

CanDIG CHORD

Canadian Health Omics Repository, Distributed

RDM-090

David Lougheed

*Canadian Centre for Computational Genomics (C3G), Montreal Node
McGill University*

Introduction to Genomics

- Interdisciplinary field of science focusing on the structure, function, evolution, mapping, and editing of genomes. A genome is an organism's complete set of DNA, including all of its genes. (*Wikipedia*)
- A genome's structural data is obtained using a sequencing machine
- Raw, compressed data for a whole genome experiment: ~100GB (very rough estimation), before downstream analysis begins!
- Decreasing cost means:
 - More experiments being done
 - New practical applications such as personalized medicine; new contexts (hospitals)
 - More sequencing occurs, and more data is generated

Genomics Metadata and Downstream Analysis

Genomics experiments don't just yield raw genomic data!

We also need to keep track of:

- Genomic metadata
 - Donor age, sex, disease
 - Sample data: tissue type, cell type, sampling time
 - Experiment descriptions
 - Processing tools used
 - Many additional fields
- Downstream analysis results

Genomics aren't the only “-omics” data!

Other types include:

- Transcriptomics: RNA instead of DNA
- Epigenomics: Epigenetic markers on the genome

These fields are even newer than genomics, and are experiencing similarly rapid growth in terms of data output.

Problem

How do we share genomic (and other -omic) datasets with others?

- Sharing data isn't as simple as uploading it to a public server!
 - Genomes are inherently unique to an individual
 - Many datasets are sensitive, and access requests must be reviewed first
 - Some data may not be able to cross jurisdictional (i.e. provincial) borders
 - Data are in myriad formats and generally de-normalized - can take years (!) to access, download, and re-process datasets that theoretically have the same types of information
 - Want persistent, citeable identifiers for these datasets
- Siloing data into individual lab or institution servers with their own data access procedures is bad for the scientific community

The CanDIG Project

- A Canadian approach to analysis of health research data
 - National scale populations
 - Provincial / institutional stewardship: local control of data and user access
- Funded four-year CFI Cyberinfrastructure project
- **Fully decentralized, distributed, and federated**
- Three currently-online sites, which authenticate their own users:
 - BC Genome Sciences Centre (BCGSC)
 - University Health Network (UHN)
 - McGill University
- **Analysis approach:** move analysis to the data
 - Access can be controlled in a fine-grained manner
 - Provide access to resulting data via APIs

About the CanDIG CHORD Project

“CHORD is a project to build a federated, Canadian, national data service for privacy-sensitive genomic and related health data.”

- “Democratize” the technology powering CanDIG
- Federated, microservices-based approach
- Data sharing, discovery, and access services for the broader Canadian health research community, built on the existing GenAP platform
 - Standard data normalization and quality control pipelines
 - Allow the use of existing resource allocations (Compute Canada) to power analysis
 - Support novel -omic data types: RNA expression data
 - Dataset sharing on a network of nodes with standardized vocabularies for dataset use conditions and consent metadata

Who's Working On It?

At the Canadian Centre for Computational Genomics (C3G) Montréal node (McGill):

- **Myself** - Primary developer of CHORD services and architecture since last June
- **Ksenia Zaytseva** - Developer of our clinical/phenotypic metadata service
- **Simon Chénard** - Just started; working on data repository / file object service (based on GA4GH DRS standard)

- **Dr. Guillaume Bourque** - PI, McGill
- **David Bujold** - Tech. Lead, McGill
- **Pierre-Olivier Quirion** (McGill)

Other members of the CanDIG technical team:

- Jonathan Dursi - Tech. Lead, SickKids
- Zoltan Bozoky (BCGSC)
- Richard de Borja (UHN)
- Carol Gauthier (UdeS)
- Jimmy Li (BCGSC)
- Dashaylan Naidoo (BCGSC)
- Amanjeev Sethi (SickKids)
- Shaikh Farhan Rashid (SickKids)

