You are free to:

Copy, share, adapt, or re-mix;

Photograph, film, or broadcast;

Blog, live-blog, or post video of;

This presentation. Provided that:

You attribute the work to its author and respect the rights and licenses associated with its components.

@bffo on

OICR
Ontario Institute for Cancer Research

The OICR and The International Cancer Genomics Consortium

October 22th 2012

B.F. Francis Ouellette francis@oicr.on.ca

- Associate Director, Informatics & Biocomputing, Ontario Institute for Cancer Research, Toronto, ON
- Associate Professor, Department of Cell and Systems Biology, University of Toronto, Toronto, ON.

@bffo on

Outline

- OICR's mission
- ICGC's goal
- OICR and ICGC: Open Access/Open Source shop

- **ICGC**: the DCC
- **OICR**: Processing Cancer Genomes
- **You**: getting access to the data

OICR's mission

To build innovative research programs that will have an impact on the prevention, early detection, diagnosis and treatment of cancer.

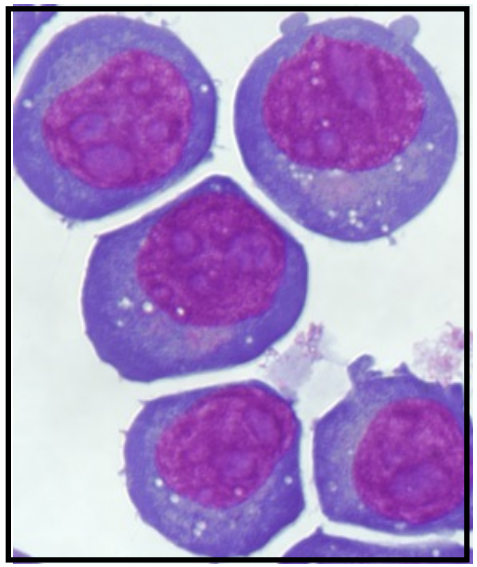
ICGC's Goal:

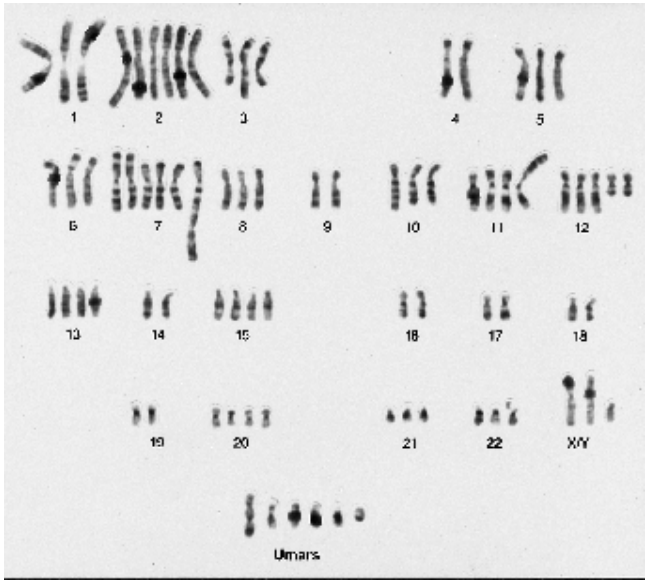
To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.

OICR
Ontario Institute for Cancer Research

Cancer

A Disease of the Genome





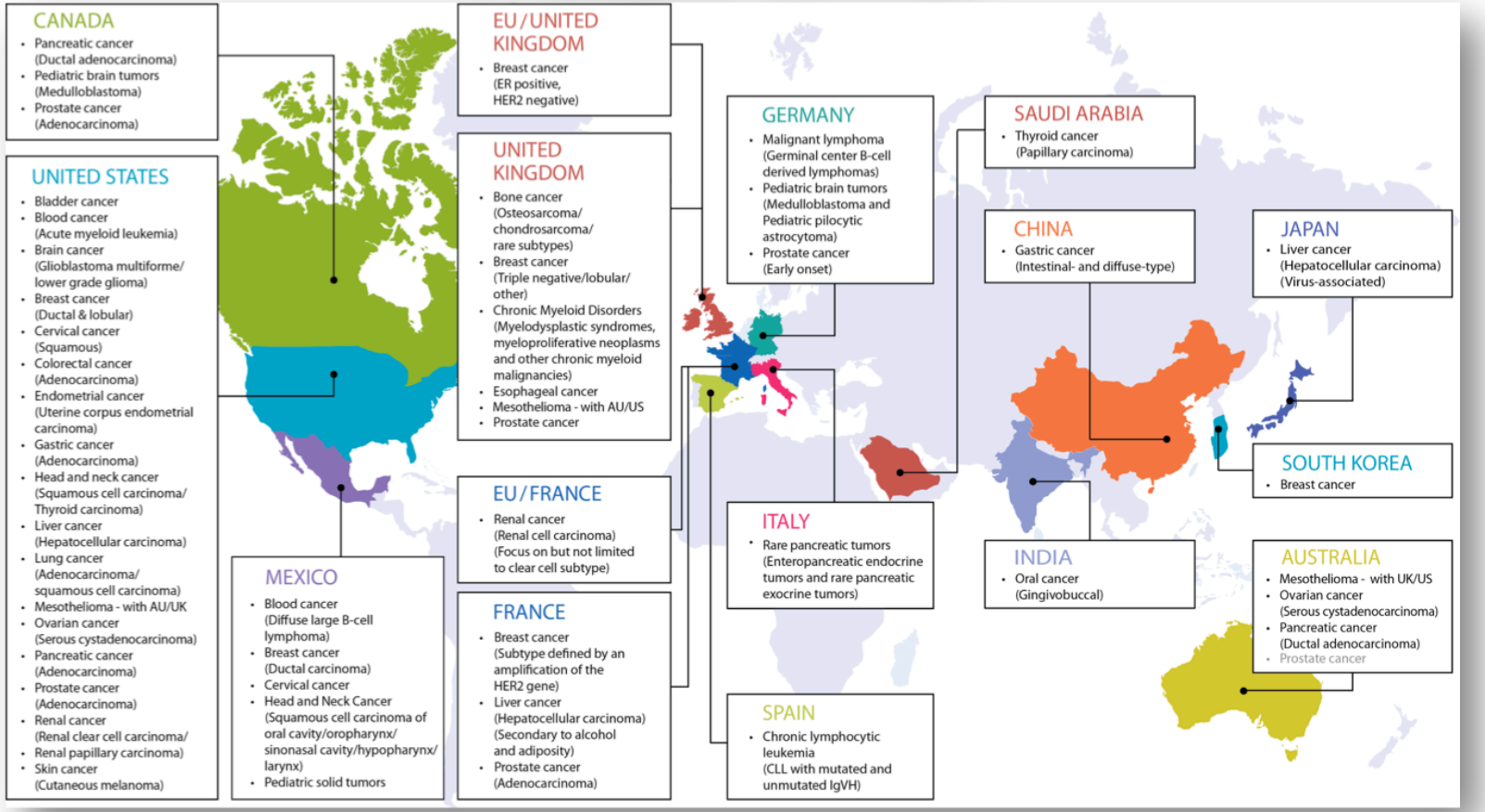
Challenge in Treating Cancer:

- Every tumor is different
- Every cancer patient is different

OICR
Ontario Institute for Cancer Research

ICGC Map – September 2012

47 projects launched



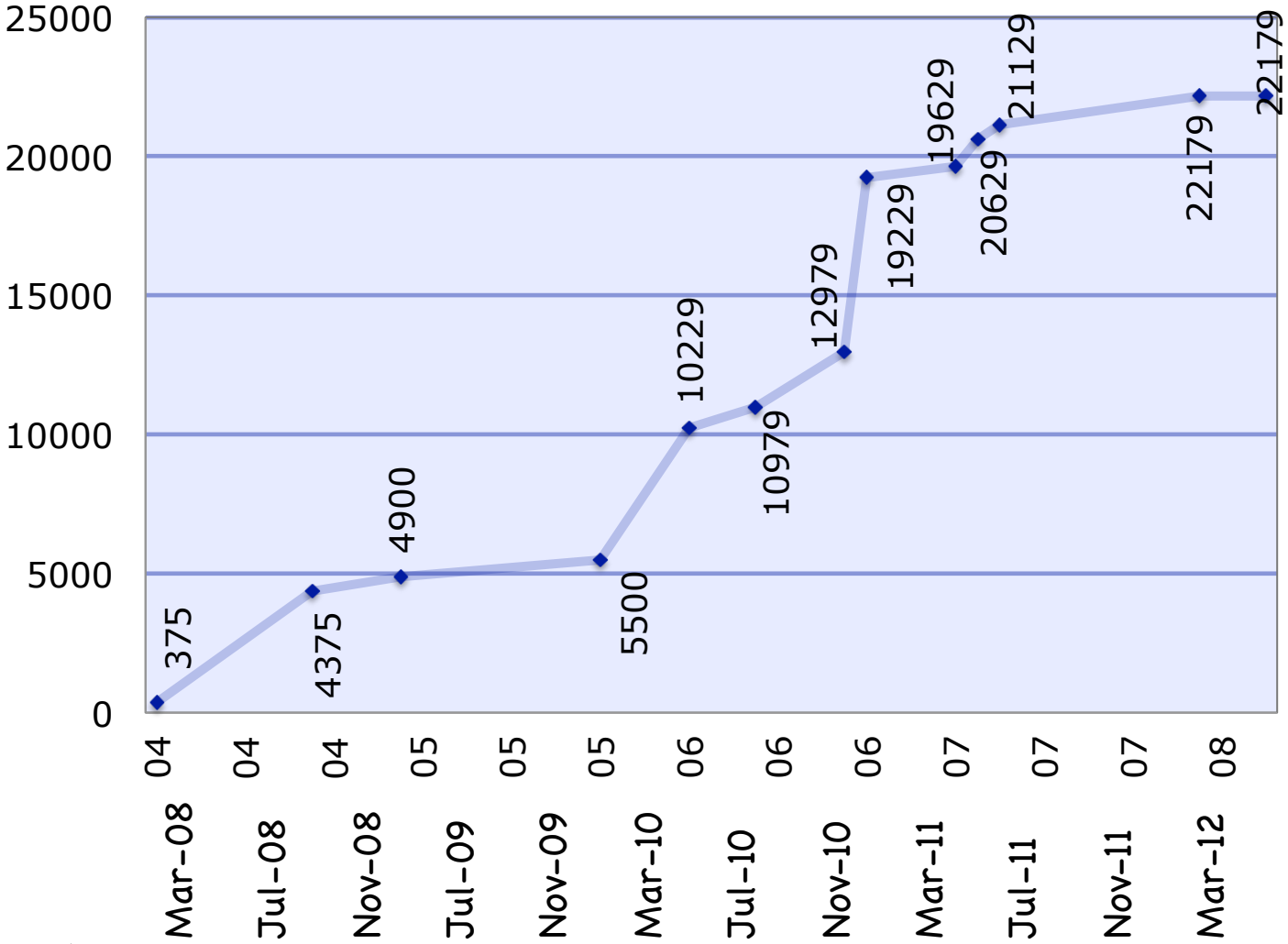
47 Projects 12 countries 23,408 tumor samples planned

