

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.



Home Connect Discover Me Search

BF Francis Ouellette
@bfft
Bioinformatics, cancer genomics, databases & tools, OA, OpenData, Open Source, teaching; likes: outdoors, cycling, & canoeing
Toronto, Ontario, Canada. <http://oicr.on.ca/person/francis-ouellette>

Edit your profile
3,413 TWEETS
2,207 FOLLOWING
2,263 FOLLOWERS

Tweets

- Following
- Followers
- Favorites
- Lists
- Recent images
- Similar to you

Tweets

- BF Francis Ouellette** @bfft 8h
.@sangerinstitute sending a nice goodbye to their @Alexbateman1 who is going 2 also be at the @emblebi pic.twitter.com/Q2yROAPE
View photo
- BF Francis Ouellette** @bfft 12h
Last minute preps for for my @FGED talk -- I'm presenting in my home town! microarrays.ca/fged/agenda.ht... #bioinformatics
Expand
- FGED Society** @FGED 27 Sep
Agenda now posted for the October FGED meeting in Toronto: fb.me/2aMz15m88
Retweeted by BF Francis Ouellette
Expand



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

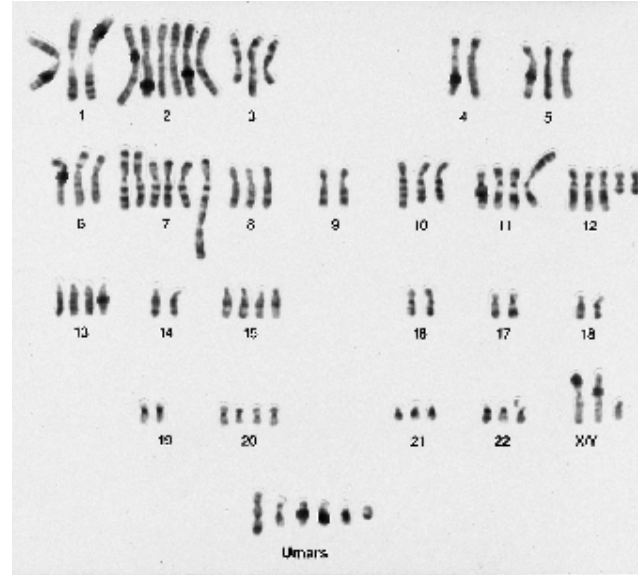
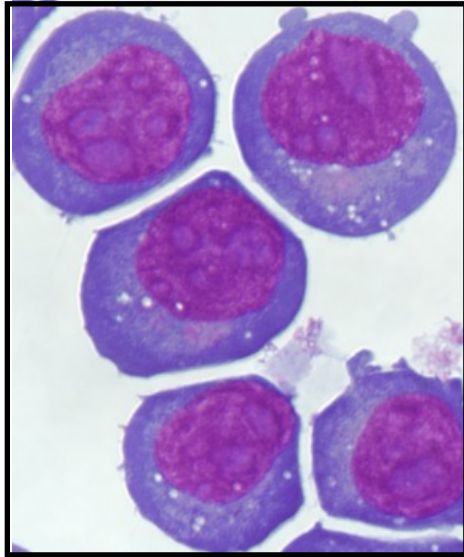
To build innovative research programs that will have an impact on the prevention, early detection, diagnosis and treatment of cancer.

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.

Cancer

A Disease of the Genome



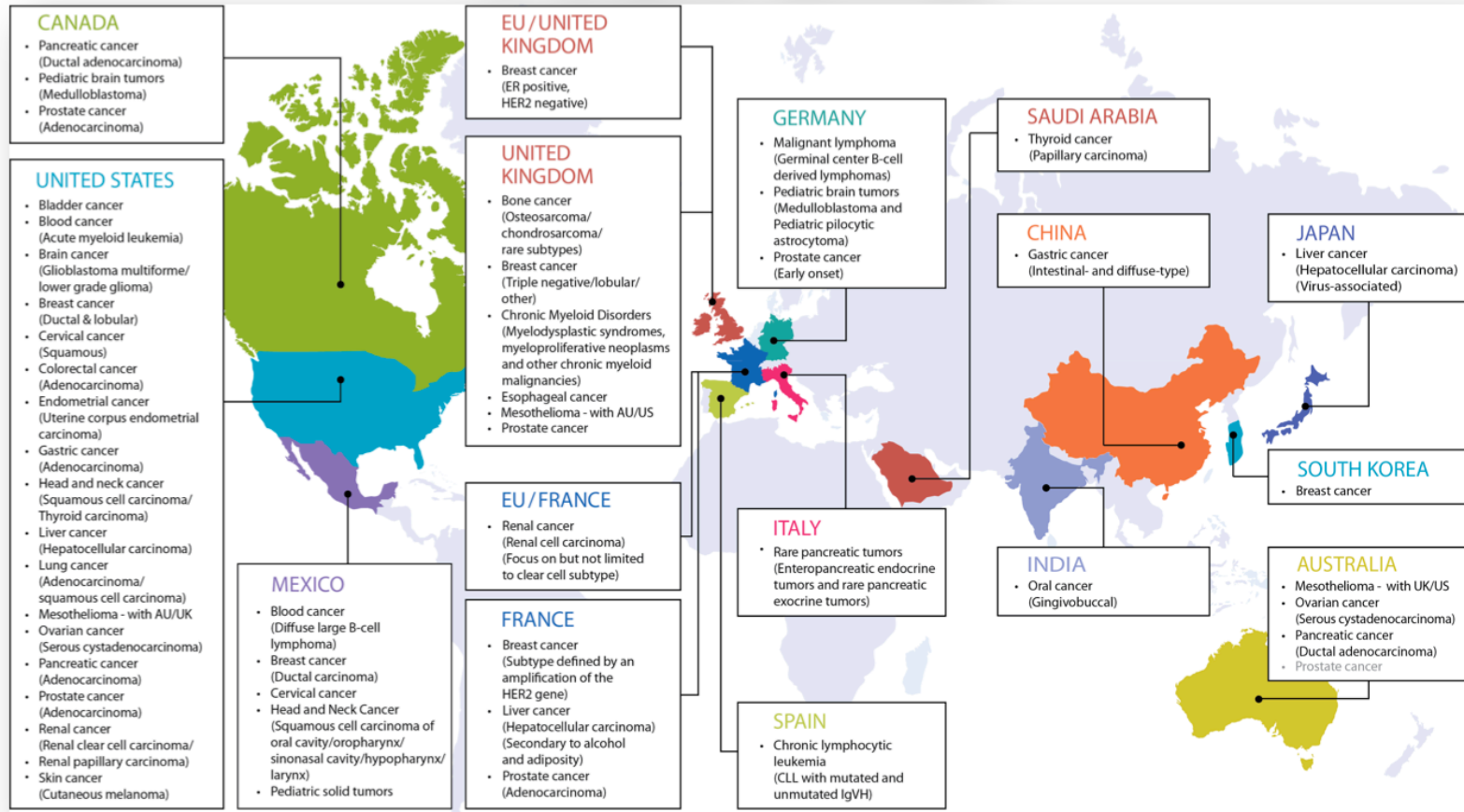
Challenge in Treating Cancer:

