Open Data is Essential for Personalized Medicine

BF Francis Ouellette
VP Scientific Affairs, Génome Québec
October 17, 2017

https://goo.gl/8U1QJa
This presentation is on:

https://www.slideshare.net/
Attribution-Share Alike 2.5 Canada

You are free:

- to Share — to copy, distribute and transmit the work.
- to Remix — to adapt the work.

Under the following conditions:

- Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

- Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Your fair dealing and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full licence) available in the following languages:

- English
- French

Learn how to distribute your work using this licence.
Open Data is Essential for Genomics
Open Data is Essential for Genomics

E-mail: francis@genomequebec.com

@bffo
Times I’ve been in Italy

• Trieste 1996: Last Yeast Genome Meeting
• Naples 2005: NETTAB “Workflows management: new abilities for the biological information overflow”
• Rome 2017: Elixir
• Palermo 2017: NETTAB
Outline

• What I do
• Open Data in genomics
• Final thoughts
But first, a little about me …

… an unfinished story!
Open Data is Essential for Genomics

http://goo.gl/dJlur
Open Data is Essential for Genomics

http://goo.gl/LwVOZ
Open Data is Essential for Genomics

http://goo.gl/QI6aL
Open Data is Essential for Genomics

http://goo.gl/mYHFO
Open Data is Essential for Genomics

http://goo.gl/Jc5TK
Open Data is Essential for Genomics

1993-1997

https://goo.gl/3PFr7L
Open Data is Essential for Genomics
A Decade of Growth for GenBank and NCBI

Congressional legislation establishing NCBI was signed on November 4, 1988.

Open Data is Essential for Genomics
A Decade of Growth for GenBank and NCBI

Base Pairs of DNA (in millions)

Congressional legislation establishing NCBI was signed on November 4, 1988


BLAST Entrez dbEST GenBank at NCBI

3-D Structures Network Genomes Taxonomy

BankIt dbSTS

Human Genome LocusLink RefSeq dbSNP

PubMed PSI-BLAST Gapped BLAST COGs VAST ePCR

UniGene GeneMap OMIM Sequin Cn3D

Open Data is Essential for Genomics
Open Data is Essential for Genomics
Open Data is Essential for Genomics
Open Data is Essential for Genomics
Open Data is Essential for Genomics
Open Data is Essential for Genomics

https://www.ubc.ca/
Open Data is Essential for Genomics

Canadian Bioinformatics Workshops

The Canadian Genetic Diseases Network and the Biotechnology Human Resources Council are piloting an integrated series of practical computer-based workshops on the following dates:

- **Bioinformatics**
  - Ottawa
  - Aug 7-19, 2000

- **Genomics**
  - Vancouver
  - Jan 24-29, 2000

- **Proteomics**
  - Toronto
  - May 15-20, 2000

- **Developing the Tools**
  - Montreal
  - Jun 26-Jul 1, 2000
2001: Human Genome Project

Open Data is Essential for Genomics
Open Data is Essential for Genomics
Open Data is Essential for Genomics
Toronto
Open Data is Essential for Genomics
Open Data is Essential for Genomics
Open Data is Essential for Genomics

http://goo.gl/dJIur
Open Data is Essential for Genomics
Open Data is Essential for Genomics
SABs, EBs & projects I’m on:

- Galaxy Project
- eLiXir
- SGD
- NIH Data Commons - Platform Stack
- HMP
- H3ABioNet
- PLOS Computational Biology
- bioinformatics.ca

Open Data is Essential for Genomics
Open Data is Essential for Genomics
So what unifies all of what I’ve done?
So what unifies all of what I’ve done?

Helping scientists do science.
Open Data is Essential for Genomics

https://goo.gl/Z63Wxp
Genomics

https://goo.gl/MX84KA

Open Data is Essential for Genomics
What am I calling “Genomics”?  

All “omics”
- DNA and RNA, +Epigenomics
- Proteomics, +Protein Interactions, +Pathways
- Metabolomics
- Bioinformatics/Computational Biology
- All of the related data and metadata
  - Phenotype
  - Clinical
  - Images
- New technologies …
Biological scope?

