram–negative Infections

Topic 4
DRIVE-AB
Driving re-investment in R&D and Responsible use of antibiotics

Topic 5
COMBACTE-CARE
Clinical development of antibacterial agents for Gram-negative antibiotic resistant pathogens

Topic 6
COMBACTE-MAGNET
Systemic molecules against HAIs due to clinically challenging Gram-negative pathogens

Topic 7
iABC
Inhaled antibacterials in cystic fibrosis and bronchiectasis

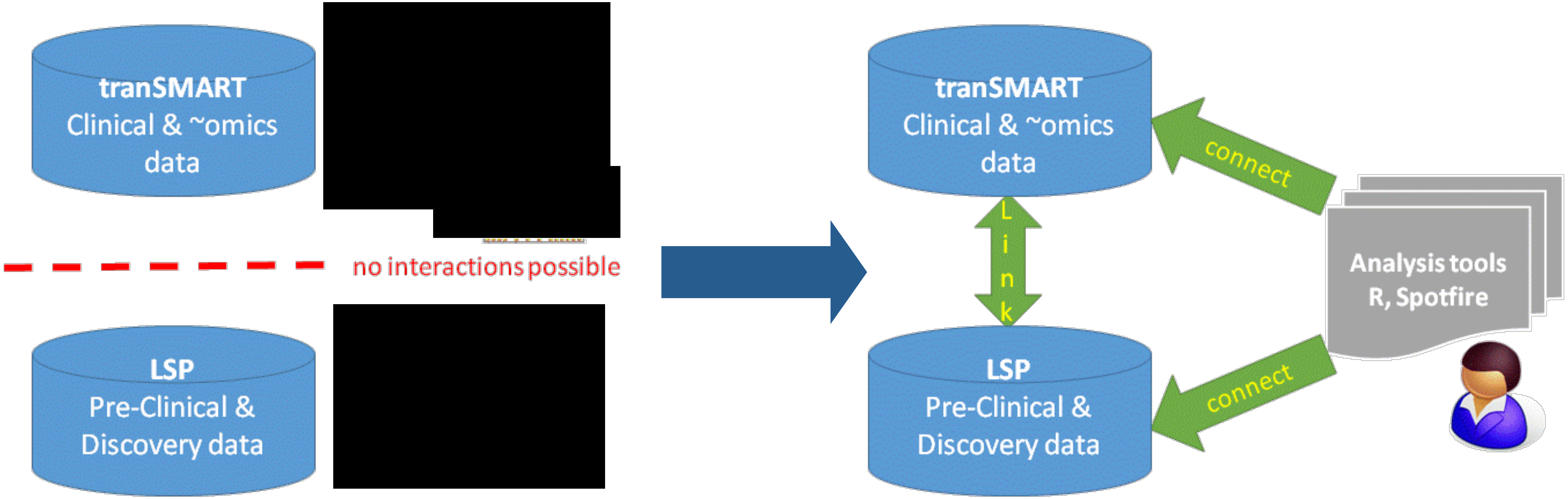
ND4BB Information Centre

Topic 2
Workpackage 6+7
ND4BB Information Centre (InfoCentre)

All data generated is submitted and accessible to all consortium members

ND4BB Information Centre

Connecting clinical and pre-clinical data around compounds



tranSMART
Clinical & ~omics
data

[redacted]

no interactions possible

LSP
Pre-Clinical &
Discovery data

[redacted]

[redacted]

tranSMART
Clinical & ~omics
data

Link

LSP
Pre-Clinical &
Discovery data

connect

connect

Analysis tools
R, Spotfire

Integration via R

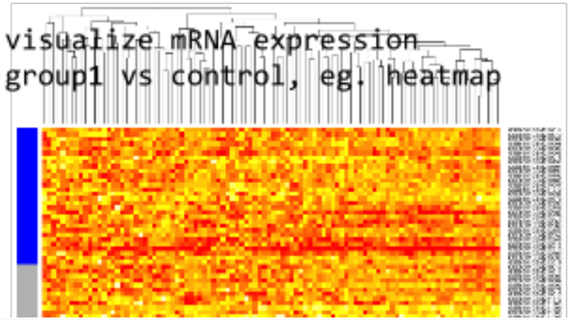
```
connectToLSP('https://LSP.server')
comps <- getCompounds()
```

Table with compound info

```
conc <- getCompoundConcepts(comps$id[7])
```

Concepts in tranSMART linked to compound

```
group1 <- getSubjectsByConceptLinks(conc[1])
expression <- getHighDimdata(concept.in.same.study)
```



visualize mRNA expression group1 vs control, eg. heatmap

```
connectToTransmart('https://tranSMART.server')
studies <- getStudies()
```