OICR Policies on Open Access Publication and Data Retention

- To allow and promote access to research outputs funded by OICR, thus increasing the diffusion and impact of the research process.
- All papers will be freely available through the internet within six (6) months of publication.
- OICR will not violate the Publisher's embargo policy on free access
- OICR encourages OA publication, but is also developing an Institutional Repository (IR) where research output will be found

ICGC – March 2012

Commitments for 22,179 tumor genomes!



New

Saudi Arabia
Thyroid

South Korea
Breast

AU/UK/US
Mesothelioma

OICR
Ontario Institute for Cancer Research

Completeness of Data for Genomic Analysis Types in DCC Datasets (ICGC 10)

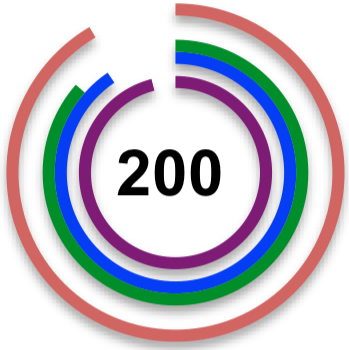
Brett Whitty

- Simple Somatic Mutations
- Copy Number Alterations
- Structural Variation
- Gene Expression
- miRNA Expression
- Splicing Variation
- DNA Methylation

OICR
Ontario Institute for Cancer Research

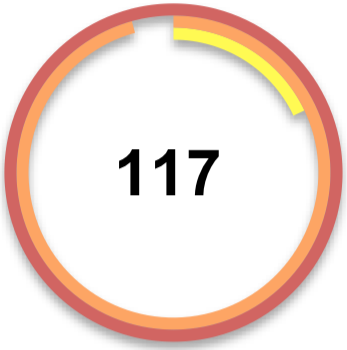
Completeness of Genomic Analysis Data Types in DCC Datasets

Brett Whitty



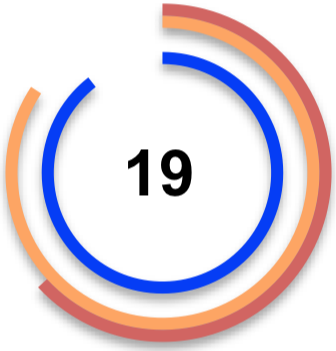
Acute Myeloid Leukemia (TCGA, US)

Bladder Urothelial Carcinoma (TCGA, US)



Breast Carcinoma (WTSI, UK)

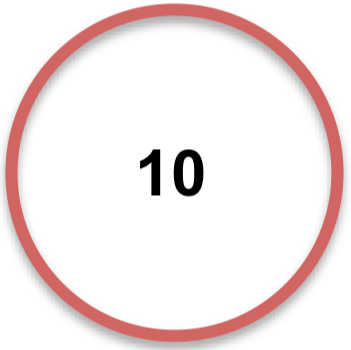
Breast Invasive Carcinoma (TCGA, US)



Cervical Squamous Cell Carcinoma (TCGA, US)

Chronic Lymphocytic Leukemia (ISC/MICINN, ES)

Colon Adenocarcinoma (TCGA, US)



Gastric Cancer (CCGC, CN)

Simple Somatic Mutations | Copy Number Alterations | Structural Variation | Gene Expression | miRNA Expression | Splicing Variation | DNA Methylation

Completeness of Genomic Analysis Data Types in DCC Datasets

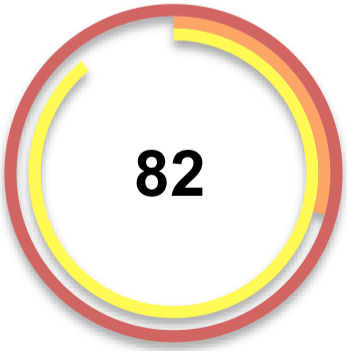
Brett Whitty

Glioblastoma Multiforme (TCGA, US)

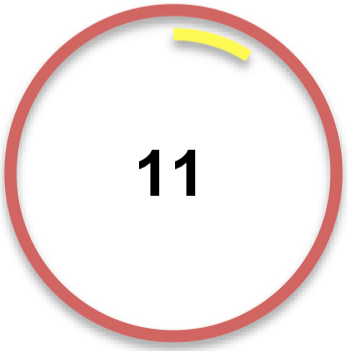
Head and Neck Squamous Cell Carcinoma (TCGA, US)

Kidney Renal Clear Cell Carcinoma (TCGA, US)

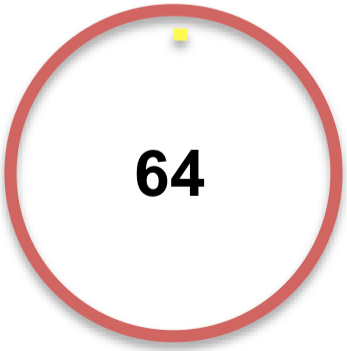
Kidney Renal Papillary Cell Carcinoma (TCGA, US)



Liver Cancer (INCa, FR)



Liver Cancer (NCC, JP)



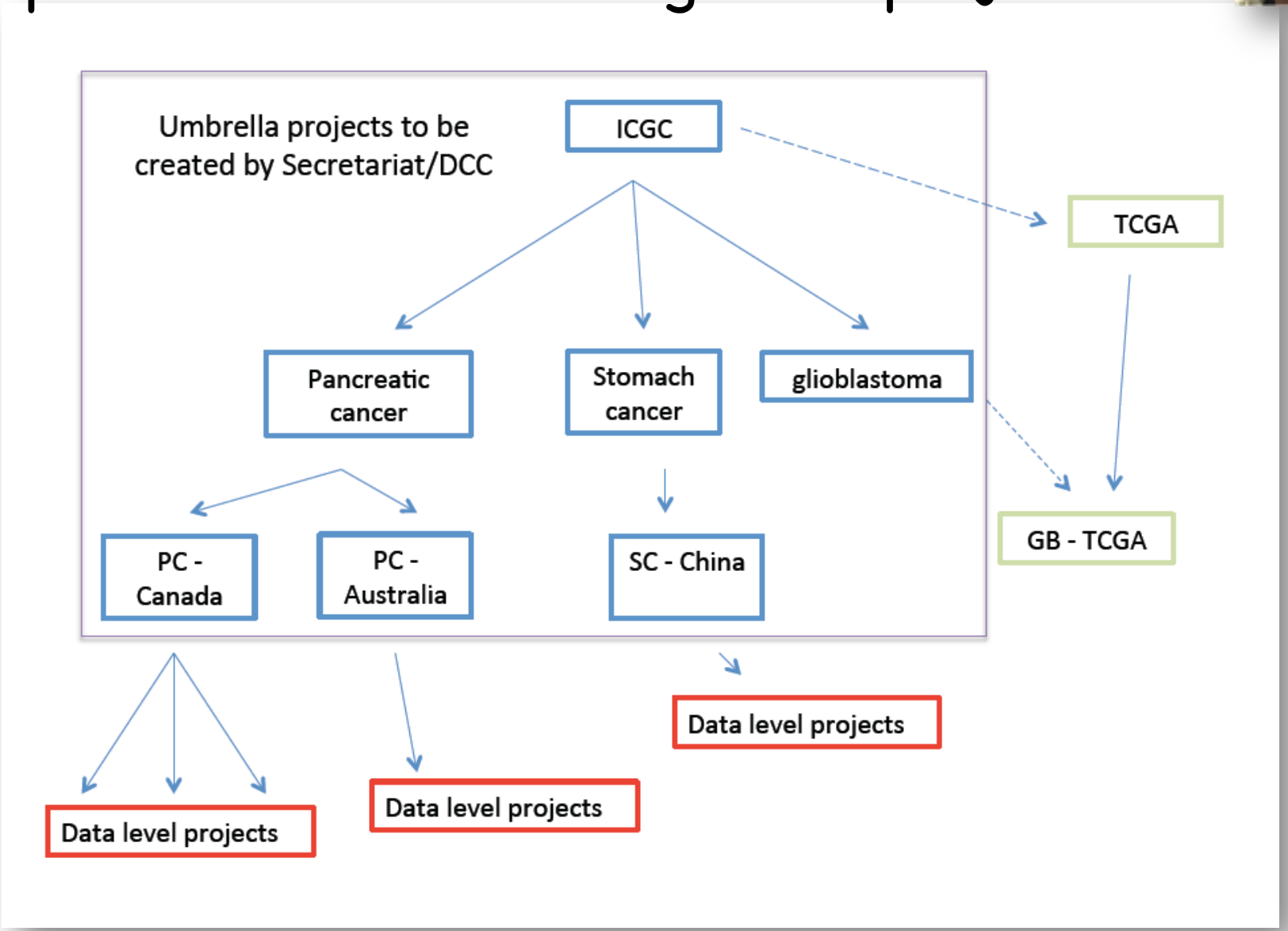
Liver Cancer (RIKEN, JP)