• Anything with DNA or RNA or protein
Open Data is Essential for Genomics
Example of one of a challenge for all of us?

The integration of genomic data with deep learning and artificial intelligence.
AI, Big Data, Deep Computing

- Artificial Intelligence / Deep Learning and the Big Data Hype?

https://goo.gl/WHg36Q
What do we need for that?

https://goo.gl/JWpXj2
What do we need for that?

https://goo.gl/JWpXj2
What else?

• Data has to be FAIR
  – TO BE FINDABLE
  – TO BE ACCESSIBLE
  – TO BE INTEROPERABLE
  – TO BE RE-USABLE

• https://www.force11.org/group/fairgroup/fairprinciples
Big data examples

- Genomic sequences
- Imaging
- Population scale collected wearable data
Data Center for all in Québec?

• Health Care in Canada is governed province by province.
• Génome Québec is working with various ministries to set something that could be useful/centralized and make genomic data usable for research (controlled access).
• Needs to include clinical data
“Building a data centre is like making pancakes, you always need to throw away the 1st one”

Robert Grossman
Frederick H. Rawson Professor and the Director of the Center for Data Intensive Science (CDIS) at the University of Chicago
http://rgrossman.com/
Sharing all data types, including clinical data?

Prepublication data sharing

Rapid release of prepublication data has served the field of genomics well. Attendees at a workshop in Toronto recommend extending the practice to other biological data sets.

Open discussion of ideas and full disclosure of supporting facts are the bedrock for scientific discourse and new developments. Traditionally, published papers combine the salient ideas and the supporting facts in a single discrete ‘package’. With the advent of methods for large-scale and high-throughput data analyses, the generation and transmission of the underlying facts are often replaced by an electronic process that involves sending information to and from scientific databases. For such data-intensive projects, the standard requirement is that all relevant data must be made available on a publicly accessible website at the time of a paper’s publication.

One of the lessons from the Human Genome Project (HGP) was the recognition that making data broadly available before publication can be profoundly valuable to the scientific enterprise and lead to public benefits. This is particularly the case when there is a community of scientists from many disciplines, a group of scientists, ethicists, lawyers, journal editors and funding representatives. The goal was to reaffirm and refine, where needed, the policies related to the early release of genomic data, and to extend, if possible, similar data-release policies to other types of large biological data sets — whether from proteomics, biobanking or metabolite research.

Building on the past

By design, the Toronto meeting continued policy discussions from previous meetings, in particular the Bermuda meetings (1996, 1997 and 1998) and the 2003 Fort Lauderdale meeting, which recommended that rapid prepublication release be applied to other data sets whose primary utility was a resource for the scientific community, and also established the responsibilities of the resource producers, resource users, and the funding agencies. A similar 2008 Amsterdam meeting extended the principle of prepublication release outside the major sequencing centres and funding agencies remain unaware of the details of these policies, and so one goal of the Toronto meeting was to reaffirm the existing principles for early data release with a wider group of stakeholders.

In Toronto, attendees endorsed the value of rapid prepublication data release for large reference data sets in biology and medicine that have broad utility and agreed that prepublication data release should go beyond genomics and proteomics studies to other data sets — including chemical structure, metabolomic and RNA interference data sets, and to annotated clinical resources (cohorts, tissue banks and case-control studies). In each of these domains, there are diverse data types and study designs, ranging from the large-scale ‘community resource projects’ first identified at Fort Lauderdale (for which meeting participants endorsed prepublication data release) to investigator-defined projects (for which release of data is often a condition of funding).

https://goo.gl/ofEPEX
Authors present at the “Toronto meeting”

Writing Group Members: Ewan Birney, Thomas J. Hudson, Eric D. Green, Chris Gunter, Sean Eddy, Jane Rogers, Jennifer R. Harris and S. Dusko Ehrlich.