Table with study info

```
tree <- getTreeForStudy(studies$id[2])
```

StudyTree
- endpoint
- treatment
  - group1
  - group2

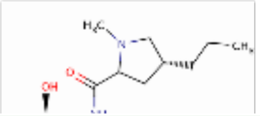
```
tree$endpoint$getObservations()
```

Subject observations

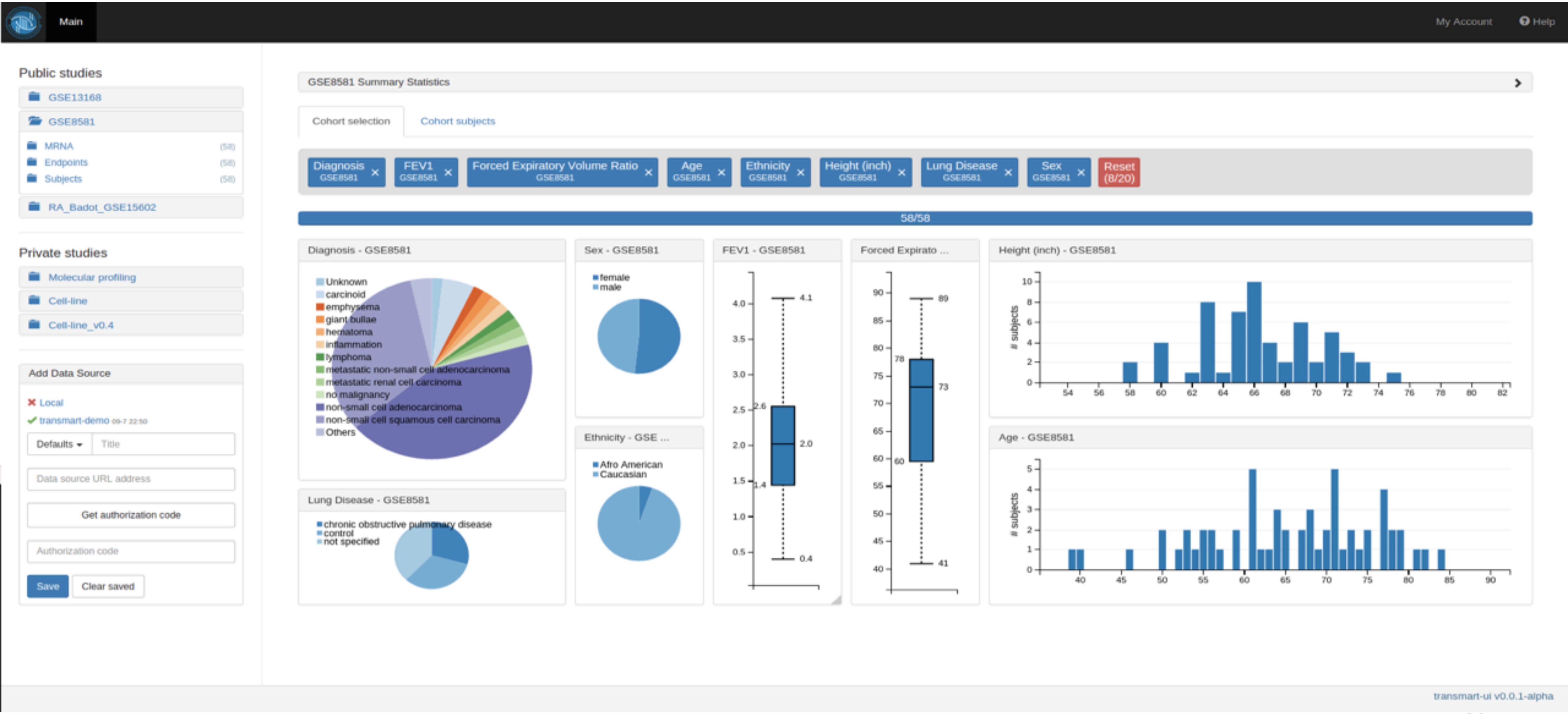
```
c1 <- tree$...$group2$fetchLSPCompoundInformation()
```