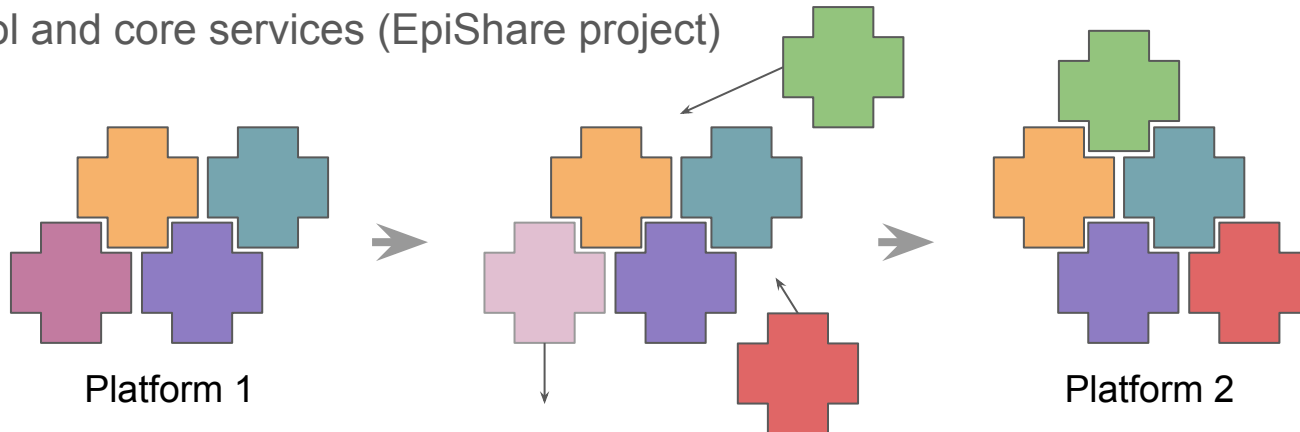
Other collaborators:

- Romain Grégoire (McGill)
- David Anderson (McGill)
- Michel Barrette (UdeS)

Why Microservices?

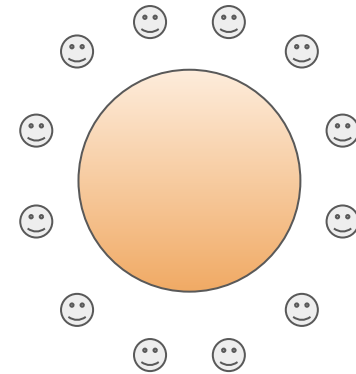
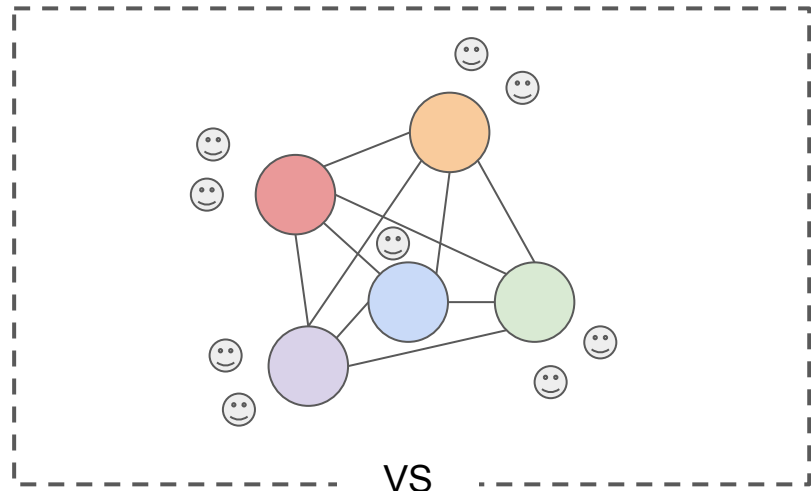
Microservices are small web applications that come together to power a much larger platform – **composable blocks** that:

- Enable **code re-use** between the main CanDIG platform and CanDIG CHORD while keeping project-specific details independent
- Allow **new platform creation** while sharing an underlying microservice communication protocol and core services (EpiShare project)



Why Federated?