Additional Authors:

Open data critical to progress in Science
One example: GenBank

**GenBank** sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations.
Open data critical to progress in Science

• Without GenBank and other public sequence databases
  – There would be no BLAST
  – There would be no diagnostics DNA testing
  – There would be no understanding of the human genome (there probably would not have been a human genome to work on in the first place).
Adapted from Niko Beerenwinkel, Chris D. Greenman, Jens Lagergren

Computational Cancer Biology: An Evolutionary Perspective

Published: February 4, 2016. https://doi.org/10.1371/journal.pcbi.1004717

Open Data is Essential for Genomics
Cancer is a Disease of the Genome

Challenge in Treating Cancer:

- Every tumour is different
- Every cancer patient is different

https://www.cancer.gov/research/areas/genomics

Open Data is Essential for Genomics
Analysis Data Types

- Simple Somatic Mutations (SSM or SNV)
- Copy Number Alterations (CAN or CNV)
- Structural Variants (SV)
- Germline variants (SNPs)
- Gene Expression (micro-arrays and RNASeq)
- miRNA Expression (RNASeq)
- Epigenomics (Arrays and Methylation)
- Splicing Variation (RNASeq)
- Protein Expression (Arrays)
International Cancer Genome Consortium

- Collect ~500 tumour/normal pairs from each of 50 different major cancer types; **25,000 T/N pairs!**

- Comprehensive genome analysis of each T/N pair:
  - Genome
  - Transcriptome
  - Methylome
  - Clinical data

- Make the data available to the research community & public.

Identify genome changes

Adapted from Tom Hudson
International Cancer Genome Consortium (ICGC)

The ICGC is coordinated by OICR.
International Cancer Genome Consortium: http://icgc.org

ICGC Cancer Genome Projects
Committed projects to date: 89

Sort by: Organ System

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.

Launch Data Portal
Apply for Access to Controlled Data

Announcements
23/August/2016 - The ICGC Data Coordination Center (DCC) is pleased to announce ICGC data portal data release 22 (http://dcc.icgc.org).

Open Data is Essential for Genomics
ICGC needs to deal with different kinds of users!

• Biologists/Clinicians:
  – Web interface to processed data, providing:
    • Affected gene lists with consequences
    • Impact on pathways

• Power users:
  – Application Programming Interface (API) to get to data
  – Availability and Integration with cloud resources
ICGC Data Coordinating Centre: dcc.icgc.org
Open Data is Essential for Genomics

Data Release 25
June 8th, 2017

Donor Distribution by Primary Site

<table>
<thead>
<tr>
<th>Cancer projects</th>
<th>76</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer primary sites</td>
<td>21</td>
</tr>
<tr>
<td>Donors with molecular data in DCC</td>
<td>17,570</td>
</tr>
<tr>
<td>Total Donors</td>
<td>20,343</td>
</tr>
<tr>
<td>Simple somatic mutations</td>
<td>63,480,214</td>
</tr>
<tr>
<td>Mutated Genes</td>
<td>57,753</td>
</tr>
</tbody>
</table>

https://dcc.icgc.org/
Open Data is Essential for Genomics

https://dcc.icgc.org/icgc-in-the-cloud

The ICGB has joined forces with commercial and academic compute cloud partners to remove many of the barriers that prevent researchers from using the ICGB’s vast genomic database. The large size of the ICGB datasets means that they can take months to download and analysing them requires computing power that many research groups do not have. These partnerships will allow scientists to access and analyze ICGB datasets through multiple cloud computing platforms, enhancing collaboration and accelerating the development of new tools and treatments for cancer patients.

The Data Coordination Center has developed a unified interface for searching and accessing data across all supported clouds. Authorized users can now access the ICGB BAM, VCF and other types of files through the Repository search tool on the Portal and retrieve manifest files that allow for bulk file downloads from the cloud repositories. User authentication and authorization is achieved through the standard ICGB DACO controlled access mechanism ensuring safe access to these datasets.

