The Terry Fox Research Institute
L'Institut de recherche Terry Fox

CanDIG - National Genomics Analyses over Locally-Controlled Private Data

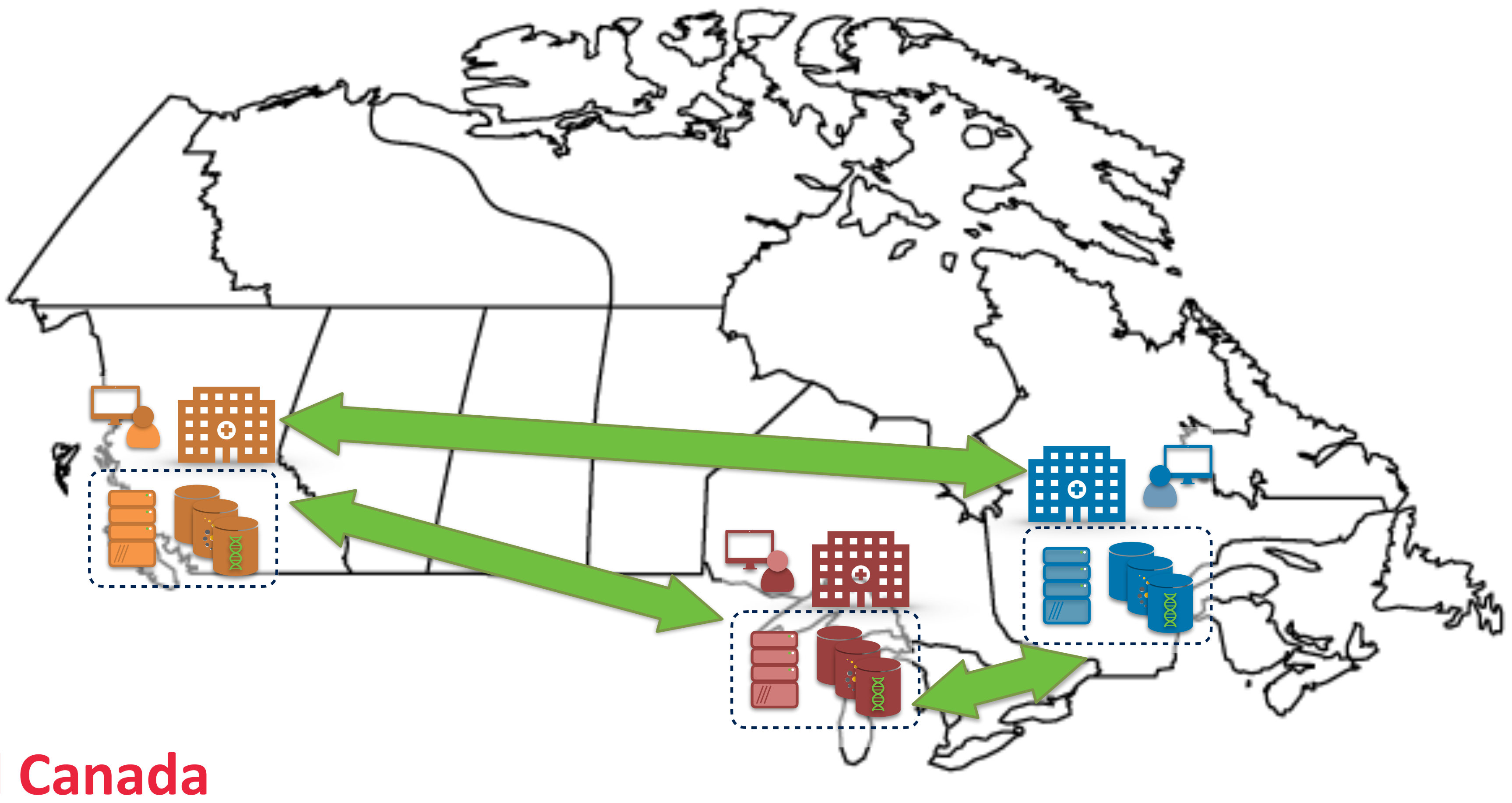
LJ Dursi[1], R deBorja[2], Z Bozoky[3], N Memon[3], Z Lu[2], K Zhu[2], J Foong[1], A Ponomarev[3], D Dupont[2], J Kim[2], C Gauthier[4], D Bujold[5], PE Jacques[6], Y Joly[7], TJ Pugh[2], G Bourque[5], SJM Jones[3], C Virtanen[2], M BRUDNO[1]