- Every tumor is different
- Every cancer patient is different

ICGC Map – September 2012



47 projects launched



47 Projects

12 countries

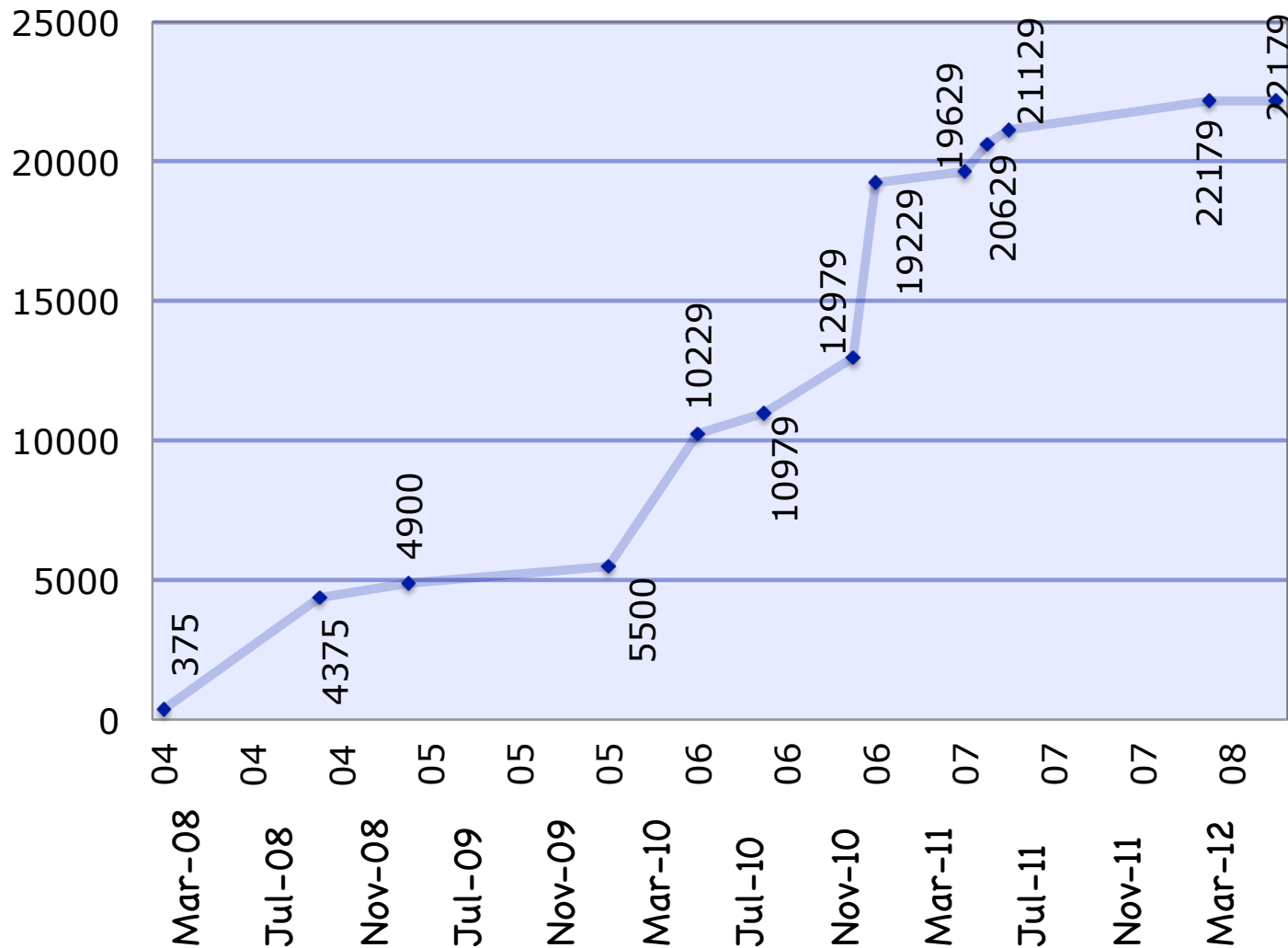
23,408 tumor samples planned

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

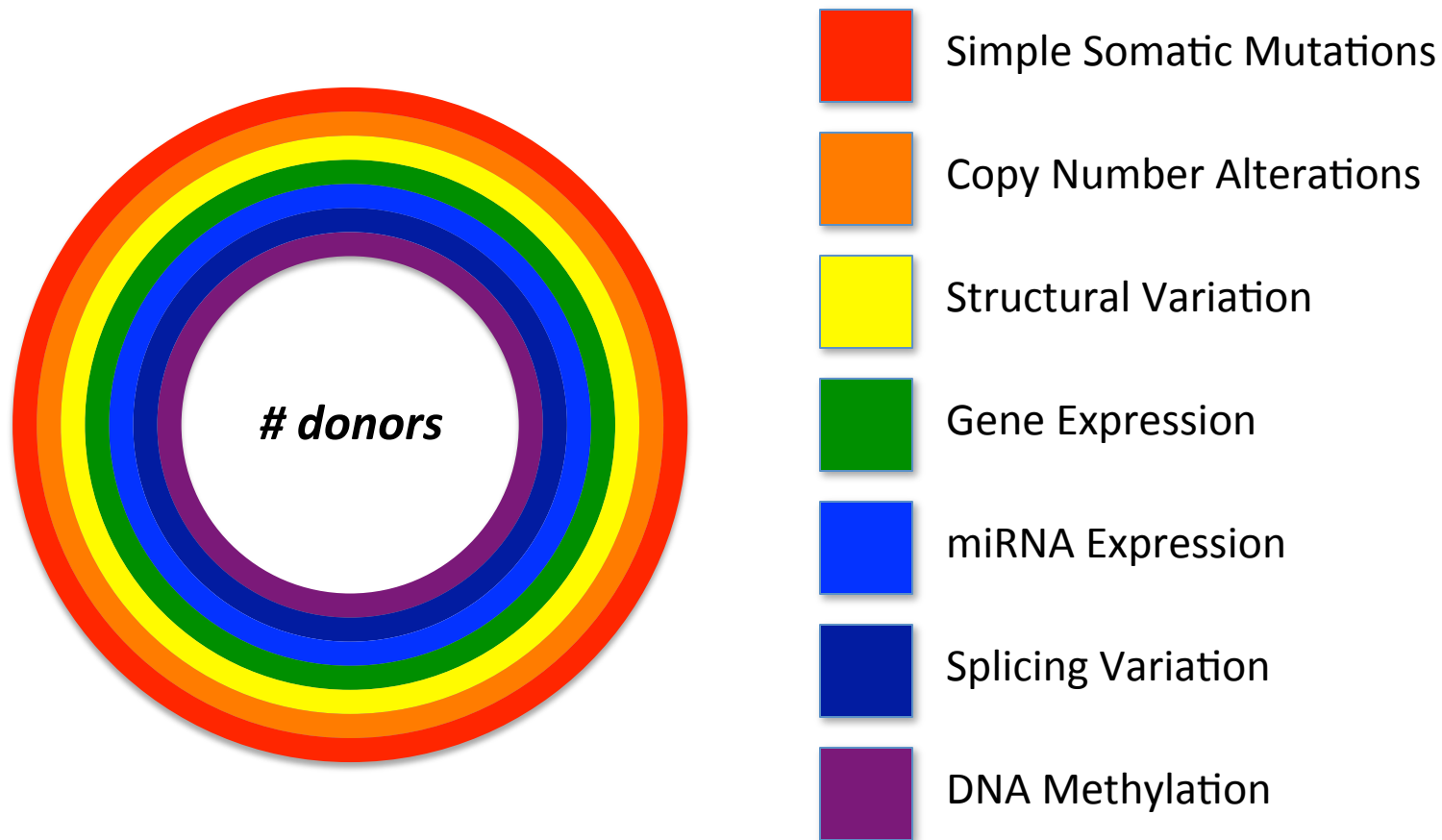
South Korea
Breast

AU/UK/US
Mesothelioma

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



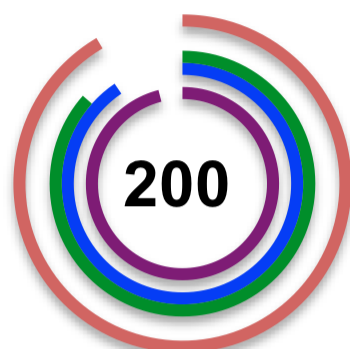
Brett Whitty



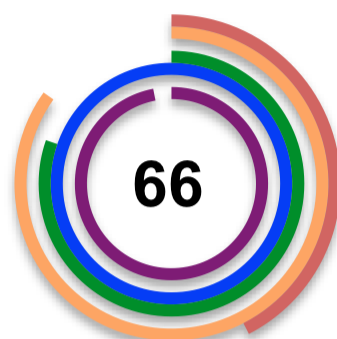
Completeness of Genomic Analysis Data Types in DCC Datasets



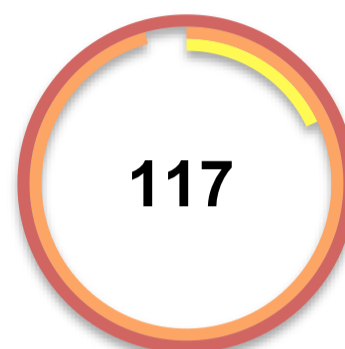
Brett Whitty



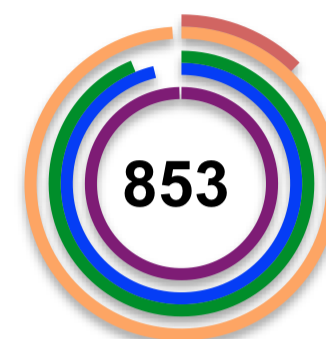
Acute Myeloid Leukemia (TCGA, US)



Bladder Urothelial Carcinoma (TCGA, US)



Breast Carcinoma (WTSI, UK)



Breast Invasive Carcinoma (TCGA, US)