Compound-info

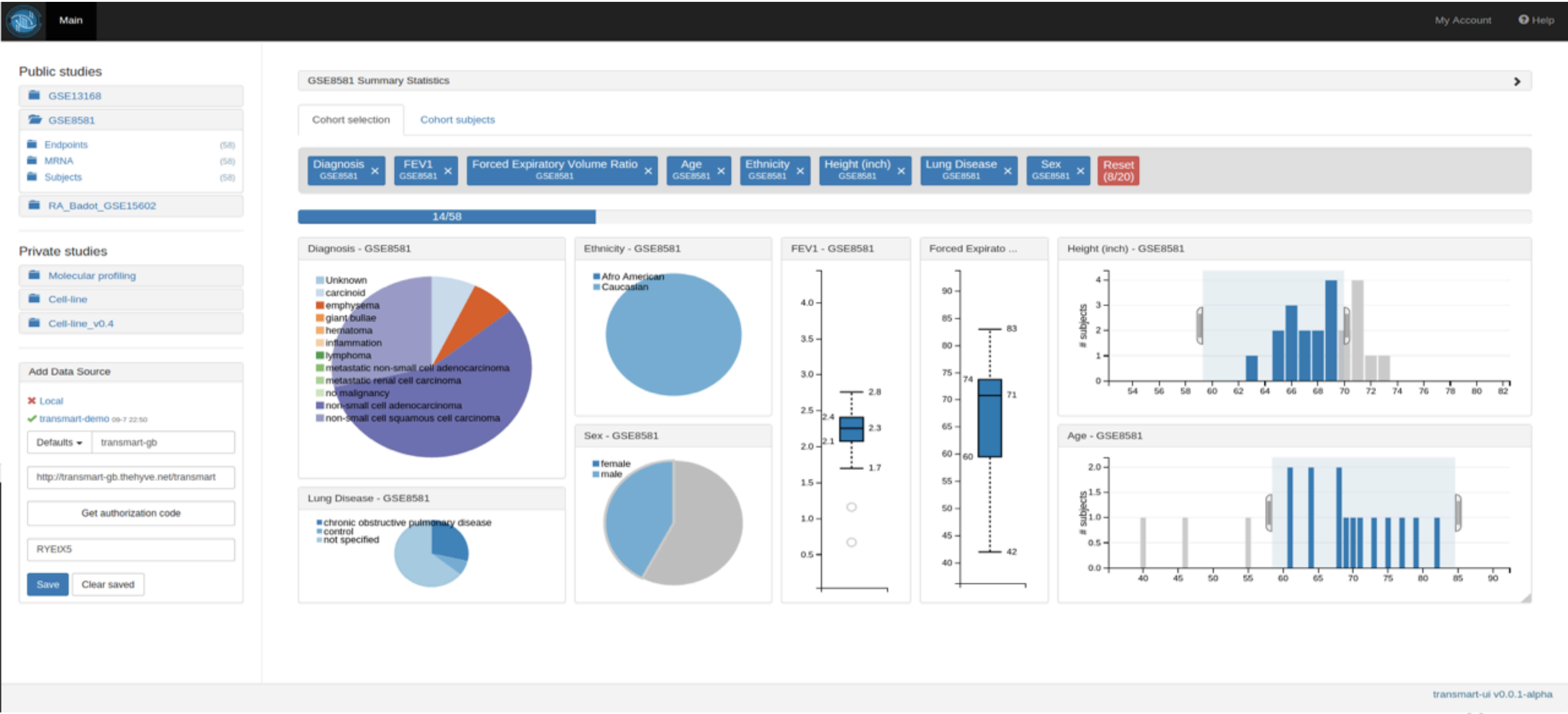
```
viewMolecule(c1)
```



New TranSMART User Interface



New TranSMART User Interface



Defaults: transmart-gb

http://transmart-gb.thehyve.net/transmart

RYEtX5

Pfizer Roche SANOFI abbvie

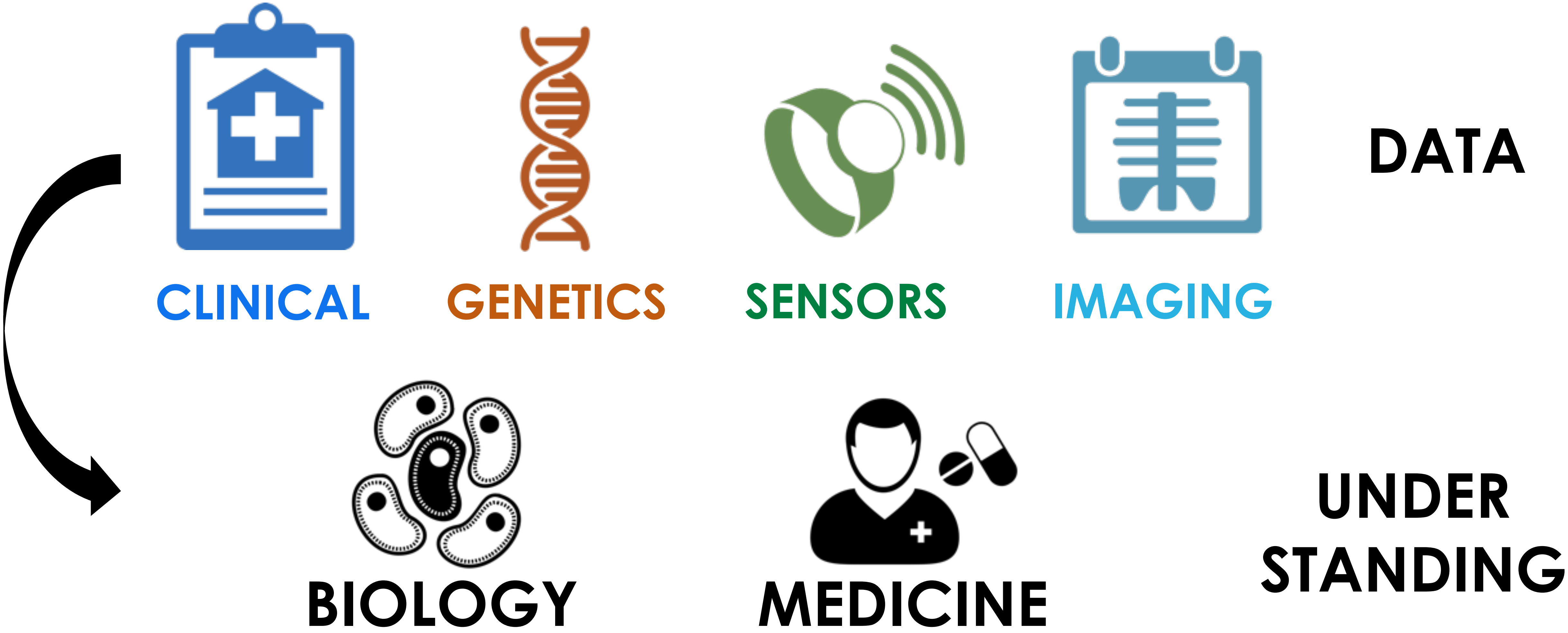
5.

tranSMART FOUNDATION

TRANSMART 17.1

TranSMART Pro Alliance

TranSMART Platform: Scientific Function



CLINICAL

GENETICS

SENSORS

IMAGING

DATA

BIOLOGY

MEDICINE

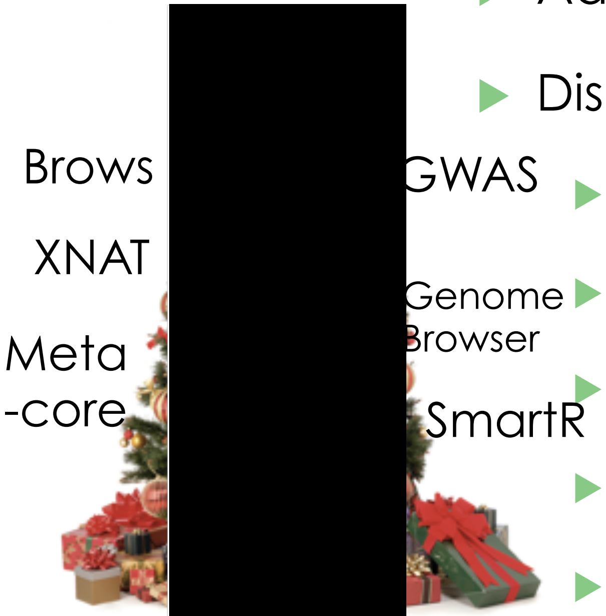
UNDER STANDING

Analytics – what makes it hard?

- **Clinical Genomics:** how can I accurately find and reliably measure clinically relevant genomic variations using high-throughput techniques? (genetic testing, population study)
- **Cancer Genomics:** how can I reliably sequence the cancer genome and compare with germline to identify cancer subspecies? (diagnostics / precision medicine)
- **Molecular Biology**: what is the relation between the properties of the DNA and transcription and translation processes in the cell? (gene-centric, epigenetics, multi-omics)
- **Evolutionary Genetics:** how can we quantify and explain inter-species and intra-species genetic variations?
- **Bioinformatics**: how can I improve signal to noise ratio in sequencing pipelines and accurately call genomic variants?
- **Computer Science:** which compute architecture fits data analysis & visualization needs?

TranSMART Product Management

- TranSMART approach so far has been: integrate all possible functionalities into each product release
  - Advantage: was needed to unify community, one product
  - Disadvantages
    - ‘Christmas Tree’ of product features: no overarching product vision
    - Integration of features becomes harder and more costly with every new release
    - Ironically, this strategy leads to lots of back- and forth porting of features in practice
    - Hard to enforce quality standards, leads to spaghetti code
    - Extremely hard to coordinate joint releases, different levels of feature maturity