Liver Hepatocellular Carcinoma (TCGA, US)

OICR Ontario Institute for Cancer Research

Simple Somatic Mutations | Copy Number Alterations | Structural Variation | Gene Expression | miRNA Expression | Splicing Variation | DNA Methylation

http://www.ncbi.nlm.nih.gov/bioproject



OICR
Ontario Institute for Cancer Research

ICGC Data Categories

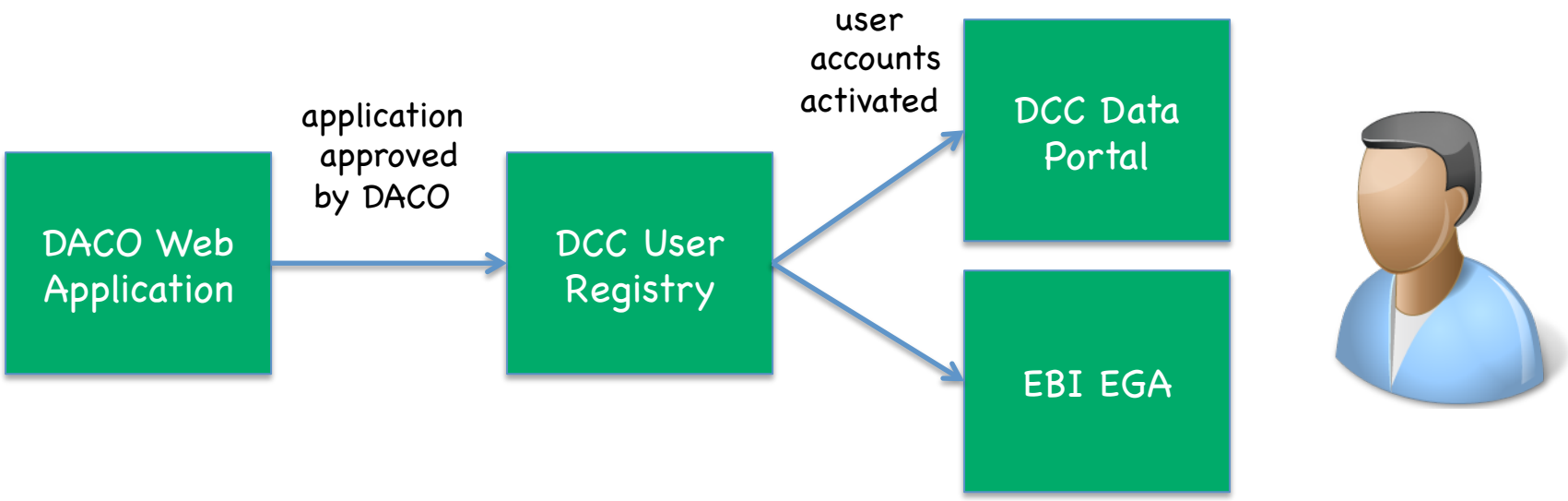
ICGC Open Access Datasets	ICGC Controlled Access Datasets
➢ Cancer Pathology Histologic type or subtype Histologic nuclear grade ➢ Donor Gender Age range ➢ RNA expression (normalized) ➢ DNA methylation ➢ Genotype frequencies ➢ Somatic mutations (SNV, CNV and Structural Rearrangement)	➢Detailed Phenotype and Outcome Data Patient demography Risk factors Examination Surgery/Drugs/Radiation Sample/Slide Specific histological features Protocol Analyte/Aliquot ➢Gene Expression (probe-level data) ➢Raw genotype calls (germline) ➢Gene-sample identifier links ➢Genome sequence files

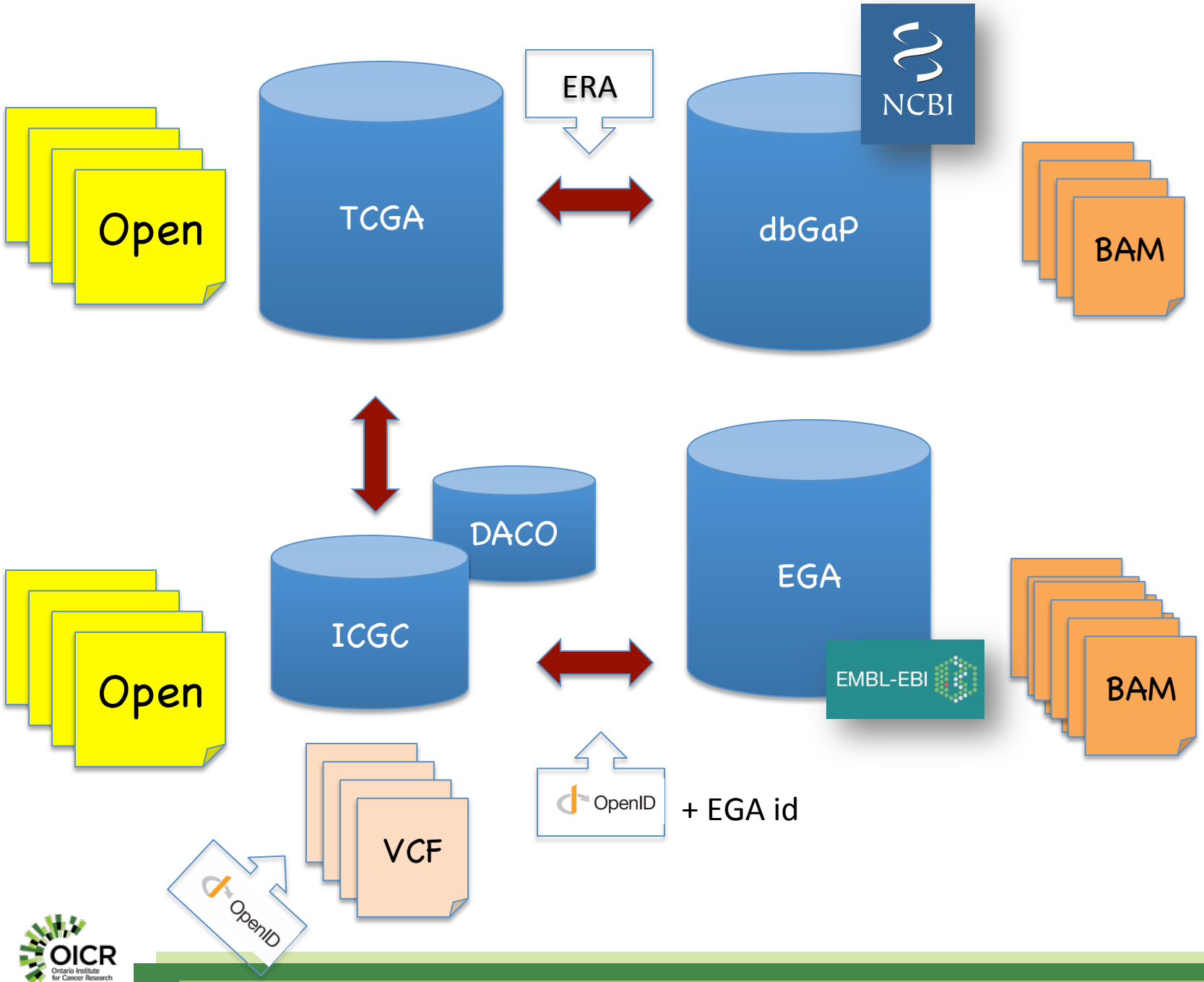
Most of the data in the portal is publically available without restriction. However, access to some data, like the germline mutations, requires authorization by the Data Access Compliance Office (DACO)