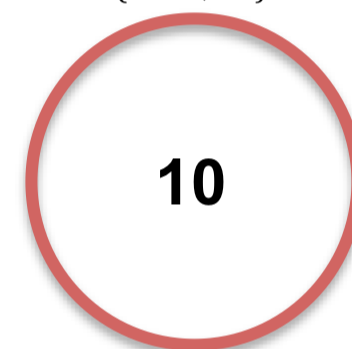
Cervical Squamous Cell Carcinoma (TCGA, US)



Chronic Lymphocytic Leukemia (ISC/MICINN, ES)



Colon Adenocarcinoma (TCGA, US)



Gastric Cancer (CCGC, CN)



Single Nucleotide 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



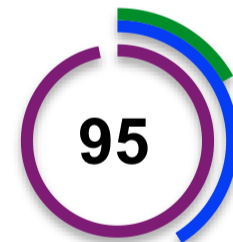
Glioblastoma Multiforme (TCGA, US)



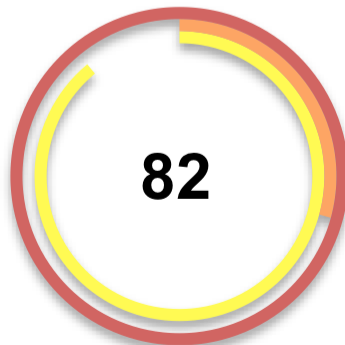
Head and Neck Squamous Cell Carcinoma (TCGA, US)



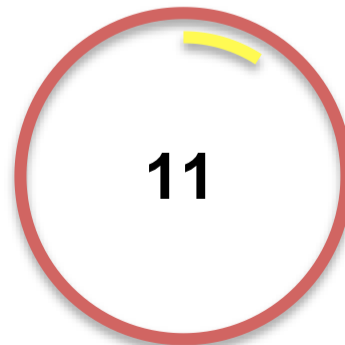
Kidney Renal Clear Cell Carcinoma (TCGA, US)



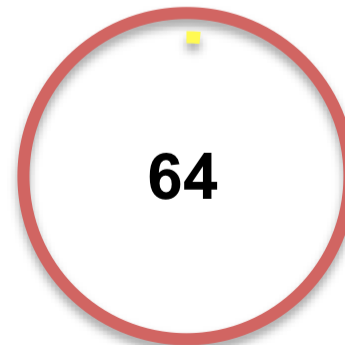
Kidney Renal Papillary Cell Carcinoma (TCGA, US)



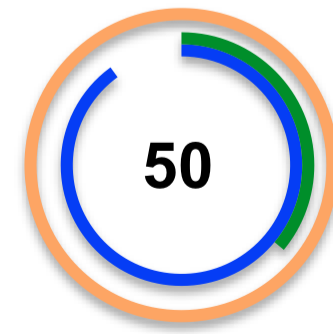
Liver Cancer (INCa, FR)



Liver Cancer (NCC, JP)



Liver Cancer (RIKEN, JP)



Liver Hepatocellular Carcinoma (TCGA, US)