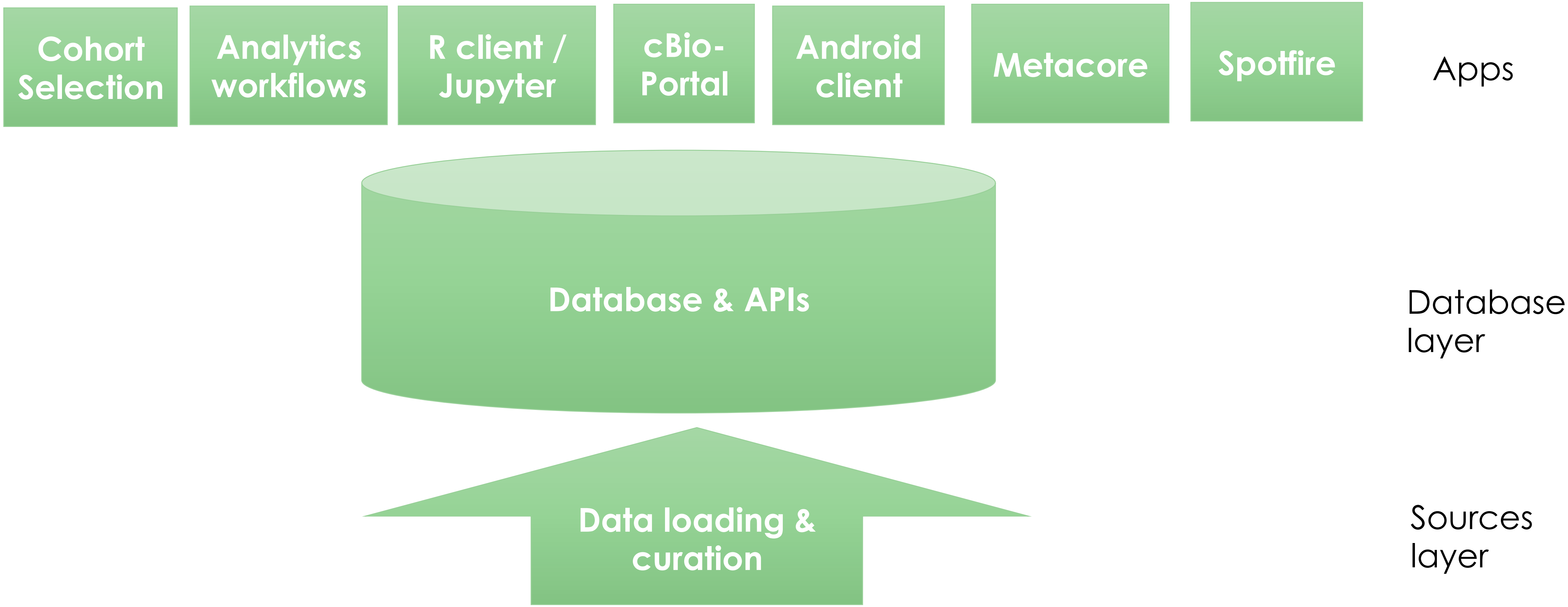


[redacted]

Improving TranSMART Product Management

- Create a tranSMART Core
  - Stable, robust, well performing, interoperable
  - meant as a base for all tranSMART instances
  - **high quality** standards, predictable release schedule
- Maintain a ‘marketplace’ of tranSMART Apps
  - Different use cases, target users, technologies
  - Different **maturity levels** (!!) – room for parallel evolution of features
  - Built on the core, leveraging the core APIs

TranSMART layers



Cohort Selection

Analytics workflows

R client / Jupyter

cBio-Portal

Android client

Metacore

Spotfire

Apps

Database & APIs

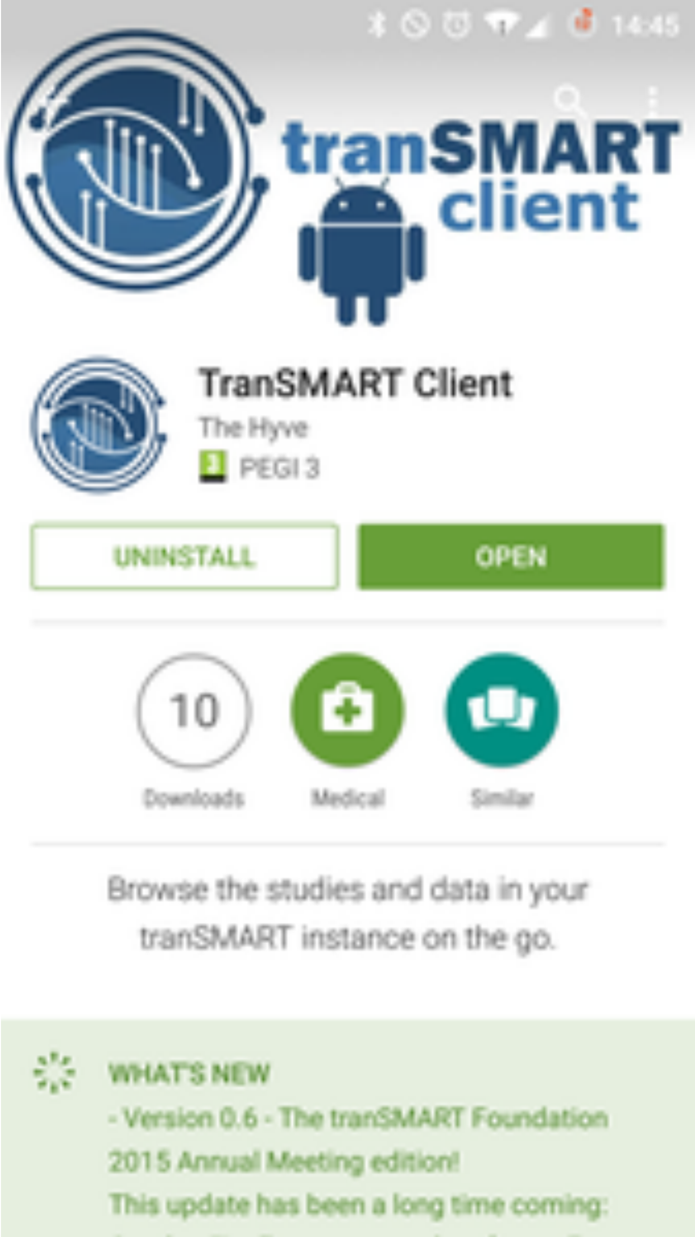
Database layer

Data loading & curation

Sources layer

Android Client

bit.ly/testtransmart



TranSMART Client

The Hyve

PEGI 3

UNINSTALL

OPEN

10 Downloads

Medical

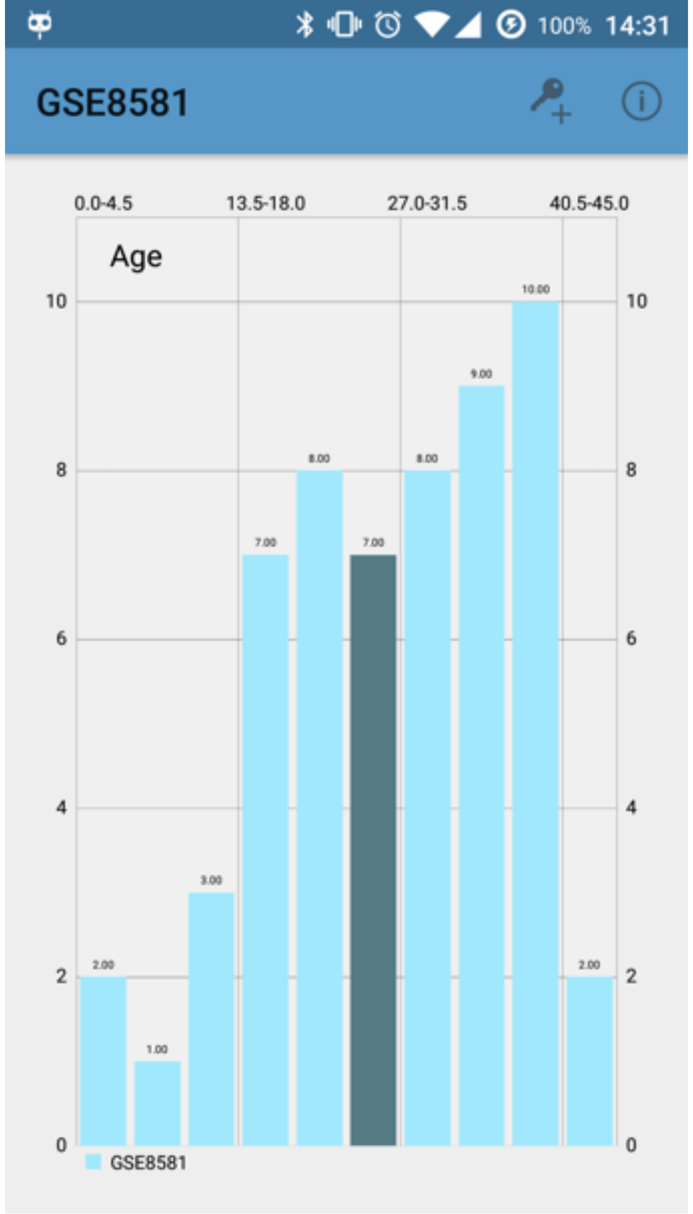
Similar

Browse the studies and data in your tranSMART instance on the go.

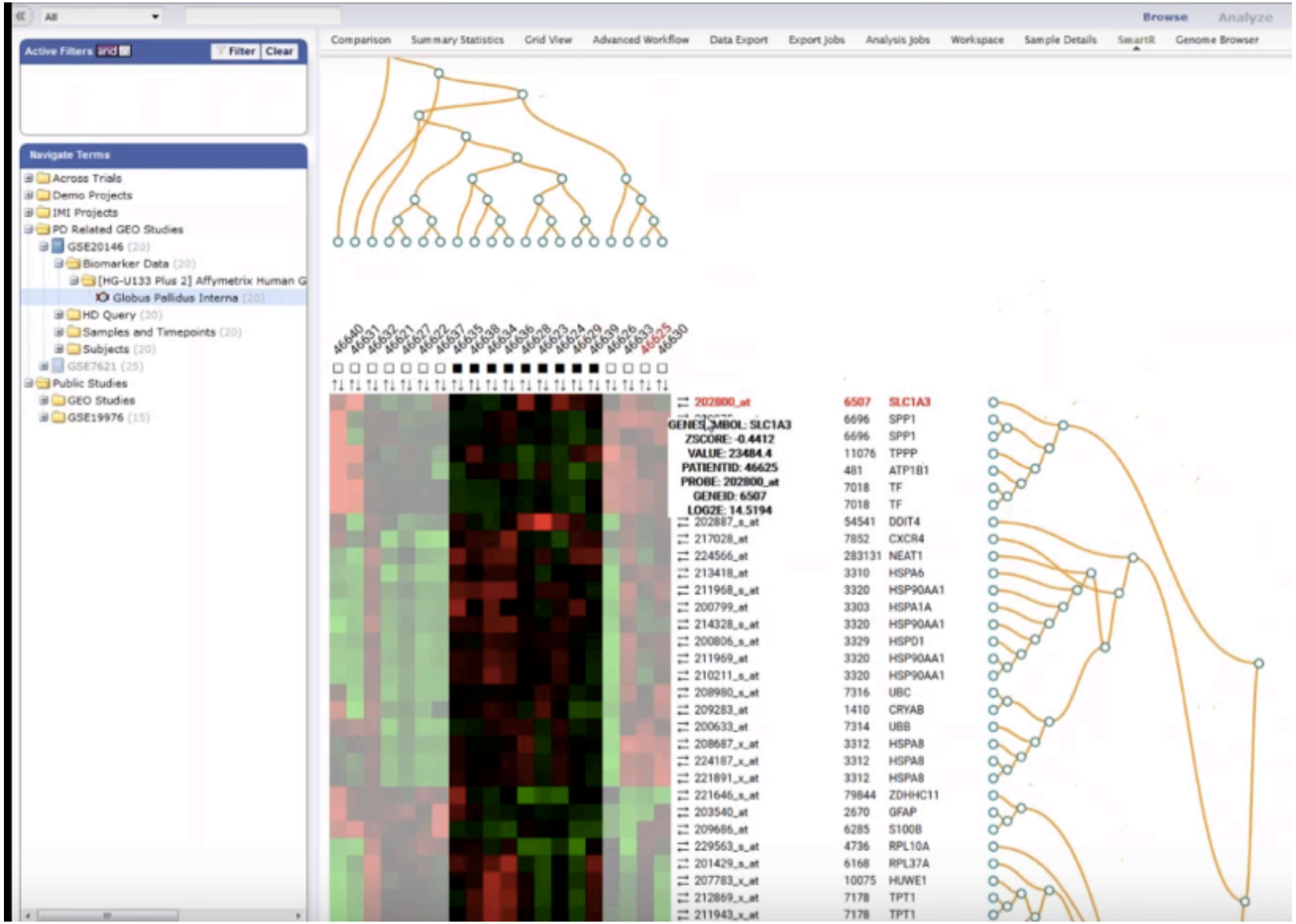
WHAT'S NEW

- Version 0.6 - The tranSMART Foundation 2015 Annual Meeting edition!

This update has been a long time coming:



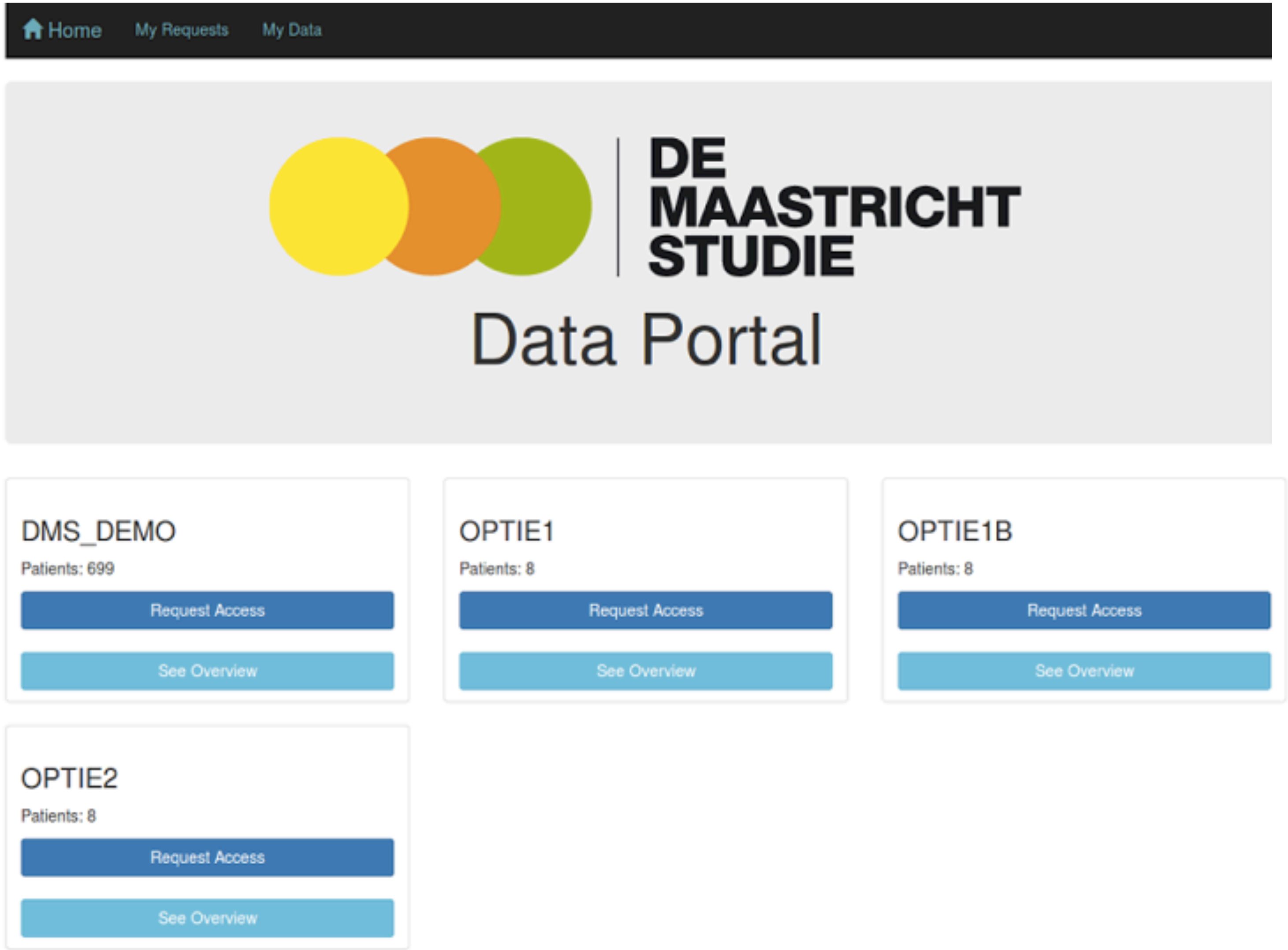
SmartR



GENESYMBOL: SLC1A3
ZSCORE: -0.4412
VALUE: 23484.4
PATIENTID: 46625
PROBE: 202800_at
GENEID: 6507
LOG2E: 14.5194

202800_at	6507	SLC1A3
[illegible]	6696	SPP1
[illegible]	6696	SPP1
[illegible]	11076	TPPP
[illegible]	481	ATP1B1
[illegible]	7018	TF
[illegible]	7018	TF
202887_s_at	54541	DDIT4
217028_at	7852	CXCR4
224566_at	283131	NEAT1
213418_at	3310	HSPA6
211968_s_at	3320	HSP90AA1
200799_at	3303	HSPA1A
214328_s_at	3320	HSP90AA1
200806_s_at	3329	HSPD1
211969_at	3320	HSP90AA1
210211_s_at	3320	HSP90AA1
208980_s_at	7316	UBC
209283_at	1410	CRYAB
200633_at	7314	UBB
208687_x_at	3312	HSPA8
224187_x_at	3312	HSPA8
221891_x_at	3312	HSPA8
221646_s_at	79844	ZDHHC11
203540_at	2670	GFAP
209686_at	6285	S100B
229563_s_at	4736	RPL10A
201429_s_at	6168	RPL37A
207783_x_at	10075	HUWE1
212869_x_at	7178	TPT1
211943_x_at	7178	TPT1