ICCG Compute Cloud Partners

Amazon Web Services is a well established commercial cloud providing a highly reliable, scalable, low-cost infrastructure platform in the cloud in 190 countries around the world. ICGB datasets are currently hosted at the US East (Northern Virginia) EC2 facility. Read more...

The Cancer Genome Collaboratory is an academic compute cloud resource built by the Ontario Institute for Cancer Research and hosted at the Compute Canada facilities. This infrastructure is still under intensive development and is currently storing only a small subset of the ICGB data for beta testing. Read more...
Cloud Computing for BIG DATA Genomics

Welcome to the Cancer Genome Collaboratory, an academic compute cloud resource that allows researchers to run complex analysis operations across large ICGC cancer genome data sets.

The Collaboratory data consists of:

<table>
<thead>
<tr>
<th></th>
<th>Donors</th>
<th>Files</th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaboratory - Toronto</td>
<td>1,949</td>
<td>48,317</td>
<td>846.8 TB</td>
</tr>
<tr>
<td>PDC - Chicago</td>
<td>885</td>
<td>5,184</td>
<td>254.19 TB</td>
</tr>
<tr>
<td>Total</td>
<td>2,834</td>
<td>53,501</td>
<td>1,100.99 TB</td>
</tr>
</tbody>
</table>

What We Offer

Cloud Infrastructure

The Collaboratory hosts an OpenStack cloud with more than 2592 CPU cores and over 4.4 PB of storage, offering resources such as:

- Compute
- Storage
- Networking

Researchers Sharing Tools

The Collaboratory offers multiple ways for researchers to share their tools including VM sharing through the OpenStack Console, and Docker container sharing through the Dockstore based on a GA4GH-compliant tool descriptor.

http://www.cancercollaboratory.org/
Some challenges:

• So, we have lots of data, is it generated the same way?
Every country/group has basically been submitting:

- Simple Somatic Mutations (SSM or SNV)
- Copy Number Alterations (CAN or CNV)
- Structural Variants (SV)
- Germline variants (SNPs)
- Gene Expression (micro-arrays and RNASeq)
- miRNA Expression (RNASeq)
- Epigenomics (Arrays and Methylation)
- Splicing Variation (RNASeq)
- Protein Expression (Arrays)
Are they all using the same pipelines?

• No
Open Data is Essential for Genomics
Steering Committee of PCAWG

- Peter Campbell, Sanger Inst.
- Gady Getz, Broad
- Jan Korbel, EMBL
- Lincoln Stein, OICR
- Josh Stuart, UCSC
PanCancer Analysis of Whole Genomes (PCAWG)

• > 2,800 T/N pairs with clinical data from 20 tumour type of whole genome analysis.
• Aligned with one standard pipeline.
• Genomic Variants determined with 3 pipelines
• 17 working groups
• > 50 Papers are being written now.
Open Data is Essential for Genomics
Deliverable for PCAWG include:

- 1\textsuperscript{st} PANCANCER analysis on > 2,800 cancer tumours from a WGS perspective
- RNA, SSM, CNV, Methylation analysis & germline
- Published (executable) pipelines
  - Docker / Dockstore
  - Multiple cloud access to data
  - Multiple portal access to data
Open Data is Essential for Genomics

https://dcc.icgc.org/pcawg
Working Groups (1/2)

1. Novel somatic mutation calling methods
2. Analysis of mutations in regulatory regions
3. Integration of transcriptome and genome
4. Integration of epigenome and genome
5. Consequences of somatic mutations on pathway and network activity
6. Patterns of structural variations, signatures, genomic correlations, retrotransposons, mobile elements
7. Mutation signatures and processes
8. Germline cancer genome
Working Groups (2/2)

9 Inferring driver mutations and identifying cancer genes and pathways
10 Translating cancer genomes to the clinic
11 Evolution and heterogeneity
12 Exploratory: portals, visualization and software infrastructure
13 Molecular subtypes and classification
14 Analysis of mutations in non-coding RNA
15 Exploratory: mitochondrial
16 Exploratory: pathogens
17 Tech Technical working group
Open Data is Essential for Genomics

https://goo.gl/AMxwSU
Open Data is Essential for Genomics

https://goo.gl/AMxwSU
Open Data is Essential for Genomics

https://goo.gl/AMxwSU
Open Data is Essential for Genomics

https://goo.gl/AMxwSU
http://dockstore.org
Docker Testing Group

• Group that to ensure all container workflow work as expected.