Single Nucleotide Mutations

Copy Number Alterations

Structural Variation

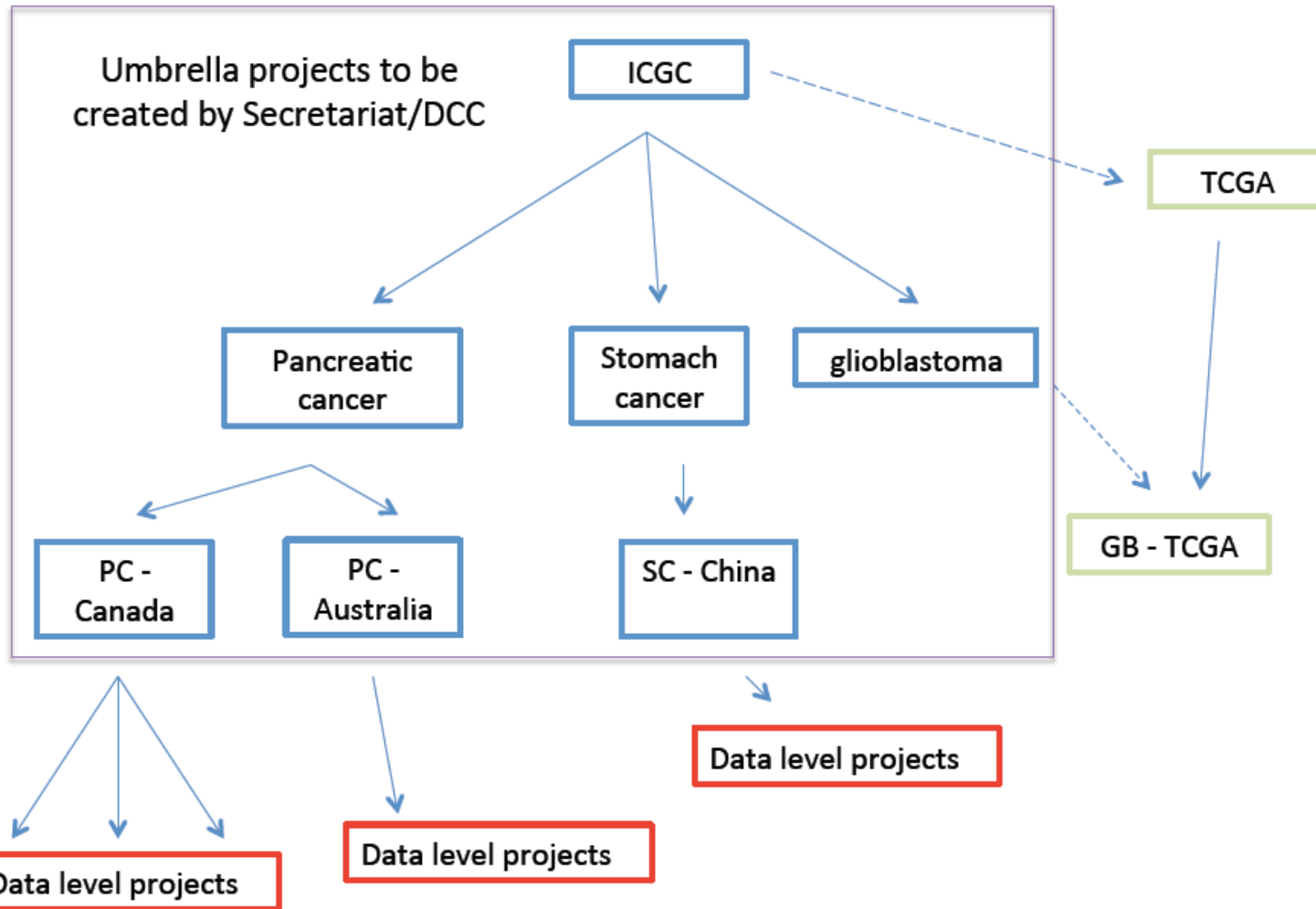
Gene Expression

miRNA Expression

Splicing Variation

DNA Methylation

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



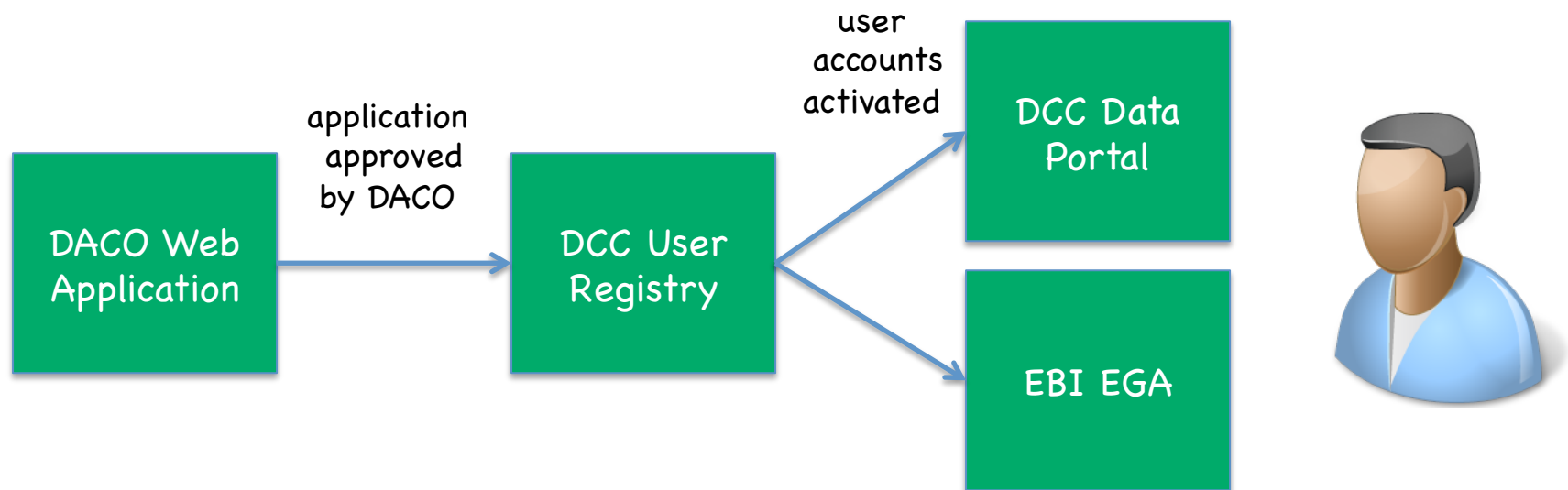
ICGC Data Categories

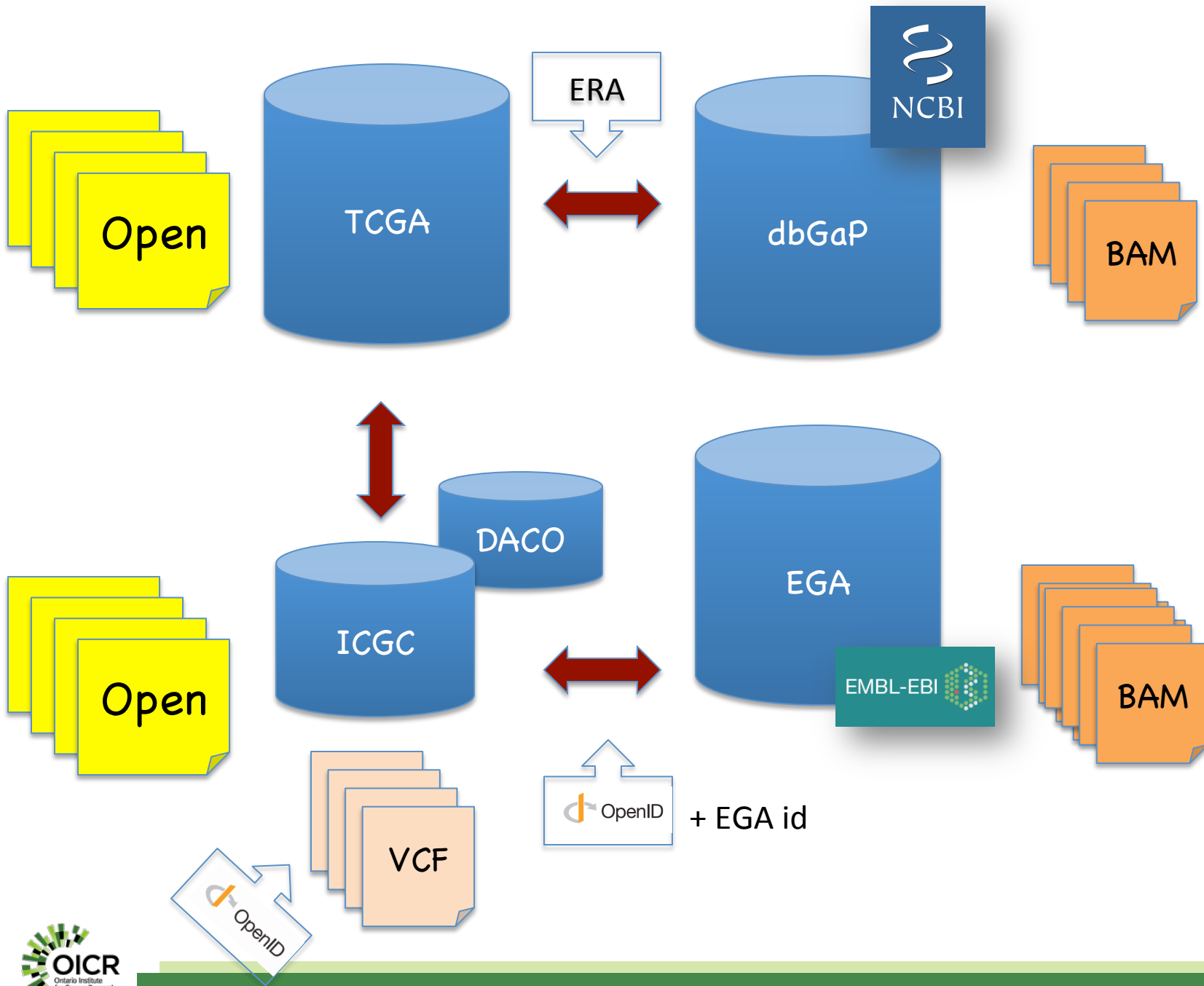
ICGC Open Access Datasets	ICGC Controlled Access Datasets
<ul style="list-style-type: none">➤ Cancer Pathology<ul style="list-style-type: none">Histologic type or subtypeHistologic nuclear grade➤ Donor<ul style="list-style-type: none">GenderAge range➤ RNA expression (normalized)➤ DNA methylation➤ Genotype frequencies➤ Somatic mutations (SNV, CNV and Structural Rearrangement)	<ul style="list-style-type: none">➤ Detailed Phenotype and Outcome Data<ul style="list-style-type: none">Patient demographyRisk factorsExaminationSurgery/Drugs/RadiationSample/SlideSpecific histological featuresProtocolAnalyte/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)