DMS Portal



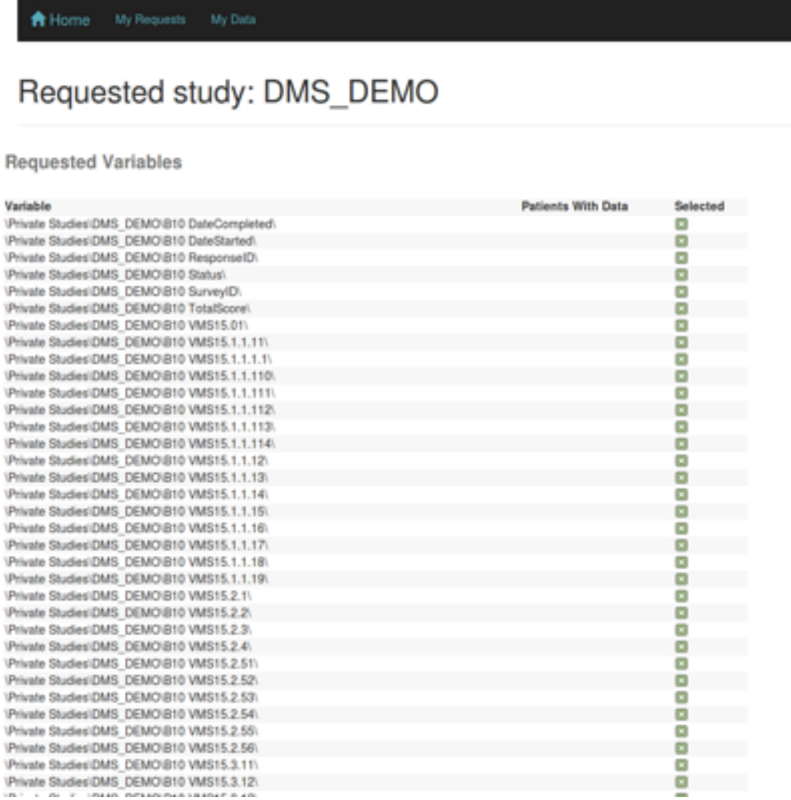
DE
MAASTRICHT
STUDIE
Data Portal

DMS_DEMO
Patients: 699

OPTIE1
Patients: 8

OPTIE1B
Patients: 8

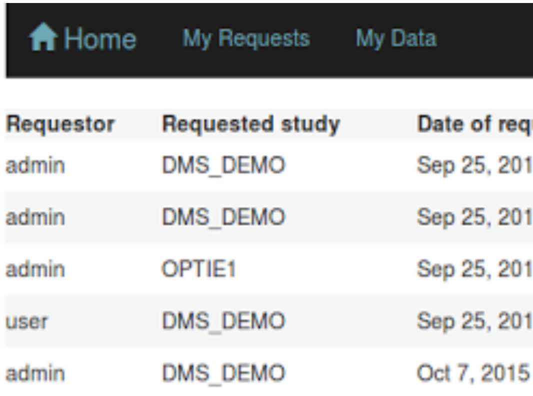
OPTIE2
Patients: 8



Requested study: DMS_DEMO

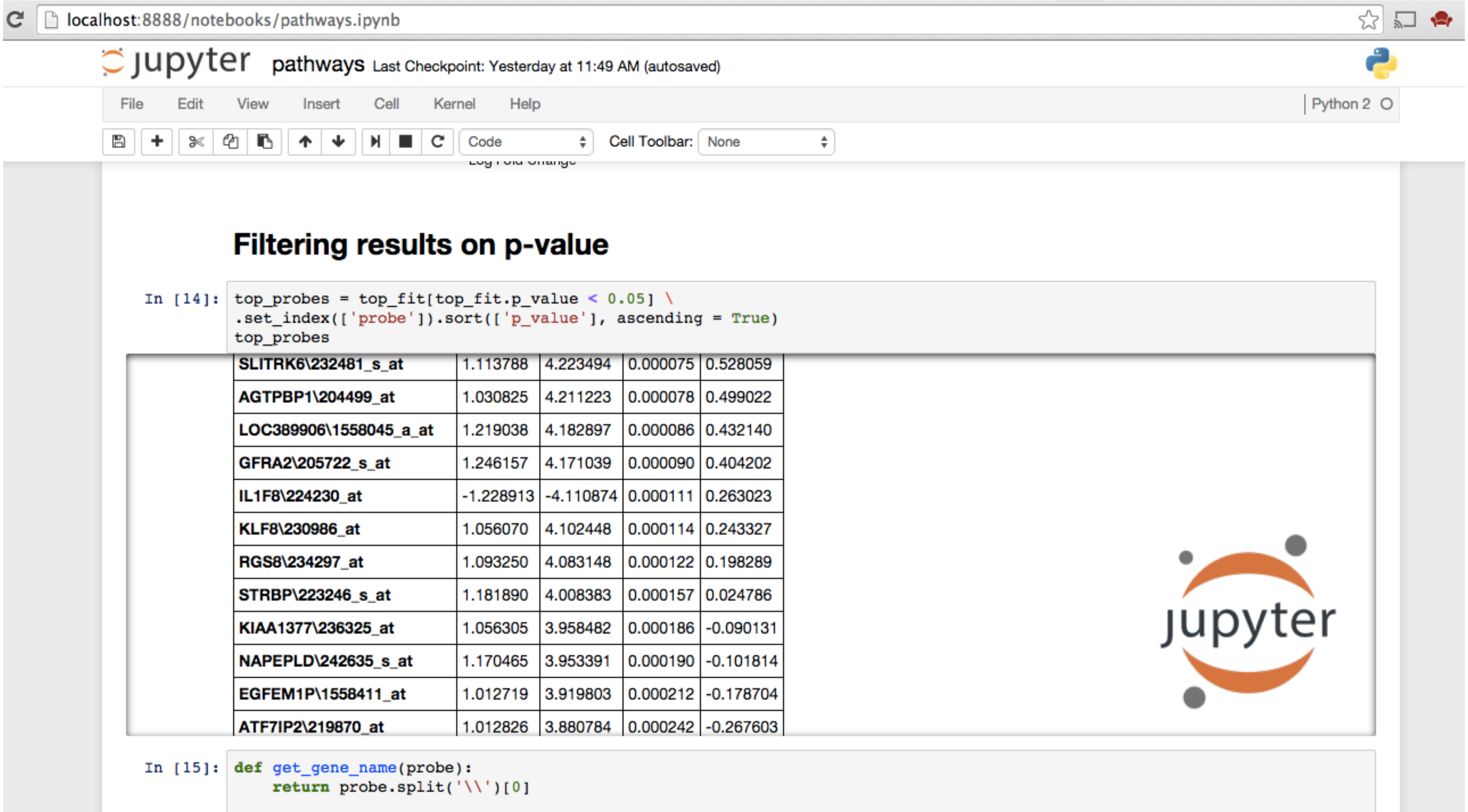
Requested Variables

Variable	Patients With Data	Selected
\Private Studies\DMS_DEMO\B10 DateCompleted\		☒
\Private Studies\DMS_DEMO\B10 DateStarted\		☒
\Private Studies\DMS_DEMO\B10 ResponseID\		☒
\Private Studies\DMS_DEMO\B10 Status\		☒
\Private Studies\DMS_DEMO\B10 SurveyID\		☒
\Private Studies\DMS_DEMO\B10 TotalScore\		☒
\Private Studies\DMS_DEMO\B10 VMS15.01\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.11\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.1.1\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.110\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.111\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.112\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.113\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.114\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.12\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.13\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.14\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.15\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.16\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.17\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.18\		☒
\Private Studies\DMS_DEMO\B10 VMS15.1.1.19\		☒
\Private Studies\DMS_DEMO\B10 VMS15.2.1\		☒
\Private Studies\DMS_DEMO\B10 VMS15.2.2\		☒
\Private Studies\DMS_DEMO\B10 VMS15.2.3\		☒
\Private Studies\DMS_DEMO\B10 VMS15.2.4\		☒
\Private Studies\DMS_DEMO\B10 VMS15.2.51\		☒
\Private Studies\DMS_DEMO\B10 VMS15.2.52\		☒
\Private Studies\DMS_DEMO\B10 VMS15.2.53\		☒
\Private Studies\DMS_DEMO\B10 VMS15.2.54\		☒
\Private Studies\DMS_DEMO\B10 VMS15.2.55\		☒
\Private Studies\DMS_DEMO\B10 VMS15.2.56\		☒
\Private Studies\DMS_DEMO\B10 VMS15.3.11\		☒
\Private Studies\DMS_DEMO\B10 VMS15.3.12\		☒



Requestor	Requested study	Date of req
admin	DMS_DEMO	Sep 25, 201
admin	DMS_DEMO	Sep 25, 201
admin	OPTIE1	Sep 25, 201
user	DMS_DEMO	Sep 25, 201
admin	DMS_DEMO	Oct 7, 2015

Jupyter



Filtering results on p-value

In [14]:

```
top_probes = top_fit[top_fit.p_value < 0.05] \
.set_index(['probe']).sort(['p_value'], ascending = True)
top_probes
```

SLITRK6\232481_s_at	1.113788	4.223494	0.000075	0.528059
AGTPBP1\204499_at	1.030825	4.211223	0.000078	0.499022
LOC389906\1558045_a_at	1.219038	4.182897	0.000086	0.432140
GFRA2\205722_s_at	1.246157	4.171039	0.000090	0.404202
IL1F8\224230_at	-1.228913	-4.110874	0.000111	0.263023
KLF8\230986_at	1.056070	4.102448	0.000114	0.243327
RGS8\234297_at	1.093250	4.083148	0.000122	0.198289
STRBP\223246_s_at	1.181890	4.008383	0.000157	0.024786
KIAA1377\236325_at	1.056305	3.958482	0.000186	-0.090131
NAPEPLD\242635_s_at	1.170465	3.953391	0.000190	-0.101814
EGFEM1P\1558411_at	1.012719	3.919803	0.000212	-0.178704
ATF7IP2\219870_at	1.012826	3.880784	0.000242	-0.267603