<table>
<thead>
<tr>
<th>Workflow/Tool</th>
<th>Dockstore</th>
<th>Latest DOI</th>
<th>Version</th>
<th>Github</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>uay.io/pancanne/r/pcawg-bwa-mem-workflow</td>
<td>377</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>uay.io/pancanne/r/pcawg-dkfz-workflow</td>
<td>2376</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>uay.io/pancanne/r/pcawg-sanger-cgp-workflow</td>
<td>2162</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>uay.io/pancanne/r/pcawg_delly_workflow</td>
<td>2166</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

https://goo.gl/AMxwSU
Access to Data?

- Human Data
- Patients consented to have their DNA looked at so people could understand cancer
- Need to have a system to maximize people's gift to science.
Open Data is Essential for Genomics
Open Data is Essential for Genomics

Data Access Compliance Office

1. Create your account
   Register, to receive your account
   Learn How

2. Fill your application
   Your application, tells us who you are
   Learn How

3. Submit to DACO
   Send your application
   Learn How

Identify yourself

Fill out detail form which includes:
- Contact and Project Information
- Information Technology details and procedures for keeping data secure
- Data Access Agreement

All of these documents are put into a PDF file that you print and get your institution to sign off on your behalf
Open Data is Essential for Genomics
Open Data is Essential for Genomics
https://icgc.org/daco/approved-projects

314 groups
Open Data is Essential for Genomics
Challenge:

- Open Data and controlled access data
- Not enough eyeballs on the data
- Eyeballs on the data needed to make discoveries.

https://goo.gl/ogbWXG
Culture of Sharing Openly

• Public Funding agencies
• Consortiums
• Mentors
• Peers
• New generation (vs my old generation)
• Has to become the norm
Final thoughts …

• Access to data is essential for science
• Getting data that is FAIR is hard work
• It is essential to share the work you do if you want to be recognized, get tenure, get a job or a promotion.
• Human data is more complicated, but don’t let that get in the way!
• There is a lot of material out there, learn from it (& cite your sources)!
Last message to students and young PDFs and investigators:
Last message to students and young PDFs and investigators:

Be open so people can see how great you are!
Open Data is Essential for Genomics

Acknowledgments

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

ICGC/OICR
Project leaders:
Tom Hudson
John McPherson
Lincoln Stein
Jared Simpson
Paul Boutros
Vincent Ferretti
Francis Ouellette
Jennifer Jennings

DCC Software Developer
Vincent Ferretti
Dusan Andric
Phuong-My Do
Francois Gerthoffert
Terry Lin
Michael Moncada
Vitalii Slobodianyk
Bob Tiernay
Douglas Wong
Linda Xiang
Junjun Zhang

Research IT/Systems
David Sutton,
Bob Gibson
David Magda
Rob Naccarato
Brian Ott
Gino Yearwood

Ouellette Lab
Alysha Moncrieffe
Ann Meyer
Zhibin Lu

Web Dev
Joseph Yamada
Kaman Wu
Kim Cullion
Koji Miyauchi
Miyuki Fukuma

ICGC DCC Biocuration
Hardeep Nahal
Marc Perry

EGA
Jordi Rambla De Argila
Arcadi Navarro
Audald Iloret
Mauricio Moldes

… and all the patients and their families that are putting their hopes into our work!
ÉQUIPE DES AFFAIRES SCIENTIFIQUES

B.F. Francis Ouellette

Hélène Fournier
Diane Bouchard
Diana Iglesias
Annina Spilker

Micheline Ayoub
Joël Savard
Cristina Ciurli
Grazie