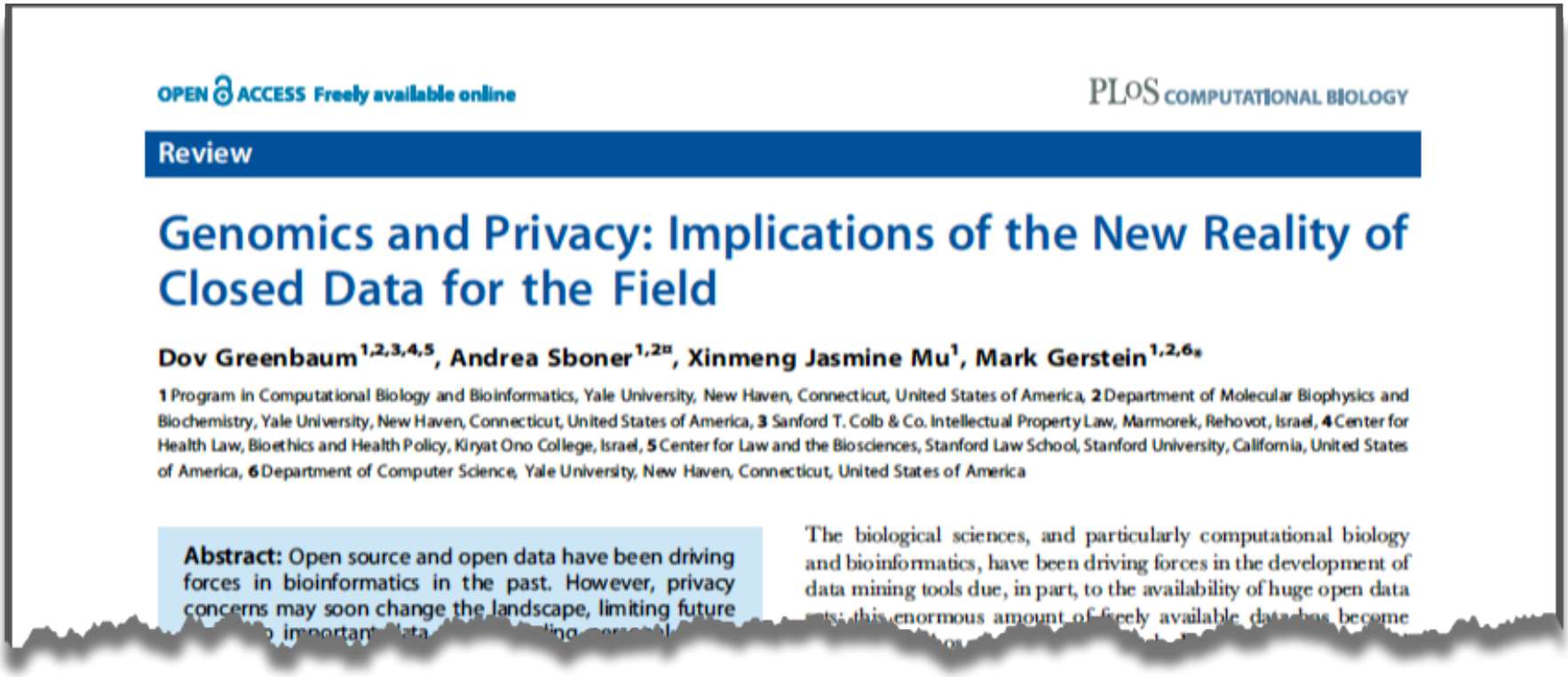
OICR
Ontario Institute for Cancer Research

DACO/DCC User Data Access Process

- Users approved through DACO are now automatically granted access to ICGC controlled access datasets available through the ICGC Data Portal and the EBI's EGA repository




Open
TCGA
ERA
NCBI
dbGaP
BAM
DACO
ICGC
EGA
Open
EMBL-EBI
BAM
OpenID
+ EGA id
VCF
OpenID



OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Review

Genomics and Privacy: Implications of the New Reality of Closed Data for the Field

Dov Greenbaum[1,2,3,4,5], **Andrea Sboner**[1,2¤], **Xinmeng Jasmine Mu**[1], **Mark Gerstein**[1,2,6*]

1 Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America, **2** Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America, **3** Sanford T. Colb & Co. Intellectual Property Law, Marmorek, Rehovot, Israel, **4** Center for Health Law, Bioethics and Health Policy, Kiryat Ono College, Israel, **5** Center for Law and the Biosciences, Stanford Law School, Stanford University, California, United States of America, **6** Department of Computer Science, Yale University, New Haven, Connecticut, United States of America

Abstract: Open source and open data have been driving forces in bioinformatics in the past. However, privacy concerns may soon change the landscape, limiting future

The biological sciences, and particularly computational biology and bioinformatics, have been driving forces in the development of data mining tools due, in part, to the availability of huge open data

"The administrative efforts to access private genetic data exact a real cost and create a drag on research efforts creating friction in the depositing, accessing, and analyzing of data. With many academics risk averse and cost conscious the time and effort often necessary to access this data will cut down on potential research efforts."

OICR
Ontario Institute for Cancer Research

John McPherson

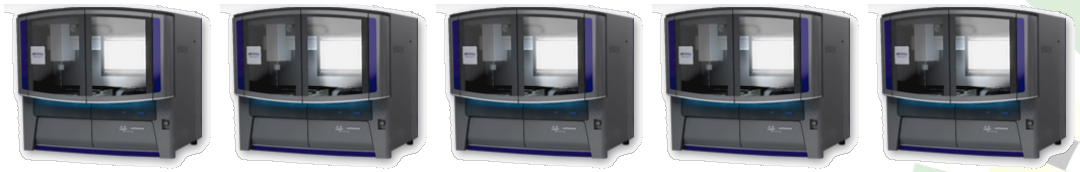
OICR Sequencing/Biocomputing Platform

Illumina HiSeq 2000

GAII

Ion Torent

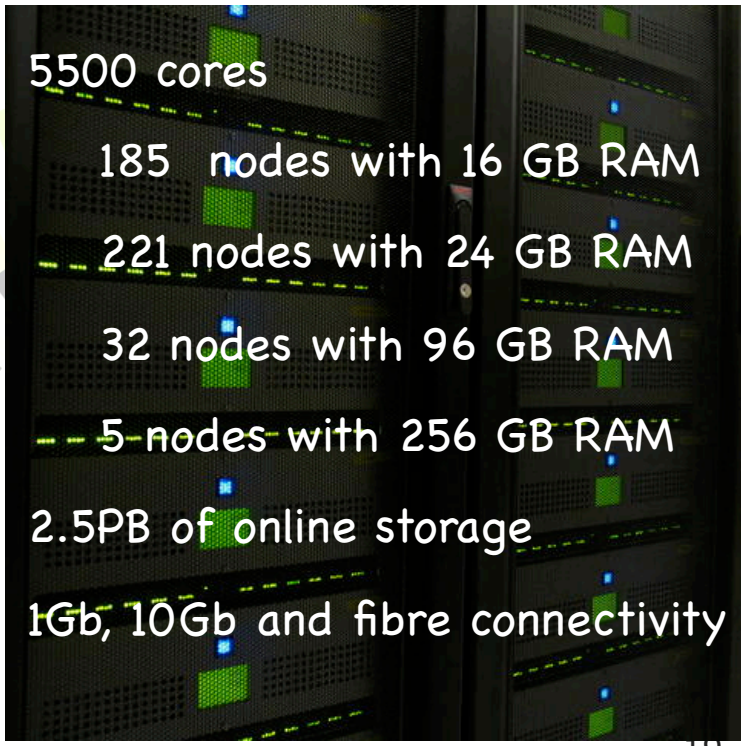
MiSeq



Life Tech Solid 5500

Pac Bio

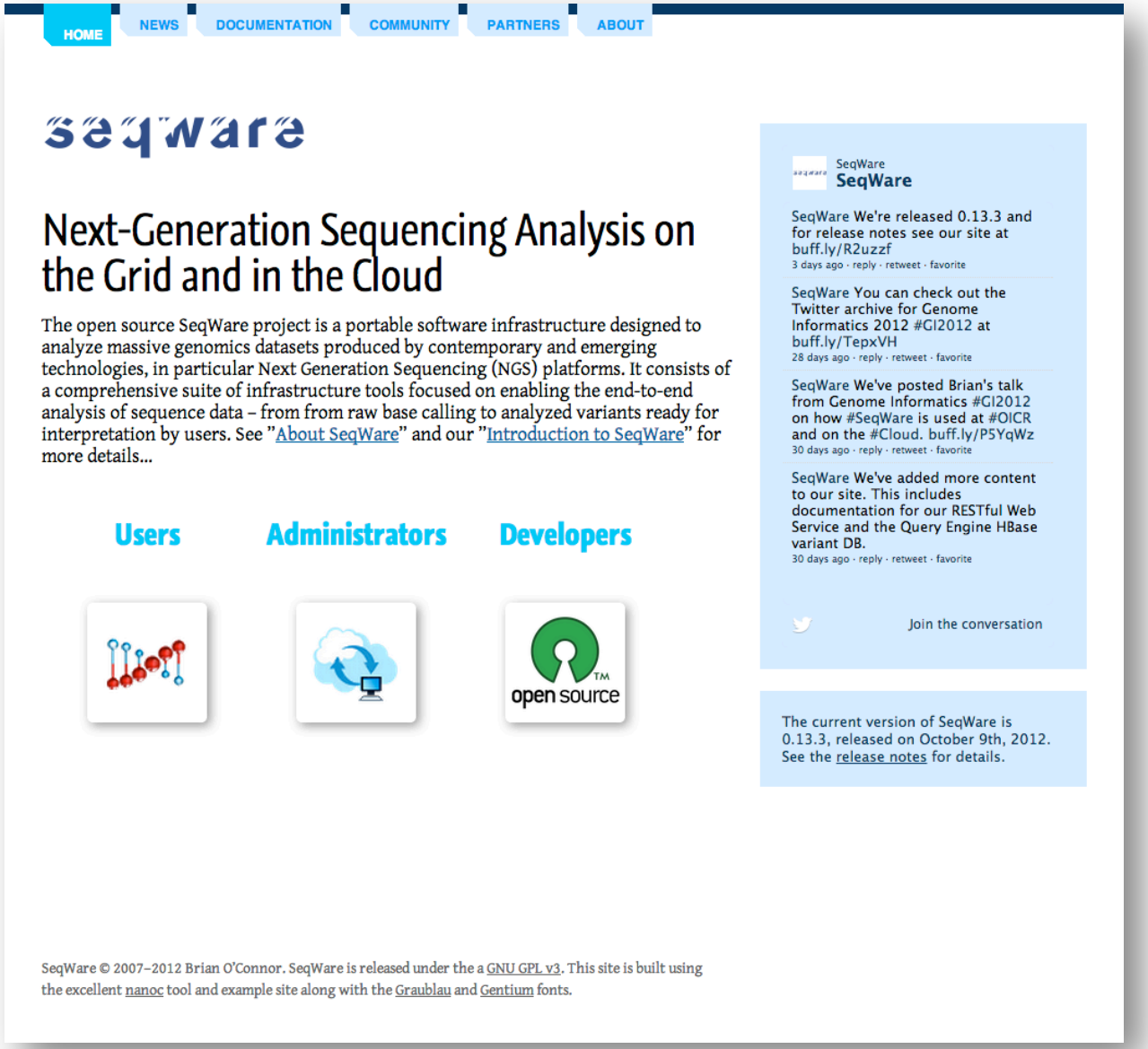
> 17 terabases per month
> 2,800 human genomes capacity and growing (70 genomes at 40X)