- Enables users to maintain fine-grained control over their own data while participating in data sharing
 - All data can pass through the same normalization/QC pipelines
- Resource allocations from e.g. Compute Canada can be used at a node level
- Some data may not be able to be stored in another jurisdiction - maintain locality while optionally participating in network



(these people shouldn't be smiling)

The Current Platform

Currently self-hosted, but will be available as a zero-configuration app on GenAP in a few months.

- 8 microservices
- Clinical/phenotypic data support
- Genomic variant data support

The screenshot displays the CHORD Dashboard interface. At the top, there is a navigation bar with the following items: CHORD, Dashboard (active), Data Discovery, Data Manager, Peers, Notifications (3), and a user profile for dlogheed. The main content area is titled "Dashboard Node status and health monitor".

Overview

Node URL: <http://1.chord.dlogheed.com/> | Projects: 2 | Datasets: 3 | Network Size: 3

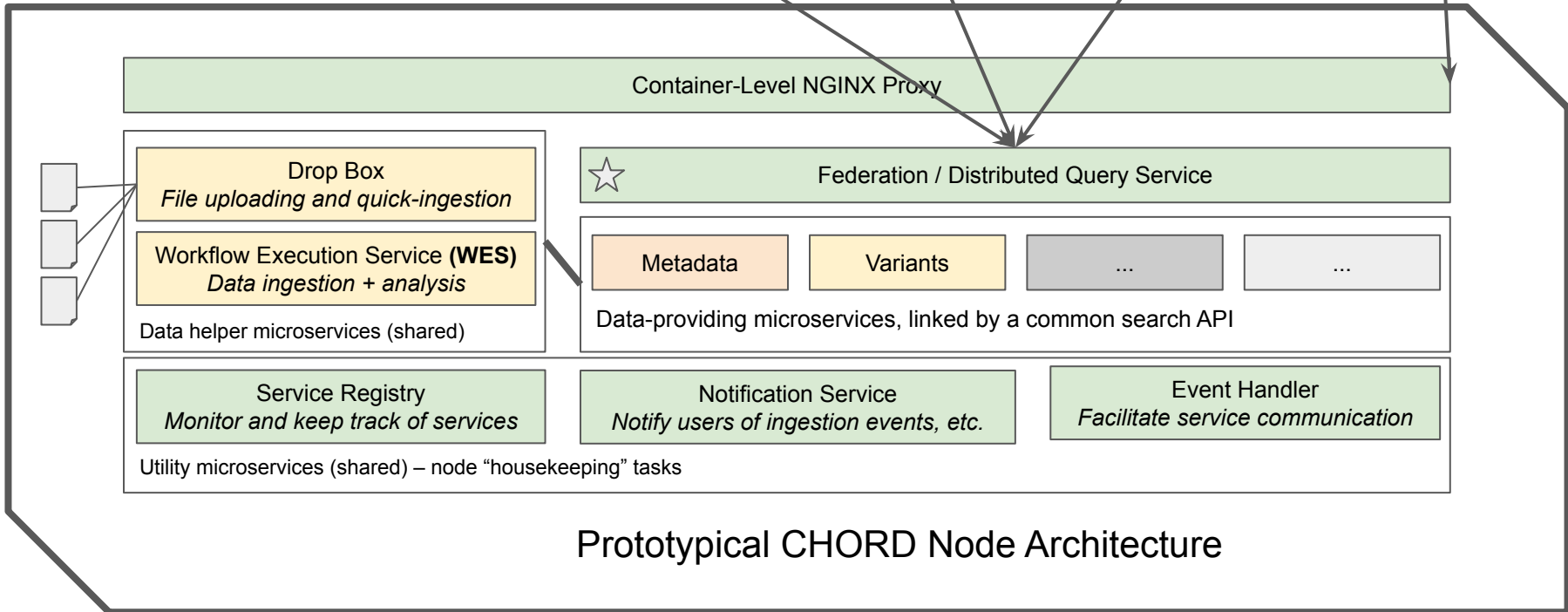
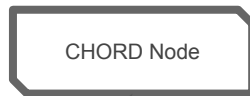
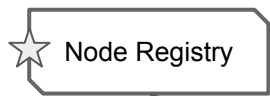
Services