[1]Centre for Computational Medicine, The Hospital for Sick Children, [2]University Health Network, Princess Margaret Cancer Centre, [3]Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, [4]Centre de calcul scientifique, Université de Sherbrooke, [5]McGill University and Genome Quebec Innovation Centre, [6]Departement de biologie, Université de Sherbrooke, [7]Centre of Genomics and Policy, McGill University

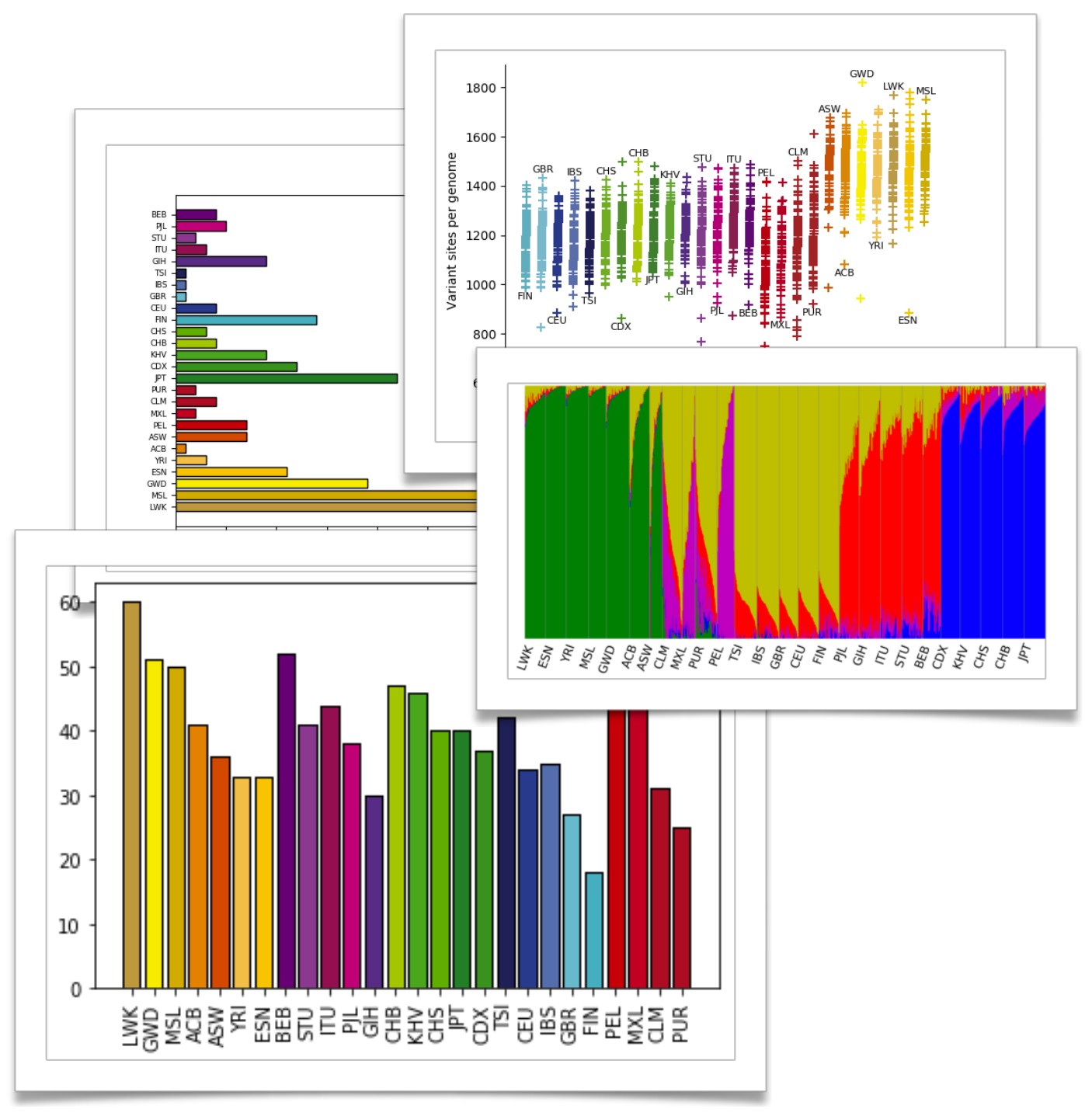
National Architecture Driving International Standards