OICR data analysis pipeline

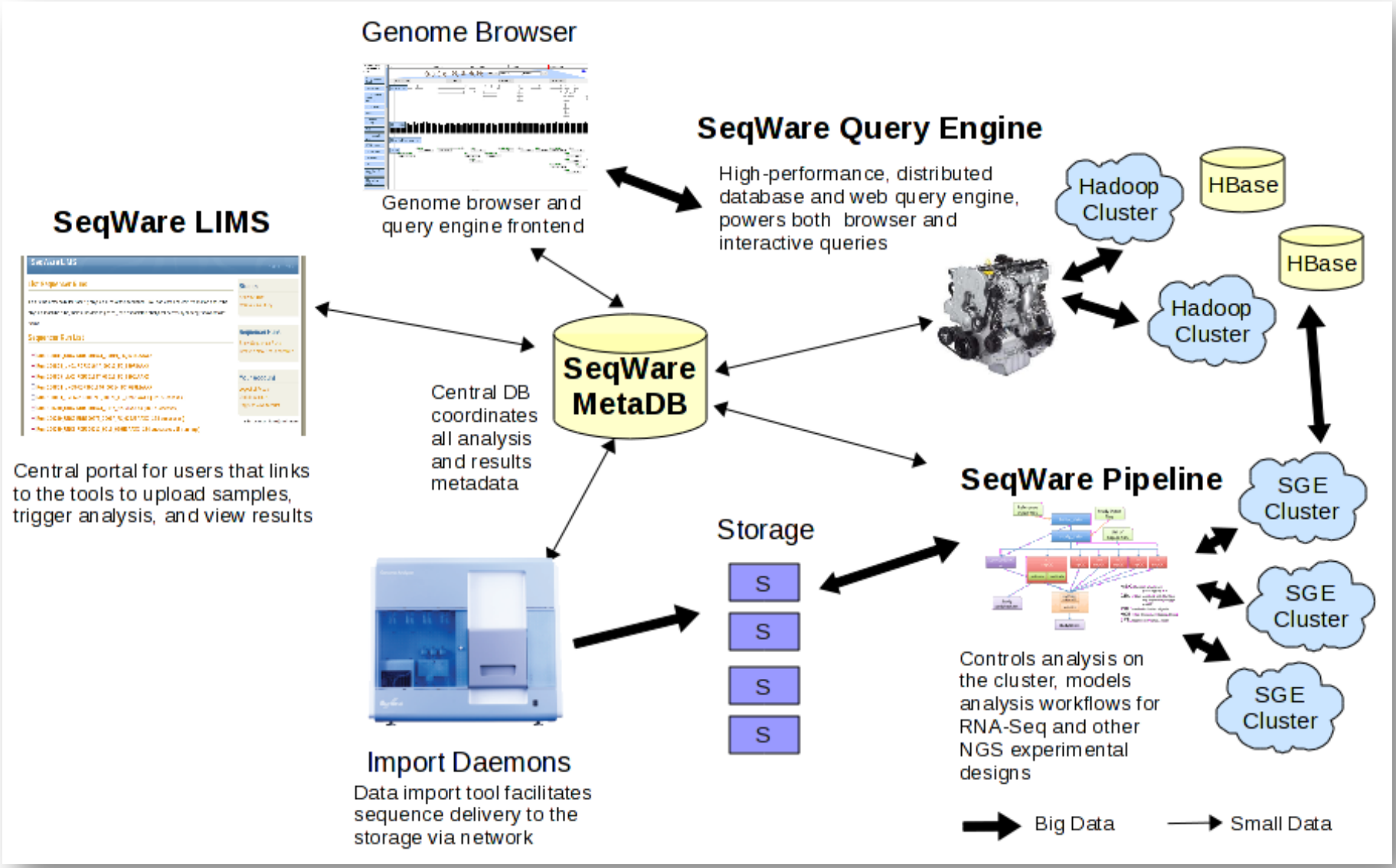
- Like most genome/bioinformatics centers, we are fully dependent on OS NGS bioinformatics tools.
- We all depend on:
  - SeqAnswers.com
  - biostars.org
- Pipelines are necessary because they:
  - Are more scalable
  - Are more recordable
  - Are more reproducible
  - Are more robust
  - … and can keep you sane!