Artifact	Name	Version	URL	Data Service?	Status
service-registry	CHORD Service Registry	0.1.1	http://1.chord.dlogheed.com/api/service-registry	X	HEALTHY
drop-box	CHORD Drop Box Service	0.1.0	http://1.chord.dlogheed.com/api/drop-box	X	HEALTHY
wes	CHORD WES	0.1.0	http://1.chord.dlogheed.com/api/wes	X	HEALTHY
federation	CHORD Federation	0.1.0	http://1.chord.dlogheed.com/api/federation	X	HEALTHY
notification	CHORD Notification Service	0.1.0	http://1.chord.dlogheed.com/api/notification	X	HEALTHY
event-relay	CHORD Event Relay	0.1.0	http://1.chord.dlogheed.com/api/event-relay	X	HEALTHY
metadata	Metadata Service	0.2.0	http://1.chord.dlogheed.com/api/metadata	✓	HEALTHY
variant	CHORD Variant Service	0.1.0	http://1.chord.dlogheed.com/api/variant	✓	HEALTHY

< 1 >

Copyright © 2019-2020 the Canadian Centre for Computational Genomics. chord_web is licensed under the LGPLv3. The source code is available on GitHub.

Platform Diagram



Prototypical CHORD Node Architecture

Data Discovery: Standard Search Across the Network

Standardized schemas and normalization make discovery straightforward

Data Discovery Federated data exploration

Data Type Queries

[Help](#) [Explore Data Types](#) [Add Conditions on Data Type](#)

phenopacket × variant ×

Condition 1: assembly_id GRCh37
Reference genome assembly ID.

Condition 2: chromosome 17
Reference genome chromosome identifier (e.g. 17 or X)

Condition 3: start = 41243509
1-indexed start location of the variant on the chromosome.

+ Add condition


Search

Data Access: Access Requests & Downloading

Results will (in the final release) be requestable and subsequently (post-authorization) downloadable for analysis

Results

▼ Search 1: 1 result

Node	Dataset ID ↕	Title ↕	Contact Information	Actions
	256c56b8-90dd-4843-8199-92451e46a9aa	1000 Genomes Part 1	David Lougheed david.lougheed@mail.mcgill.ca	Data Use & Consent

< 1 >

Data Management: Ingestion, QC, and Sharing

Data Manager Share data with the CHORD federation

[Projects and Datasets](#) [Files](#) [Ingestion](#) [Workflows](#) [Workflow Runs](#)

2 project 2

Node 1 Project [Edit](#) [Delete](#)

Node 1 Project

First third of the 1000 genomes

Datasets [+ Add Dataset](#)

1000 Genomes Part 1 [Ingest Metadata](#) [Edit](#) [Delete](#)

[Overview](#) [Individuals and Pools](#) [Data Tables](#) [Linked Field Sets](#) [Consent Codes and Data Use](#)

Description

This is a test dataset

Contact Information

David Loughheed
david.loughheed@mail.mcgill.ca

Created	Tables
1/29/2020, 9:07:14 AM	2

[+ Create Project](#)

Towards an Online Demo

We should have an online demo up in the next week or so with open data from the 1000 Genomes project.

Development can be followed on GitHub:

<https://github.com/c3g>

Future Plans

- Data access management and data download tools
 - Grant dataset access to other users in the CHORD federation
- Hosted version for researchers on the GenAP platform
 - Authenticate with Compute Canada credentials
 - Take advantage of CC resource allocations
 - Create a natural CHORD discovery and data-sharing network among GenAP CHORD nodes
- Federated analysis and censored results aggregation
- RNA data support
- Epigenomic data support (EpiShare project)
- A more extensive citable unique identification system for datasets

Thank you

Bonus: Demo Video