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





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*}, Ximeng Jasmine Mu¹, 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 access to important data.

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 sets; this enormous amount of freely available data has become

“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.”

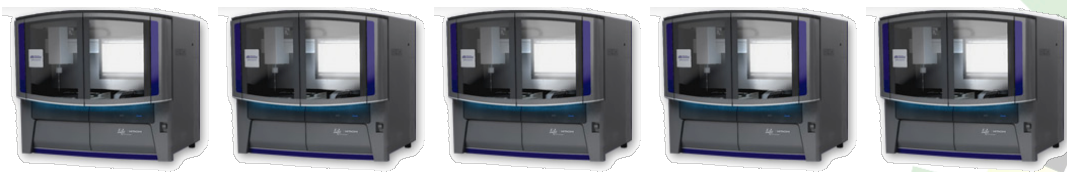
OICR Sequencing/Biocomputing Platform



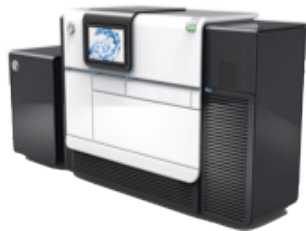
Illumina HiSeq 2000



GAII Ion Torrent MiSeq



Life Tech Solid 5500



Pac Bio

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

5500 cores
185 nodes with 16 GB RAM
221 nodes with 24 GB RAM
32 nodes with 96 GB RAM
5 nodes with 256 GB RAM
2.5PB of online storage
1Gb, 10Gb and fibre connectivity

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/

The screenshot shows the SeqWare website homepage. At the top is a navigation bar with links for HOME, NEWS, DOCUMENTATION, COMMUNITY, PARTNERS, and ABOUT. The main heading is "seqware" in a stylized font, followed by the title "Next-Generation Sequencing Analysis on the Grid and in the Cloud". Below this is a paragraph describing the project as open-source software infrastructure for NGS data analysis. Three columns are provided for "Users", "Administrators", and "Developers", each with an icon. A sidebar on the right contains a Twitter feed with three tweets about releases and documentation updates. At the bottom, a footer contains copyright information and mentions of tools like nanoc and fonts like Graublau and Gentium.

HOME NEWS DOCUMENTATION COMMUNITY PARTNERS ABOUT

seqware

Next-Generation Sequencing Analysis on the Grid and in the Cloud

The open source SeqWare project is a portable software infrastructure designed to analyze massive genomics datasets produced by contemporary and emerging technologies, in particular Next Generation Sequencing (NGS) platforms. It consists of a comprehensive suite of infrastructure tools focused on enabling the end-to-end analysis of sequence data – from raw base calling to analyzed variants ready for interpretation by users. See "[About SeqWare](#)" and our "[Introduction to SeqWare](#)" for more details...

Users **Administrators** **Developers**

SeqWare We're released 0.13.3 and for release notes see our site at buff.ly/R2uzzf
3 days ago · reply · retweet · favorite

SeqWare You can check out the Twitter archive for Genome Informatics 2012 #GI2012 at buff.ly/TepxVH
28 days ago · reply · retweet · favorite

SeqWare We've posted Brian's talk from Genome Informatics #GI2012 on how #SeqWare is used at #OICR and on the #Cloud. buff.ly/PSYqWz
30 days ago · reply · retweet · favorite

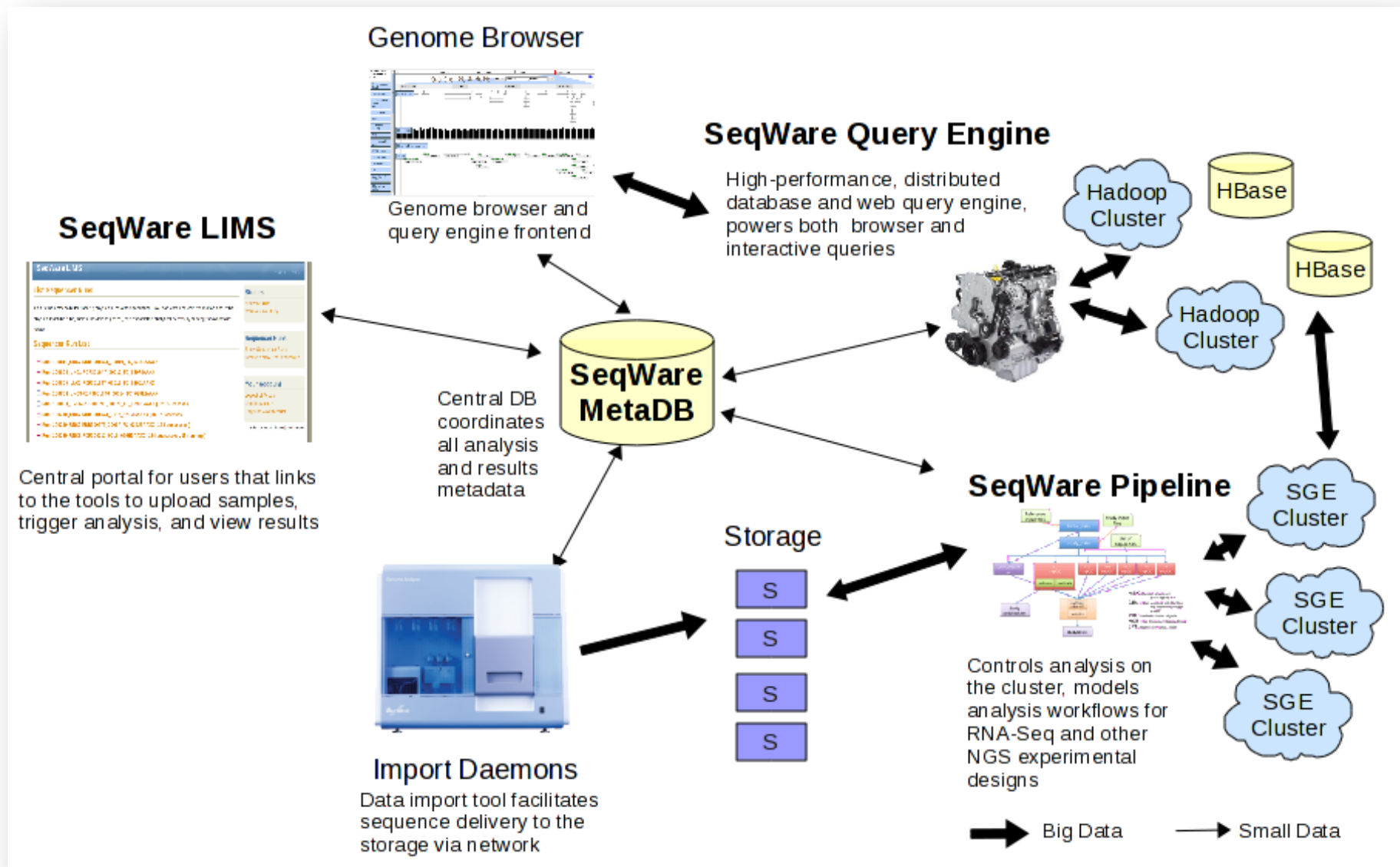
SeqWare We've added more content to our site. This includes documentation for our RESTful Web Service and the Query Engine HBase variant DB.
30 days ago · reply · retweet · favorite