In [15]:

```
def get_gene_name(probe):
    return probe.split('\\')[0]
```

TranSMART Data Layer Improvements: history & outlook

TranSMART v1.1 Release (Sep 2013)

- Postgres support
- Introduction Core API
- Removed outdated i2b2 dependencies

TranSMART v1.2 release (Aug 2014)

- RESTful API
- R client
- Spotfire client

TranSMART 16.1 release (Q1 2016)

- Improvements RESTful API
- Several RESTful API based applications e.g.:
  - new cohort selection UI
  - download portal for Data Access
  - mobile client

TranSMART 16.3 release (Q3 2016)

- Improved analytics architecture
- Improved GWAS / mutation data model
- Interactive visual analytics

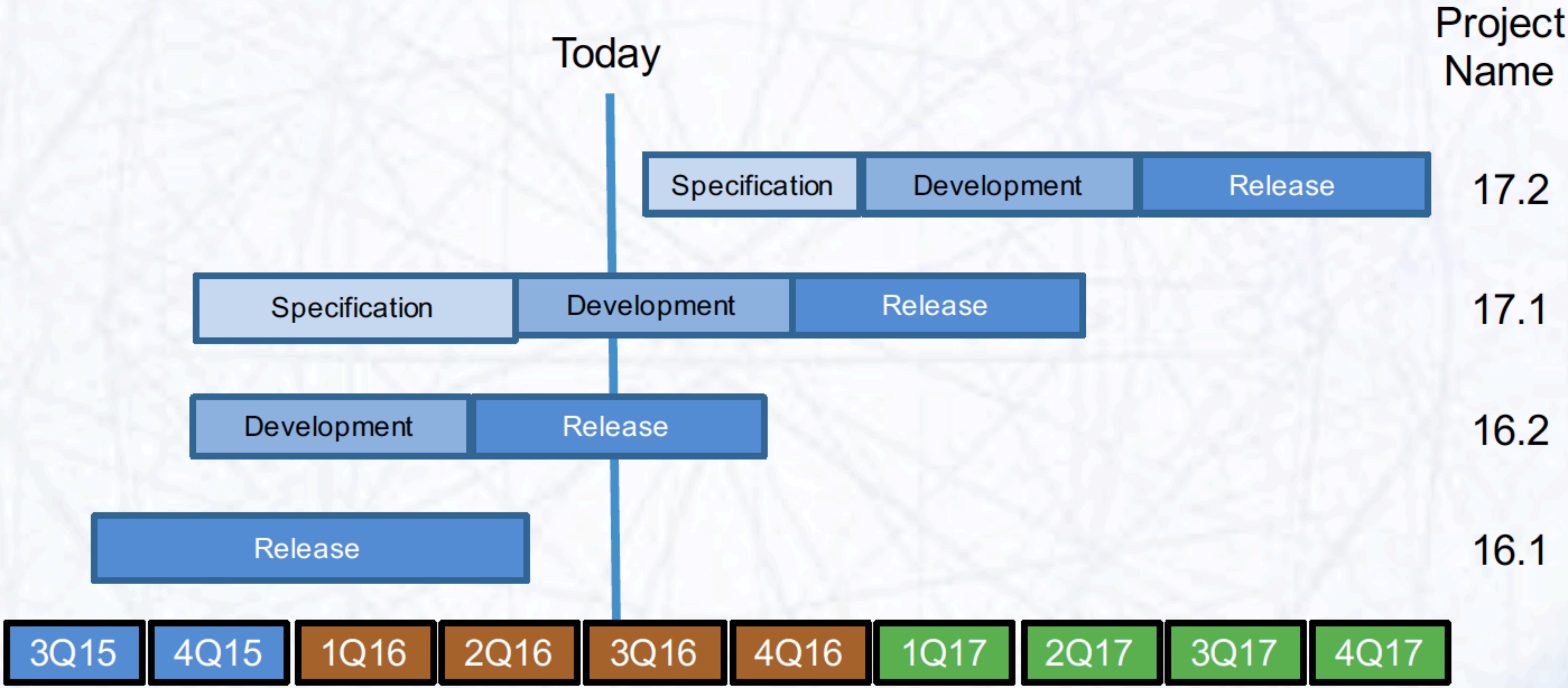
TranSMART 17.1 release (Q1 2017)

- Rebuild data layer:
- Support scalable NoSQL backends via Arvados
- Fully separate backend and frontend applications

EU Project Informatics Alignment Workshop April 2015, Imperial College London

- 18 European projects all using tranSMART to share data!
- Top requests:
  - Support for large volumes of NGS data (‘**scalable genomics backend**’)
  - Better handling of **longitudinal data**
  - Better handling of **cross-trial** data loading & browsing
  - Improve **cohort selection** on high dimensional data
  - Improve ETL, logging, auditing, visualizations etc. etc.

TranSMART Foundation Roadmap



Today

Project Name

Specification | Development | Release — 17.2

Specification | Development | Release — 17.1

Development | Release — 16.2

Release — 16.1

3Q15 | 4Q15 | 1Q16 | 2Q16 | 3Q16 | 4Q16 | 1Q17 | 2Q17 | 3Q17 | 4Q17

TranSMART 17.1

- **Clinical Data Improvements:** Changes to the data model and APIs for clinical data
  - To support **longitudinal and EHR** data, both with absolute and relative time
  - To improve **cross-study querying** support and **vocabulary** annotations
- **Scalable Genomics Backend:** Changes to the data model and APIs
  - Integration with Arvados allow storage & querying of large scale **genomics data**
- Improve the TranSMART Core with API *documentation* and *testing*
- All this while maintaining **backwards compatibility**!

Example Deployment / Architecture

Storage

Arvados Keep
or Amazon S3

Compute

Arvados Crunch
or Apache Spark
e.g. on HPC
cluster

Data warehouse
clinical + omics
studies + samples

TranSMART Core
DB

Data Processing

Arvados Workbench
ETL (e.g. Arborist)

Data Browsing

TranSMART Base UI

Data Analysis

SmartR
Base UI
R API

Integrated
Microservices

Ontologies

Persistent Identifiers
User authorization
Workflows/Protocols
Visualizations

Analysis results

Towards ‘2.0’ version of tranSMART **Platform**

SECURE

SCALABLE

MODULAR & INTEROPERABLE

DOCUMENTED

the hyve

We empower scientists by building on open source software