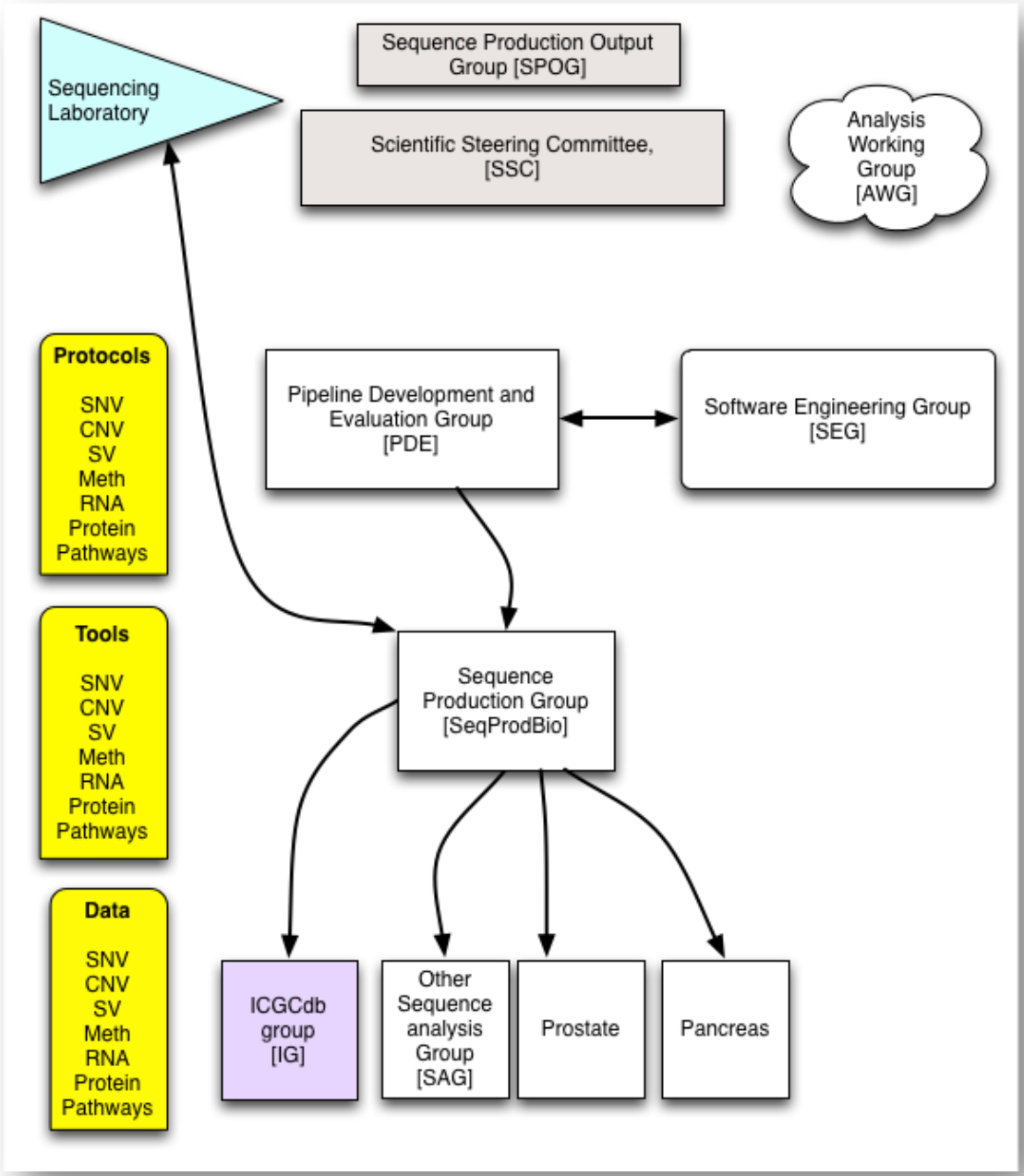
http://seqware.github.com/



OICR
Ontario Institute for Cancer Research

SeqWare: http://seqware.github.com/about/

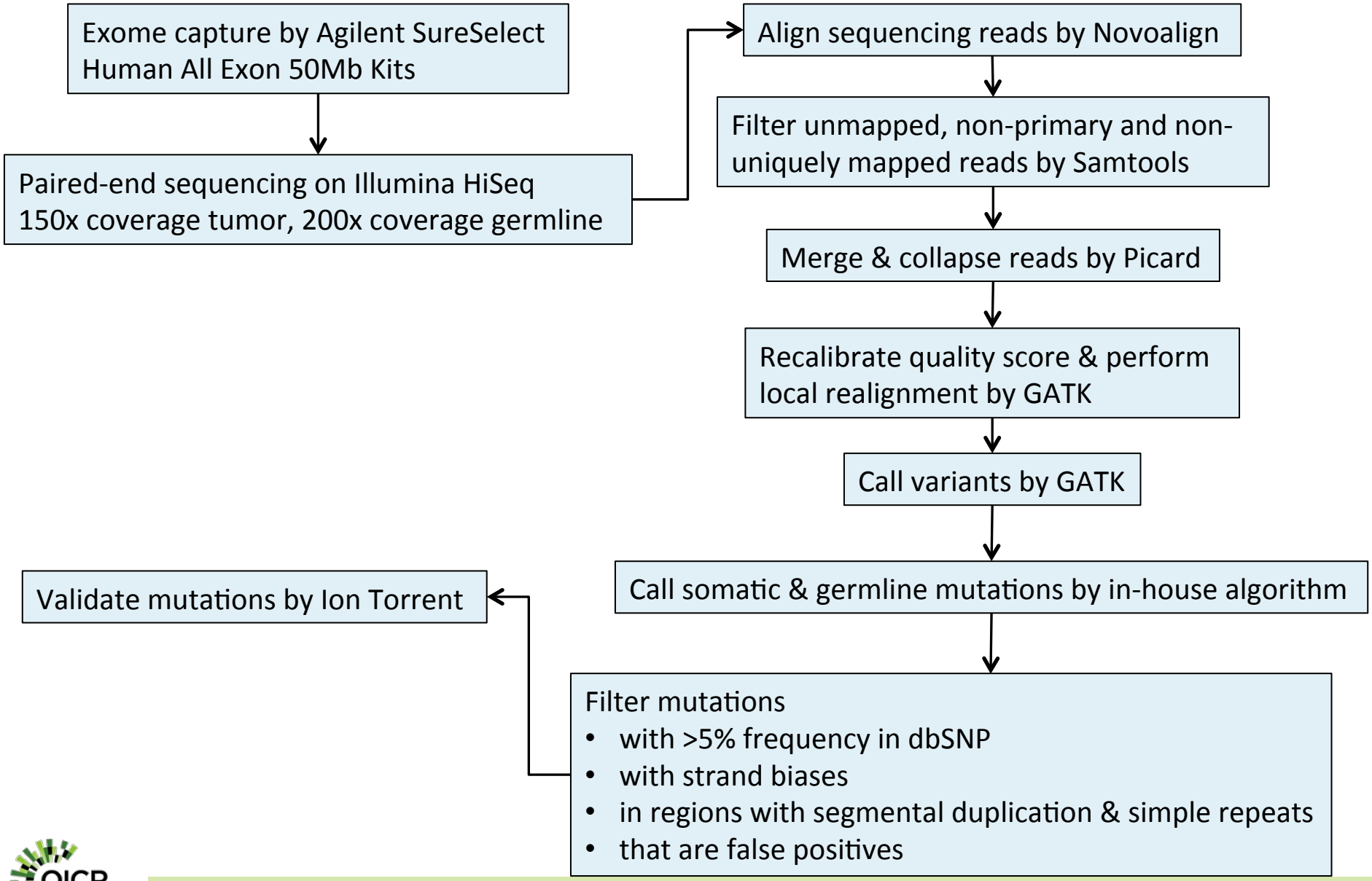



Sequencing Laboratory
Sequence Production Output Group [SPOG]
Scientific Steering Committee, [SSC]
Analysis Working Group [AWG]
Protocols
SNV
CNV
SV
Meth
RNA
Protein
Pathways
Pipeline Development and Evaluation Group [PDE]
Software Engineering Group [SEG]
Tools
SNV
CNV
SV
Meth
RNA
Protein
Pathways
Sequence Production Group [SeqProdBio]
Data
SNV
CNV
SV
Meth
RNA
Protein
Pathways
ICGCdb group [IG]
Other Sequence analysis Group [SAG]
Prostate
Pancreas

What do we do to maximize good calls?

- Minimal coverage of tumor and germline for **exome**:
  - 200x germline
  - 150x tumor
- Minimum quality score
- Simultaneous alignment of reference, normal and tumor
- Blacklist "bad" regions
- Remove suspiciously dense clusters of mutations (perhaps too aggressive)
- Validate, validate, validate!
- Future ideas
  - Assemble germline first, then align tumour to germline
  - Build patient-specific blacklist

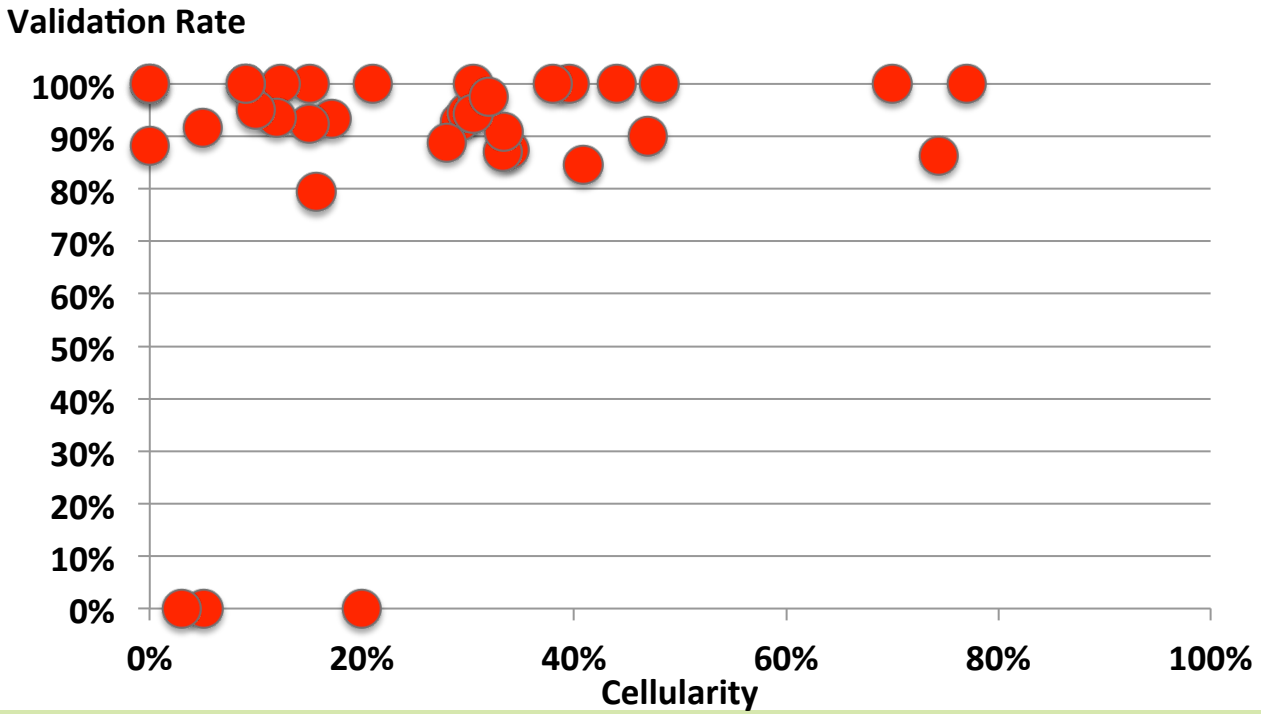
Exome Sequencing Pipeline



Validation Strategy

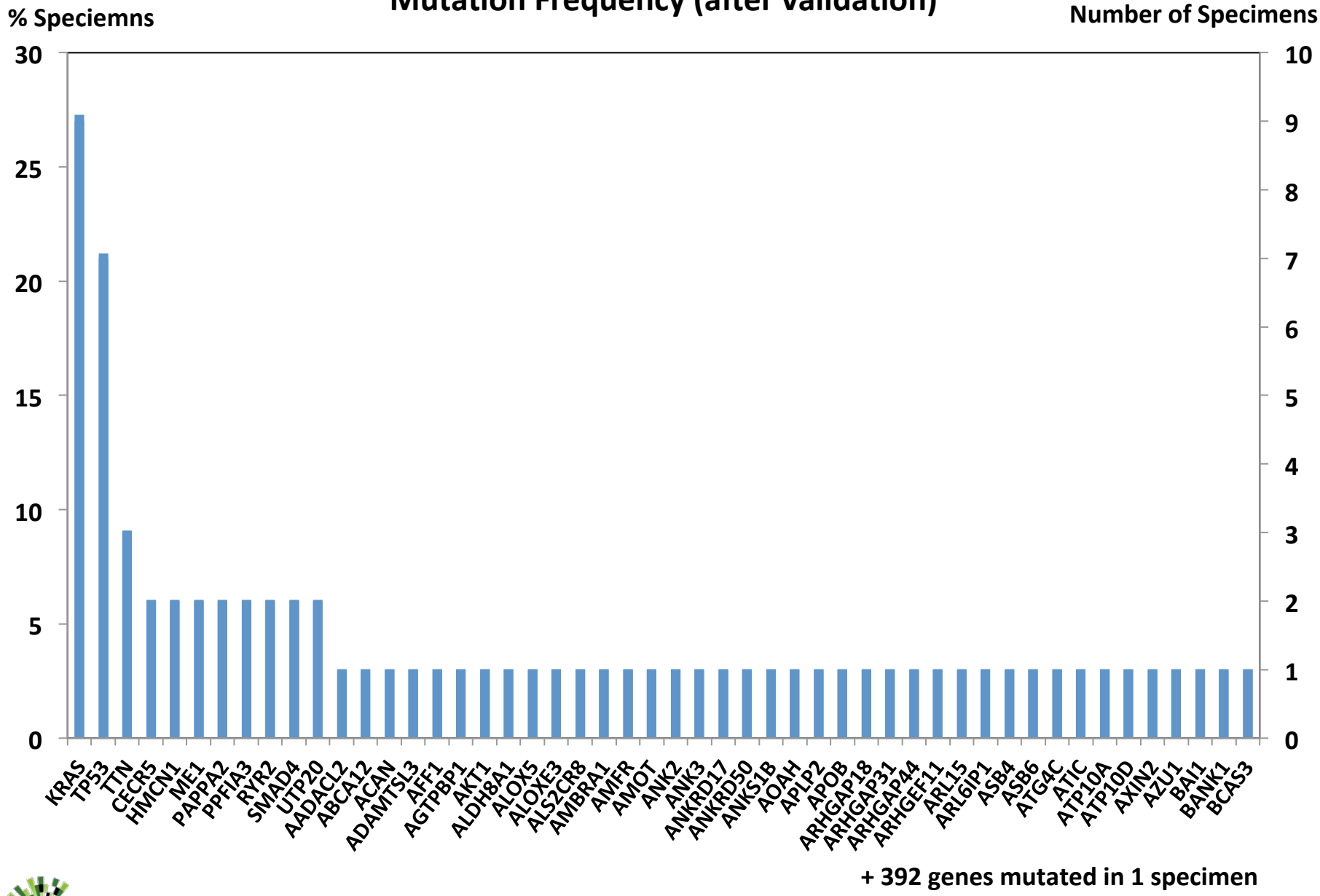
Lincoln Stein

- false positives
- false negatives
- Validation rate was an average of 87%
- No correlation between cellularity and validation rate indicating that the pipeline calls SNVs accurately irrespective of cellularity



Lincoln Stein

Mutation Frequency (after validation)

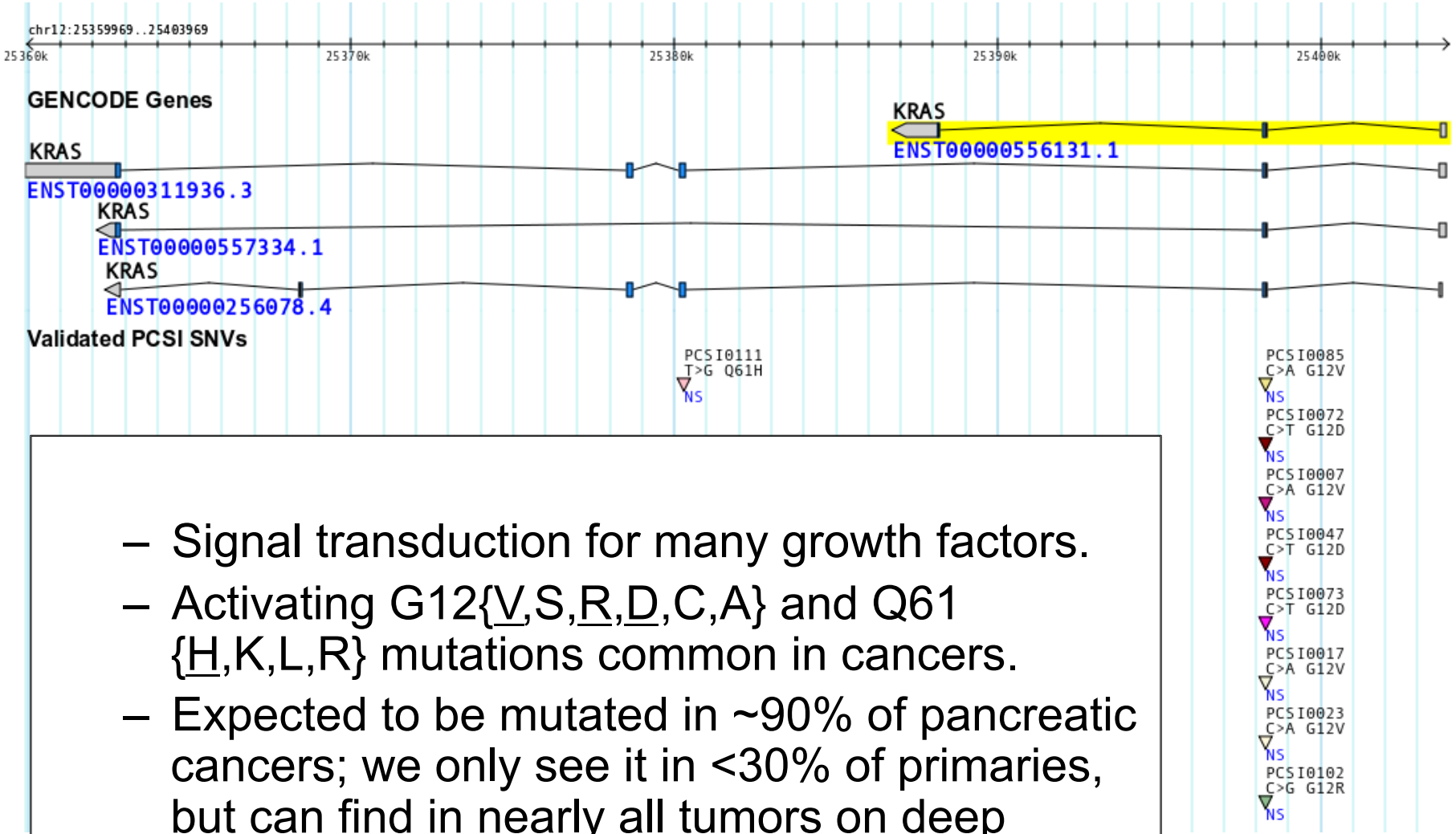


+ 392 genes mutated in 1 specimen

OICR
Ontario Institute for Cancer Research

Lincoln Stein

KRAS (mutated in 9 samples)



- Signal transduction for many growth factors.
- Activating G12{V,S,R,D,C,A} and Q61 {H,K,L,R} mutations common in cancers.
- Expected to be mutated in ~90% of pancreatic cancers; we only see it in <30% of primaries, but can find in nearly all tumors on deep sequencing (false neg rate > 60%)

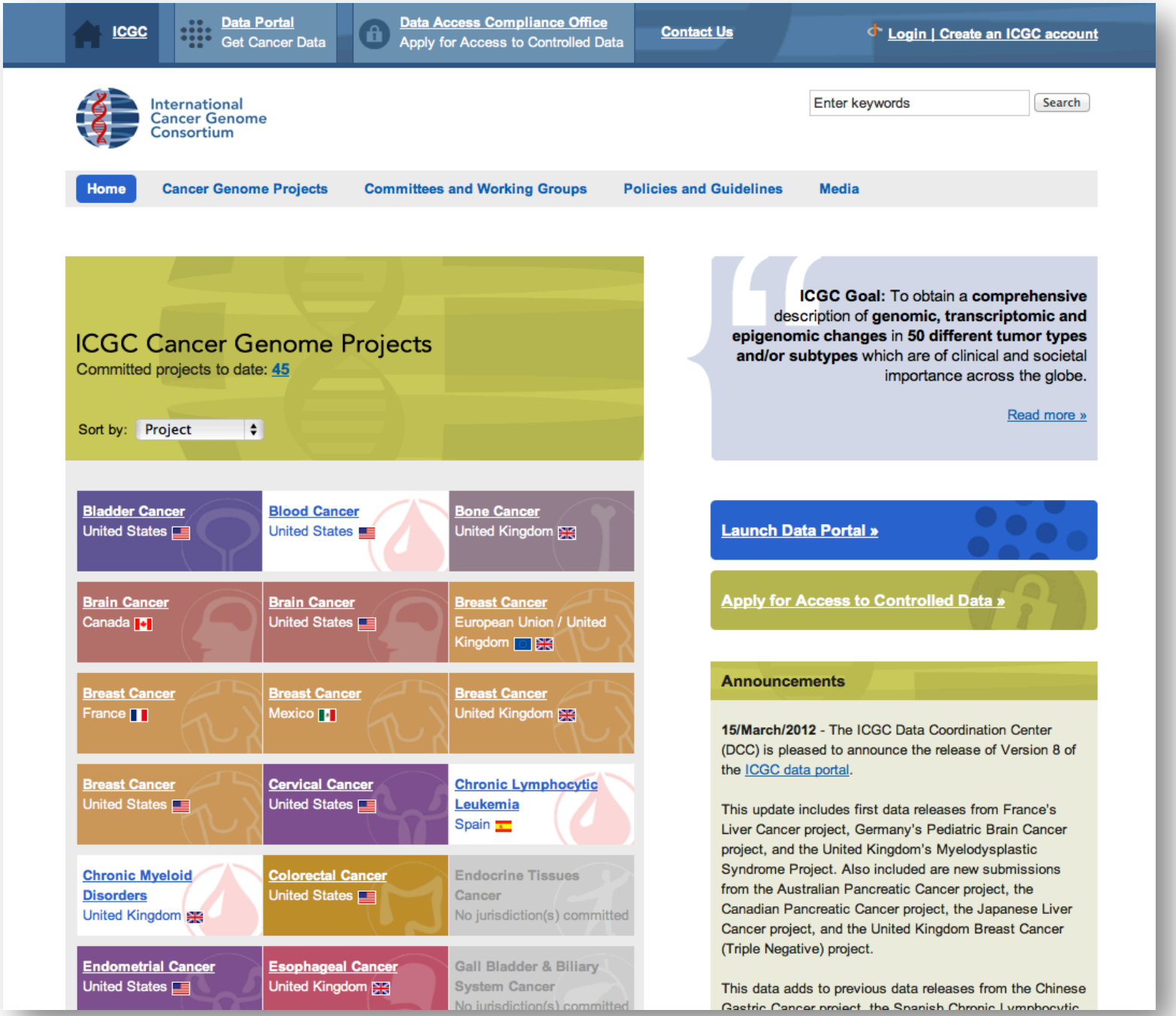
OICR
Ontario Institute for Cancer Research