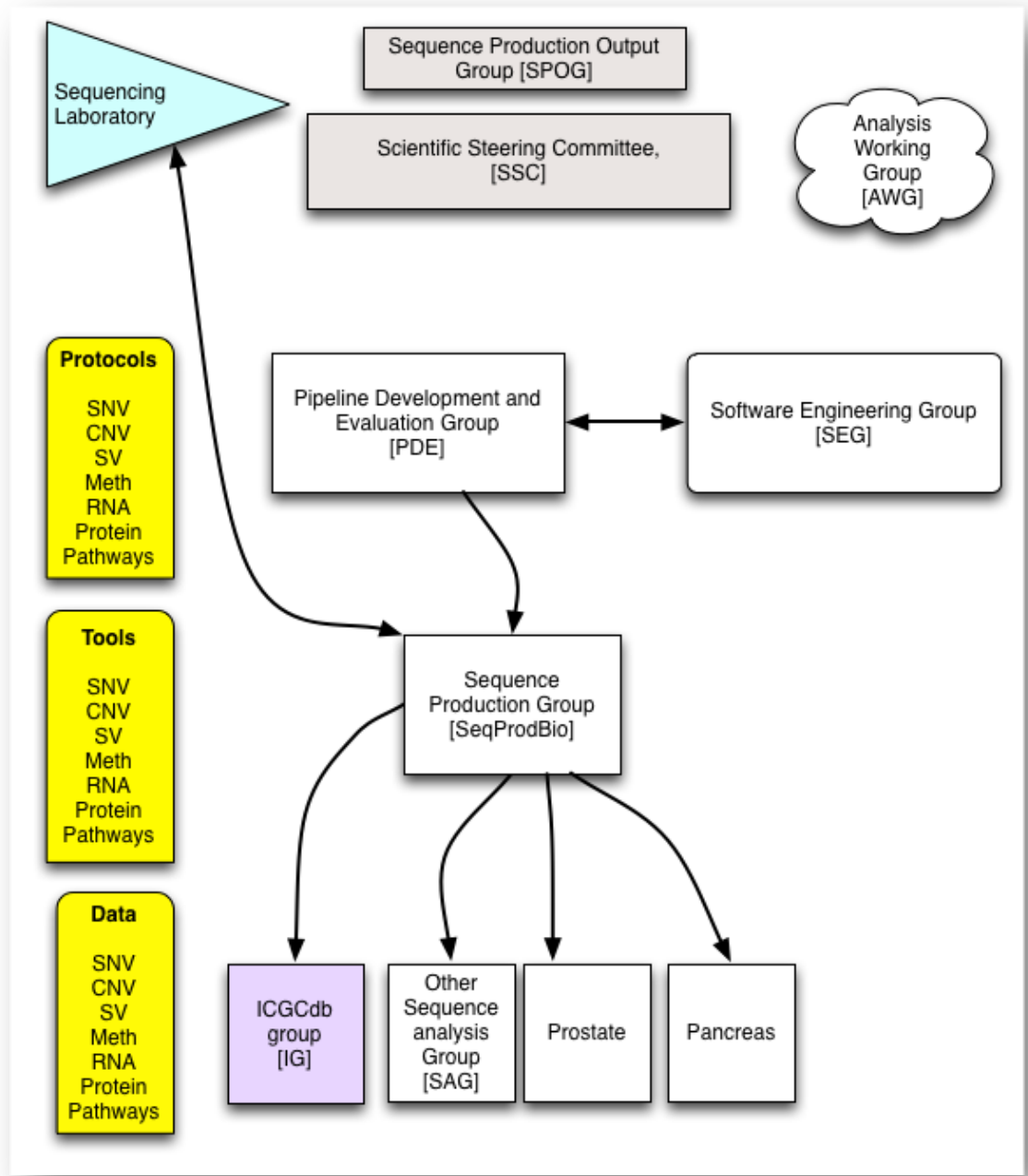
Join the conversation

The current version of SeqWare is 0.13.3, released on October 9th, 2012. See the [release notes](#) for details.

SeqWare © 2007–2012 Brian O'Connor. SeqWare is released under the a [GNU GPL v3](#). This site is built using the excellent [nanoc](#) tool and example site along with the [Graublau](#) and [Gentium](#) fonts.

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

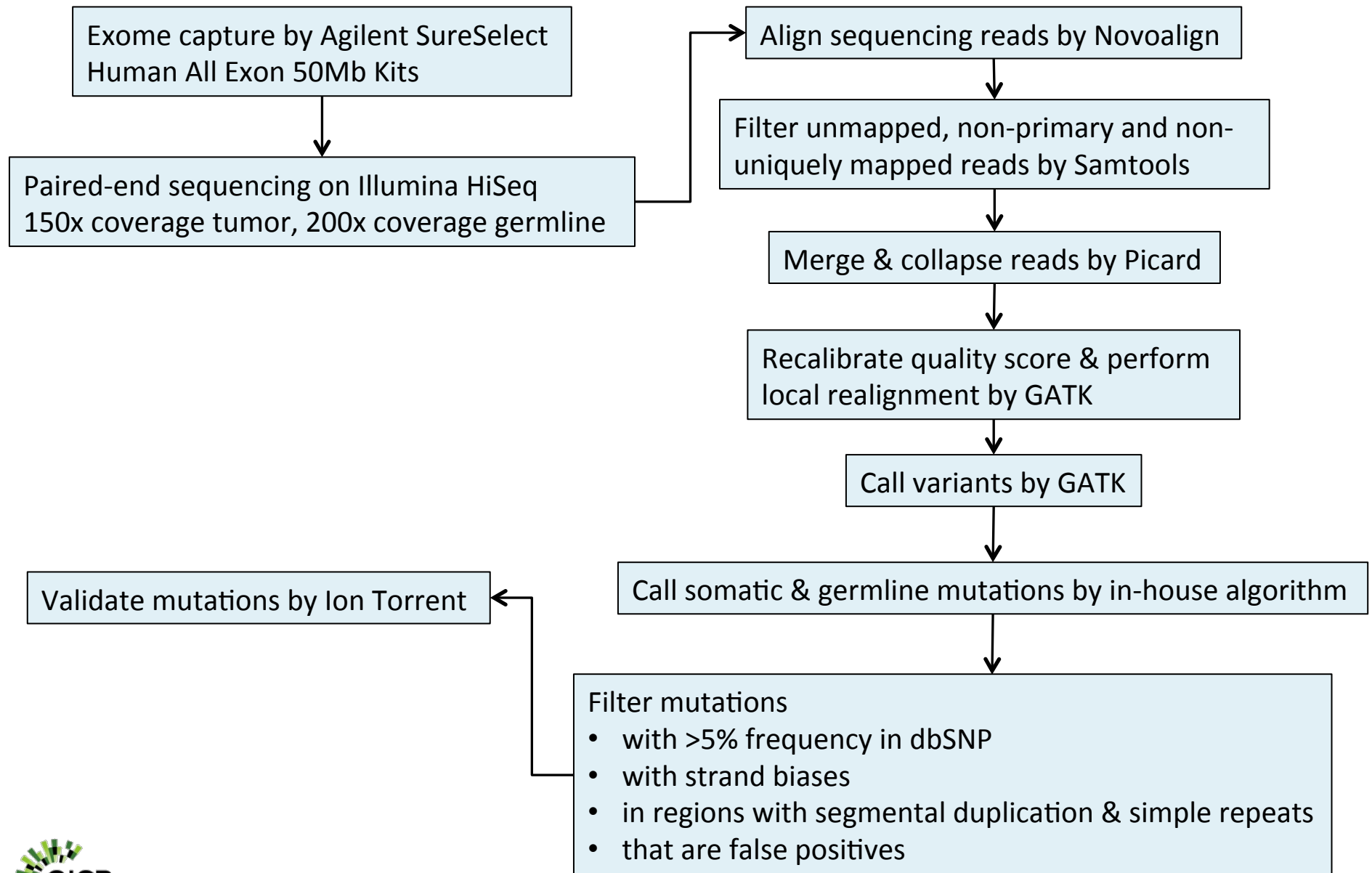




What do we do to maximize good calls?

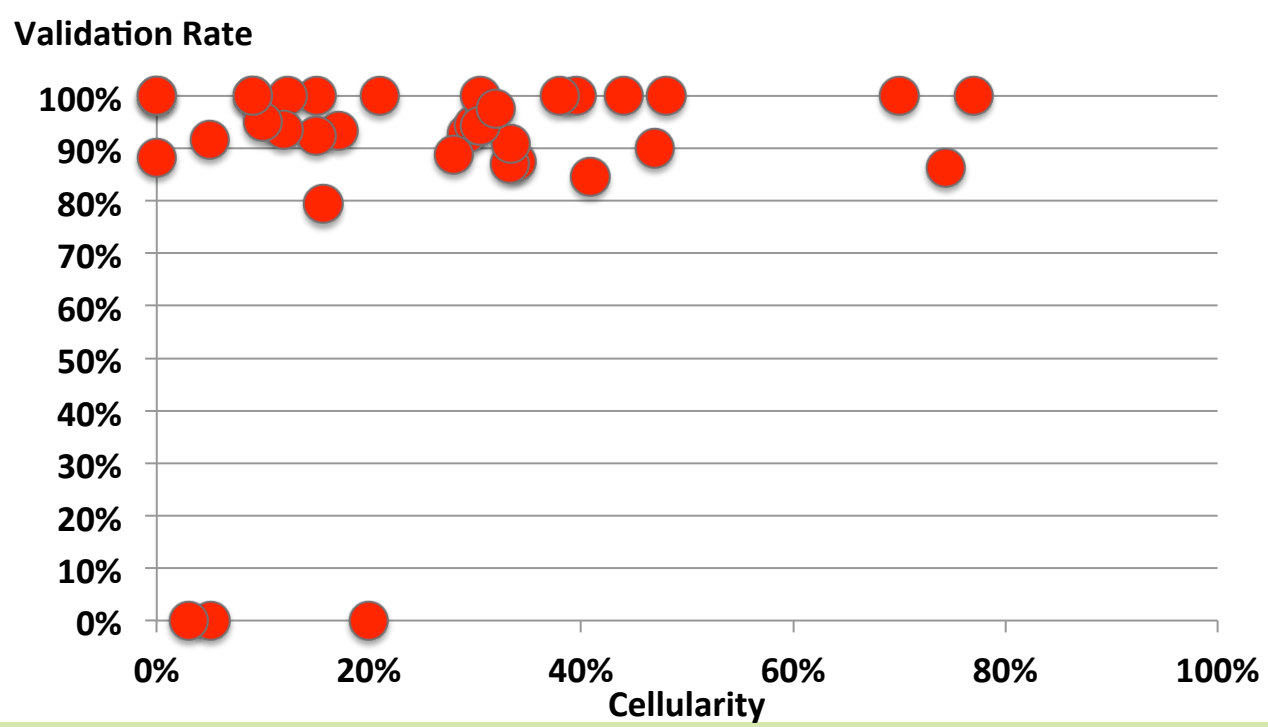
- Minimal coverage of tumor and germline for **exome**:
 - 200x germline
 - 150x tumor
- Minimum quality score
- Simultaneous alignment of reference, normal and tumor
- Blacklist “bad” regions
- Remove suspiciously dense clusters of mutations (perhaps too aggressive)
- Validate, validate, validate!
-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

Exome Sequencing Pipeline

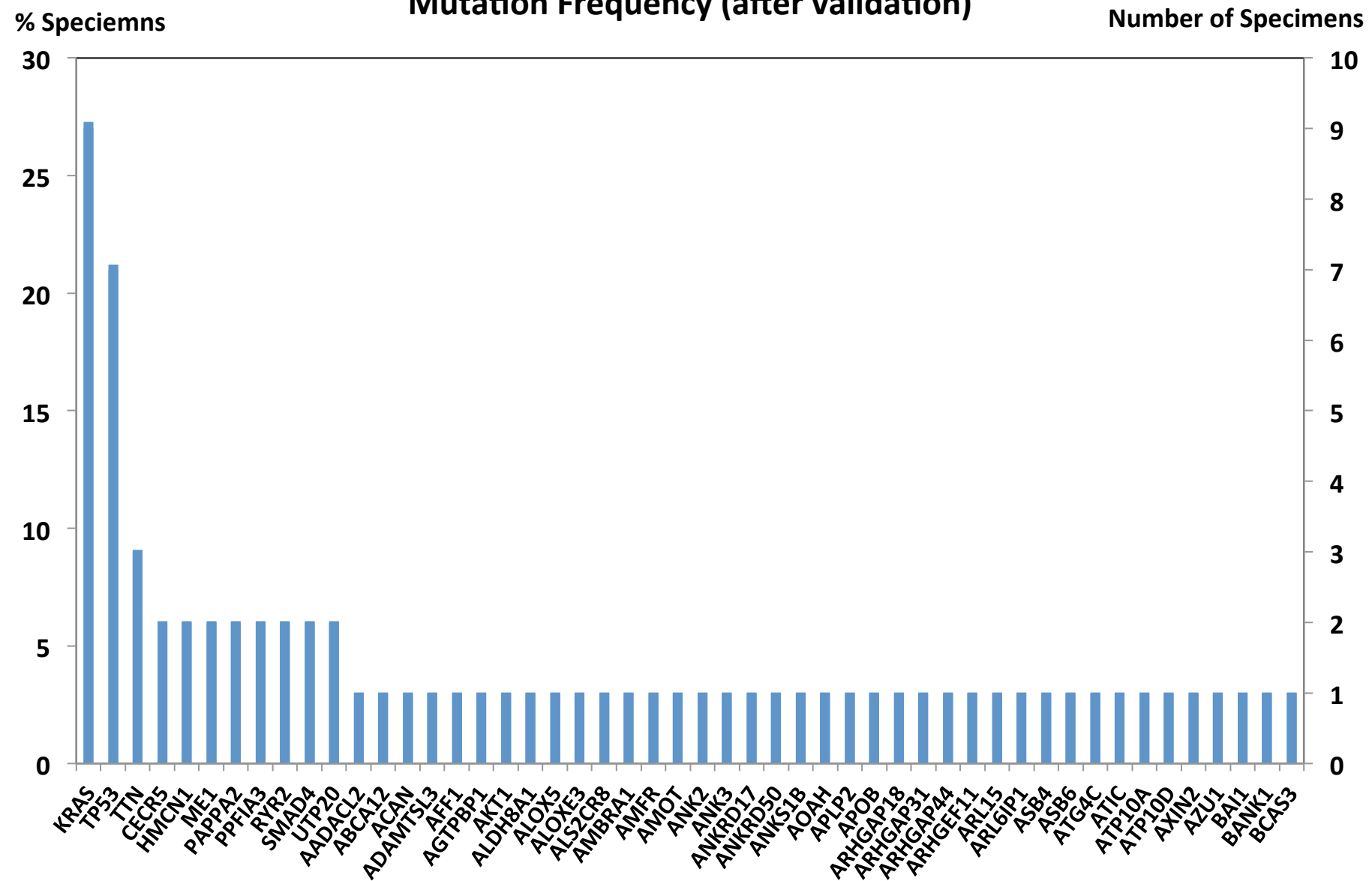


Validation Strategy

- 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

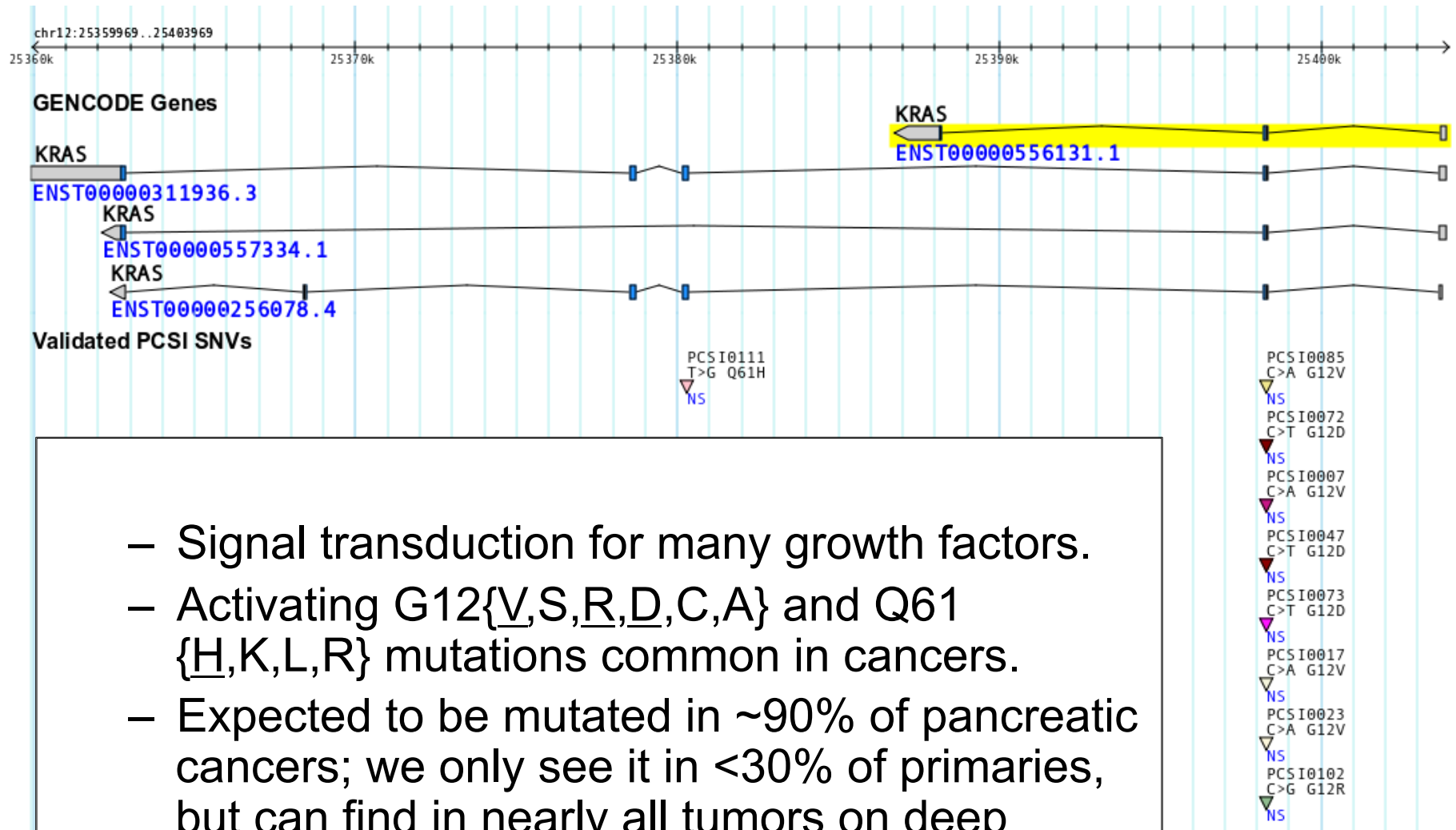


Mutation Frequency (after validation)



+ 392 genes mutated in 1 specimen

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%)

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 Cancer Genome Projects

Committed projects to date: [45](#)

Sort by:

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 No jurisdiction(s) committed

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 Gastric Cancer project, the Spanish Chronic Lymphocytic

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

The screenshot displays the ICGC Data Portal interface. At the top, it features the International Cancer Genome Consortium (ICGC) logo and the Ontario Institute for Cancer Research (OICR) logo with a Canadian flag. A navigation menu includes links for Home, ICGC Home, Publication Policy, Dataset Summary, Download Data, Documentation, and Help. Below the menu, there is a login status indicator: "Not logged in (Login) You are on the: Canada website".

The main content area is divided into several sections:

- Gene Search:** A search bar with a "Go" button and examples: TP53, ENSG00000133703, NM_000314.
- Database Search:** A section with "Quick" and "Advanced" tabs. Under "Quick", there is a list of search categories: Genes, Samples, Simple Mutations, Copy Number Alterations, Structural Rearrangements, Gene Expression, Methylation, miRNA, and Exon Junction.
- Data Summaries:** A section with "Genes" and "Pathway" tabs. Under "Genes", there is a link for "Affected Genes".
- ICGC Dataset Version 9 (August 28th, 2012):** A section containing a pie chart titled "Donors by Tissue" and the text "Cancer Projects: 36". The pie chart shows the following data:

Tissue	Number of Donors
Lung	759
Ovary	576
Pancreas	225
Rectum	166
Skin	165
Stomach	158
Thyroid	158
Uterus	425
Bladder	65
Blood	438
Brain	958
Breast	1032
Cervix	6
Colon	459
Head & Neck	283
Kidney	597
Liver	255

Total Donors: 6,590

At the bottom of the page, there is a "Powered by bio:mart" logo, a Twitter follow button for @icgc_dcc, and a footer with contact information: info@icgc.org | Contacts, © 2012 International Cancer Genome Consortium. All rights reserved. Terms & Conditions | Privacy Policy.

Acknowledgements

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

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



Informatics and Biocomputing Program at the OICR



Pascale et Maya