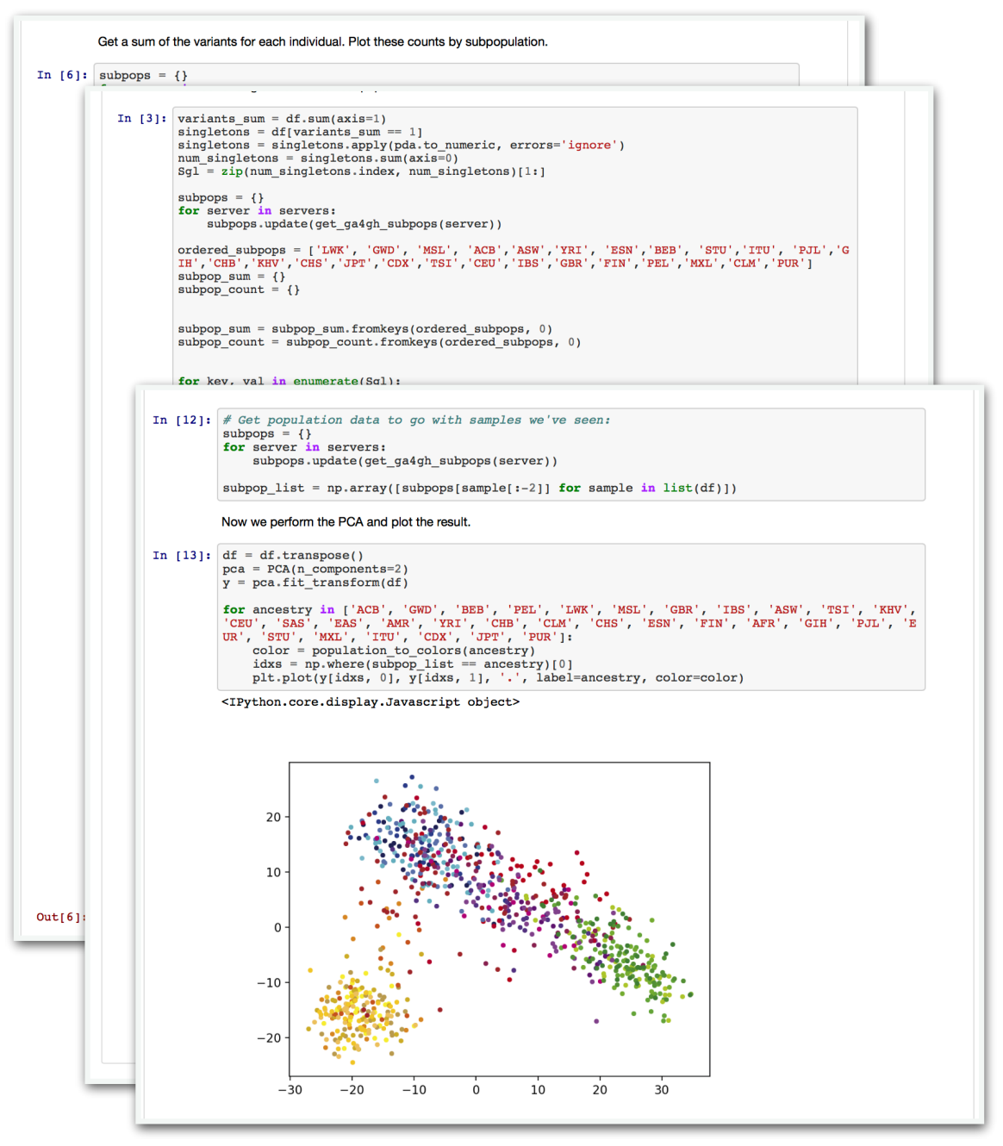
The Canadian Distributed Infrastructure for Genomics is connecting research sites across Canada, allowing national-scale, privacy-maintaining analyses of locally-controlled genomic data sets by sending computation to the data. With privacy protections and local control built in from the beginning, we make it easier for health data stewards to allow their data to be part of specific remote analyses.

CanDIG is a driver project for the international Global Alliance for Global Health (GA4GH) genomic data-sharing standardization effort, developing and using standards for data access through APIs to provide interoperability with a wide range of tools; and it uses web-era authentication and authorization standards (OpenID Connect and UMA) to ensure privacy and security of all data.