Next Steps

- SNVs
  - Deep sequencing of all primaries across all genes identified in initial screen as carrying a mutant to characterize patterns of mutation.
  - Exome sequencing of remaining specimens, including xenografts & cell lines.
  - Lab is developing protocols for laser capture in order to increase sample cellularity.
- Structural Variation
  - Exhaustive benchmarking of SV calling pipelines in progress.
- Methylation
  - Lab is testing protocols for bisulphite conversion sequencing & MeDIP.
- Transcriptome
  - RNA-seq of selected cell lines under way.

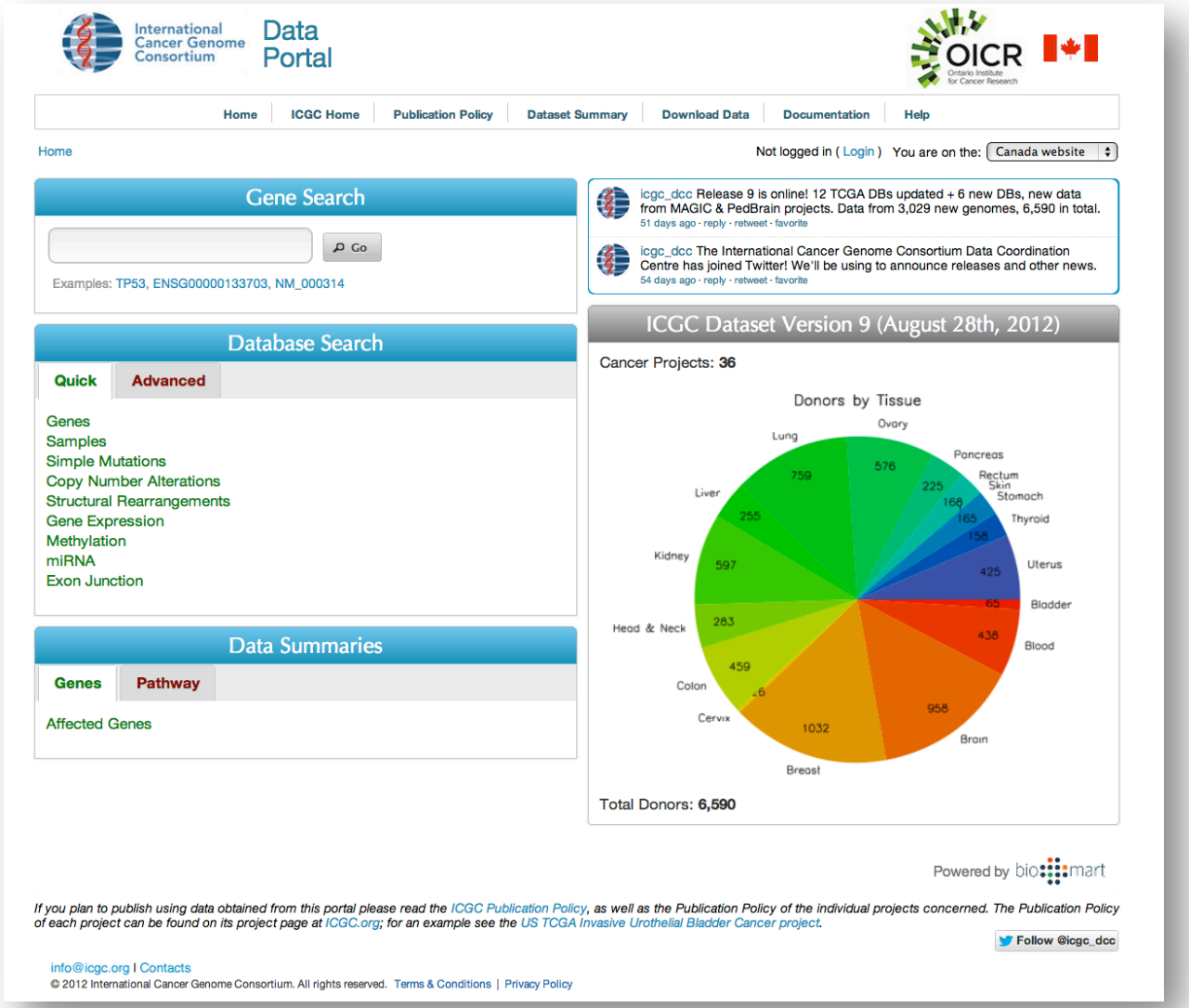
So, what next on analysis of our cancer samples?

- Doing better automation, and pipeline engineering
- We want to do more transcriptome, and integrate better with other pipelines (SNV, CNV, SV and epigenomic analyses).
- Formalizes ICGC procedures, and publish them.
- Need to consider genes that are not there (not detected, or not able to be detected), and transcriptome will help with this. Important for the network analysis.
- Also need to build models
  - That take into account low abundance and complexity of samples with low cellularity
  - That take into account the average of multiple samples (plan for 350, but will there be tumor subtypes?)
  - New project: Personal Human Proteome data


ICGC
Data Portal
Get Cancer Data
Data Access Compliance Office
Apply for Access to Controlled Data
Contact Us
Login | Create an ICGC account
International Cancer Genome Consortium
Enter keywords
Search
Home
Cancer Genome Projects
Committees and Working Groups
Policies and Guidelines
Media
ICGC Cancer Genome Projects
Committed projects to date: 45
Sort by: Project
Bladder Cancer United States
Blood Cancer United States
Bone Cancer United Kingdom
Brain Cancer Canada
Brain Cancer United States
Breast Cancer European Union / United Kingdom
Breast Cancer France
Breast Cancer Mexico
Breast Cancer United Kingdom
Breast Cancer United States
Cervical Cancer United States
Chronic Lymphocytic Leukemia Spain
Chronic Myeloid Disorders United Kingdom
Colorectal Cancer United States
Endocrine Tissues Cancer No jurisdiction(s) committed
Endometrial Cancer United States
Esophageal Cancer United Kingdom
Gall Bladder & Biliary System Cancer
ICGC Goal: To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.
Read more »
Launch Data Portal »
Apply for Access to Controlled Data »
Announcements
15/March/2012 - The ICGC Data Coordination Center (DCC) is pleased to announce the release of Version 8 of the ICGC data portal.
This update includes first data releases from France's Liver Cancer project, Germany's Pediatric Brain Cancer project, and the United Kingdom's Myelodysplastic Syndrome Project. Also included are new submissions from the Australian Pancreatic Cancer project, the Canadian Pancreatic Cancer project, the Japanese Liver Cancer project, and the United Kingdom Breast Cancer (Triple Negative) project.
This data adds to previous data releases from the Chinese

31

Data portal: http://dcc.icgc.org/



Acknowledgements

http://oicr.on.ca http://icgc.org

Project leaders at the OICR:

Tom Hudson

John McPherson

Lincoln Stein

Paul Boutros

Lakshmi Mutsawarma

Vincent Ferretti

ICGC Database Developers:

- Anthony Cros
- Jonathan Guberman
- Yong Liang
- Long Yao
- Shane Wilson
- **Zhang Junjun**
- **Brian O'Connor**

Ouellette Lab

- **Emilie Chautard**
- Michelle Brazas
- Nina Palikuca

WebDev group:

- **Joseph Yamada**
- Kamen Wu
- Miyuki Fukuma
- Salman Badr
- Stuart Lawler

Pipeline Dev. & Eval.

- Morgan Taschuk
- Peter Ruzanov
- Rob Denroche
- Zhibin Lu

ICGC DCC staff:

- **Brett Whitty**
- Marie Wong-Erasmus

Pancreatic Analysis WG

- Carson Holt
- Irina Kalatskaya
- **Christina Yung**
- Kim Begley
- Adam Wright

Sequence Informatics

- **Tim Beck**
- Tony de Bat
- Zheng Zha
- Fouad Yousif
- Xuemei Luo

SeqWare group

- **Brian O'Connor**
- Dennis Yean
- Yong Liang

Ontario

OICR Ontario Institute for Cancer Research

ICGC DCC Curation is Hiring!

- We're looking for people with a strong genomics/ bioinformatics background and experience working with large genome projects
  (with a web resource component)

Lots of data and lots of great work to do!

francis@oicr.on.ca

OICR
Ontario Institute for Cancer Research

Informatics and Biocomputing Program at the OICR


Pascale et Maya