Federated Genomics for a Federal Canada





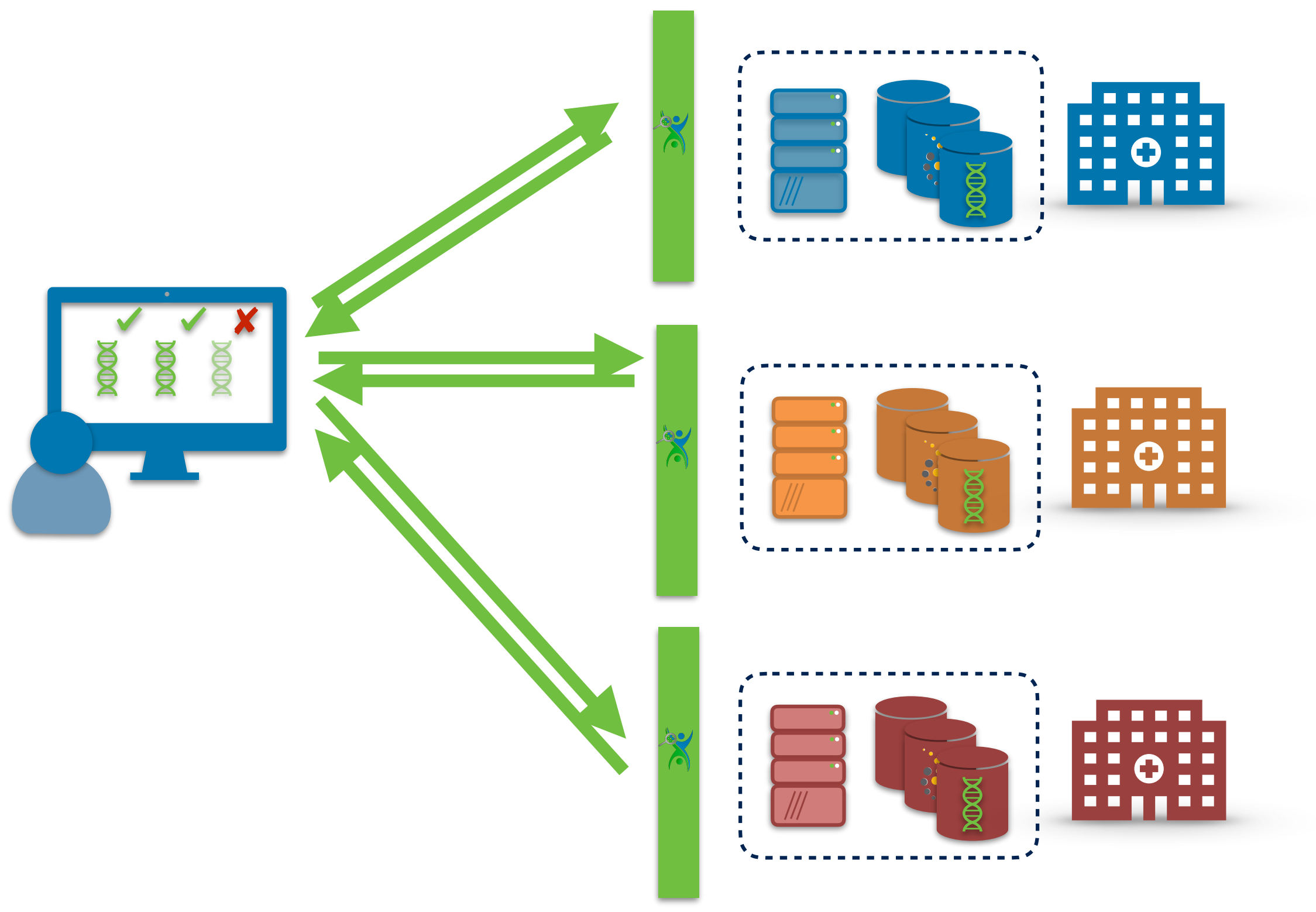
To be useful to researchers, accessing remote data has to be convenient, fast, and integrated into familiar tools. By building on existing and emerging standards being adopted internationally, we ensure support from familiar tools from samtools to the IGV genome viewer, and interactive environments like Jupyter notebook and RStudio.

To demonstrate the feasibility of computations over federated data sets, we have successfully reproduced classic works such as figures from the thousand genomes project and the recent GENIE marker paper over distributed versions of those data. In addition, we have shown that even highly restricted "differentially private" access to the data permits any important analyses.

Prototype Support for TFRI PROFYLE

PROFYLE is an urgent and important example of the multi-site, cross-Canada projects CanDIG aims to support. Working with the PROFYLE team, we have developed a proof-of-concept, API-driven project dashboard which will display the status of samples, sequencing, and analysis pipelines across the project as soon as they are made available.

Next steps will include simple interactive visualization of reads and variants of individual samples and comparisons between samples, even located at different sites.



CanDIG connects sites at McGill University, Hospital for Sick Children, UHN Princess Margaret Cancer Centre, Canada's Michael Smith Genome Sciences Centre, Jewish General Hospital and Université de Sherbrooke. For more information, visit us at http://www.distributedgenomics.ca

SickKids RESEARCH INSTITUTE UHN Princess Margaret Cancer Centre C3G McGill CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE