

Breakout Session 4: Track A

IGVF Cloud Computing

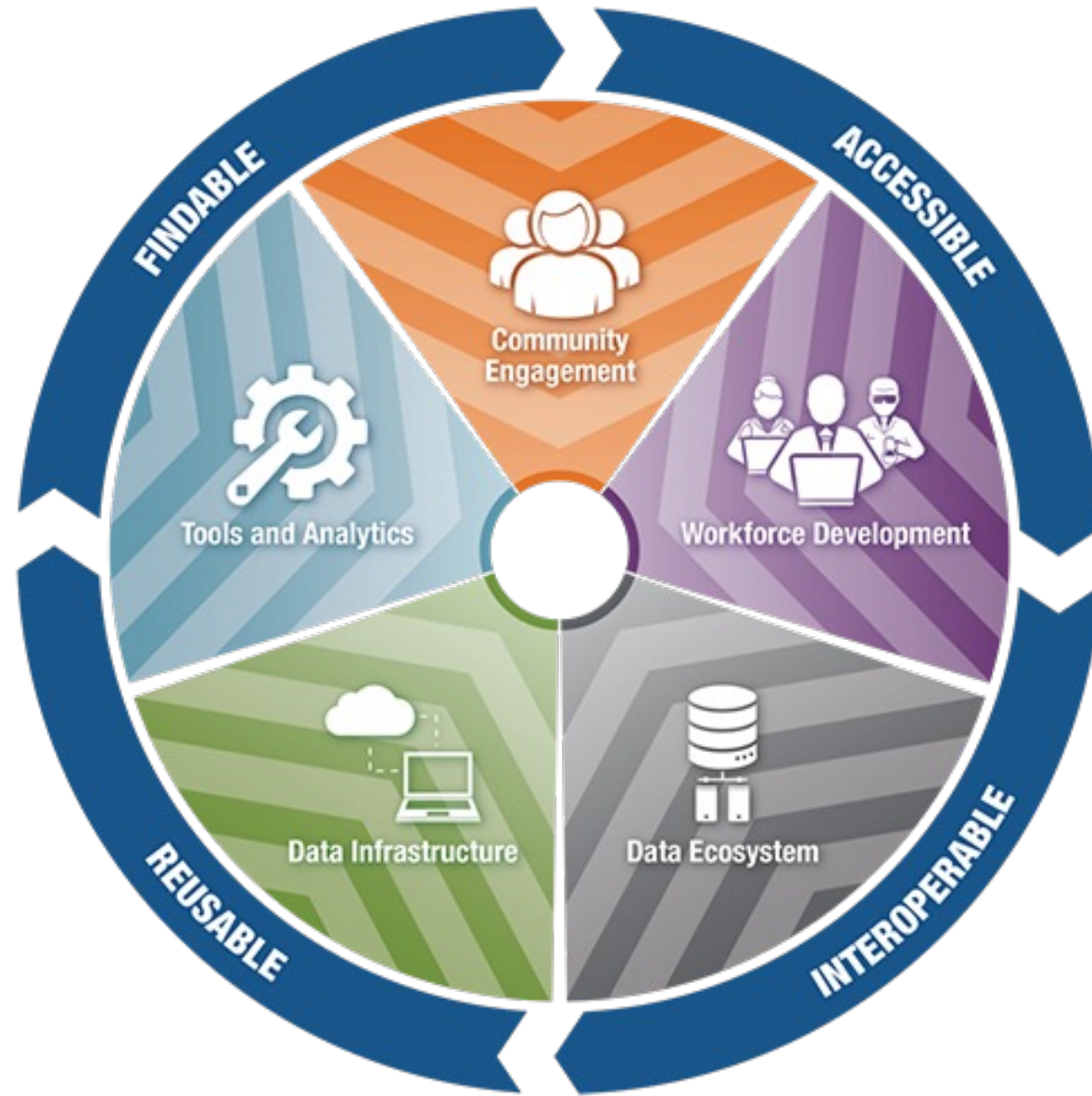
Dr. Ben Hitz
MPI, Stanford University

IGVF Cloud Computing

Ben Hitz, PI
Data Administration and
Coordination Center

January 17-18, 2024



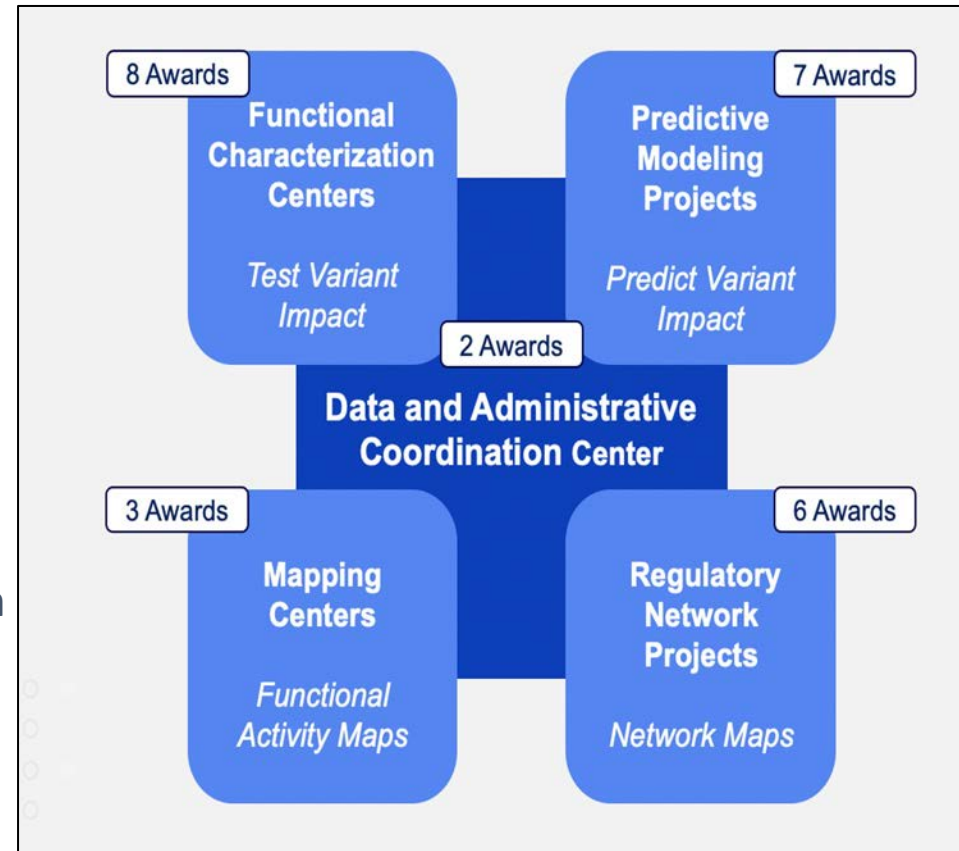


Impact of Genomic Variation on Function

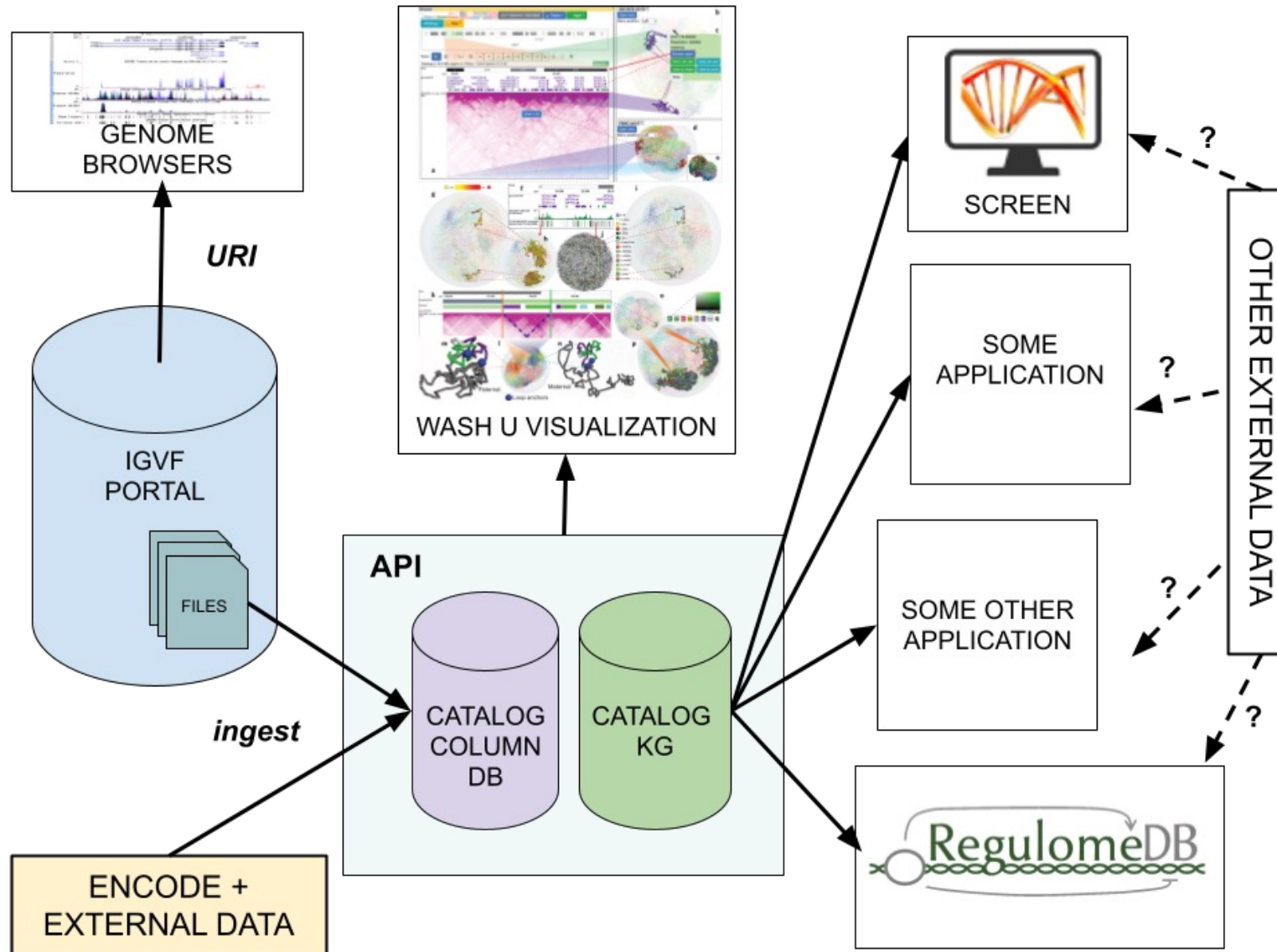
Transforming our understanding of how variation impacts function and leads to phenotypes in health and disease by:

- Using systematic perturbation to assess genome function
- Identification of where and when genes and regulatory elements function at high-resolution
- Network-understanding of genome function
- Development of predictive models of genome function
- Generation of a catalog of elements, variants and phenotypes; share data, tools, and models
- Enabling others to apply these approaches

- What is the iGVF?
 - 24 grants 75 PIs in 4 areas
 - 7 working groups 18 focus groups
 - 40-something assays coding, noncoding, MPRAs, CRISPR screens
 - Emphasis on functional characterization and perturbation and single-cell (nuclei)



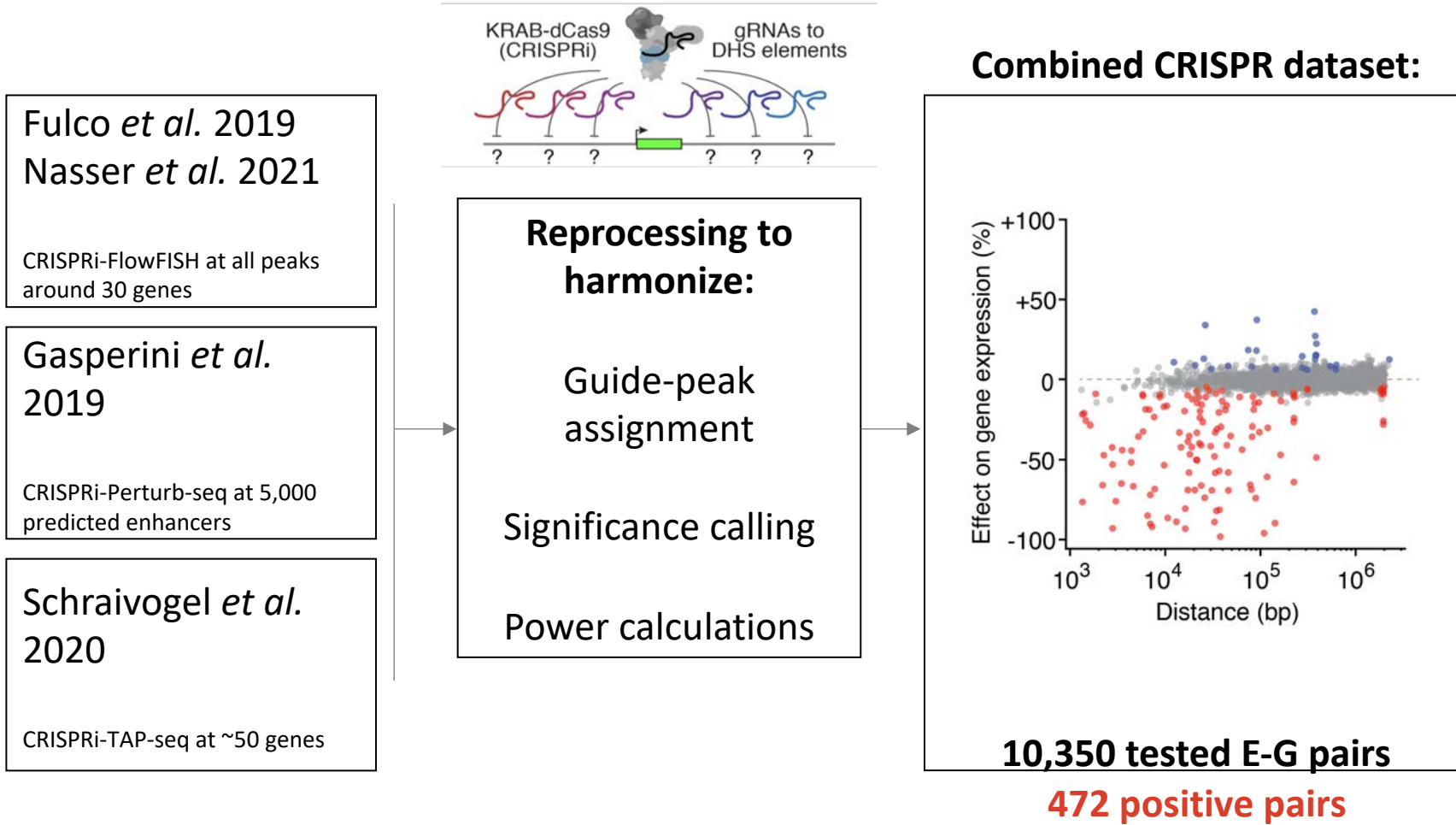
- 12+ years of supporting the ENCODE Project and iGVF
- All computing done in the cloud since 2014
 - Web development environments in AWS
 - Over 1.2PB of storage in AWS S3, mirrored in Azure
 - Long term solutions can be hardware independent
 - Computation has essentially unlimited bursting capacity
- Development of Uniform Processing Pipelines for ENCODE
 - Developed hardened pipelines for 7 assays: ChIP-seq, RNA-seq (long and short reads), ATAC-seq, DNase-seq, HiC, WGBS over 8 years
 - Almost 15,000 experiments analyzed, several Terabases of sequence run in the cloud (ca. 5,000,000 CPU*GB Hours)
 - Quality Control metrics and file provenance available on the Portal
 - Reproducible, platform independent, supported pipeline code available in Github and Docker (WDL/Cromwell)



- Output of genomics pipelines are sensitive to:
 - Choice of genome, transcriptome reference
 - Choice of software, algorithms
- Uniform processing removes this technical variation that can confound results
- Uniform processing allows uniform quality metrics to be calculated
- Genomics is hard enough as it is

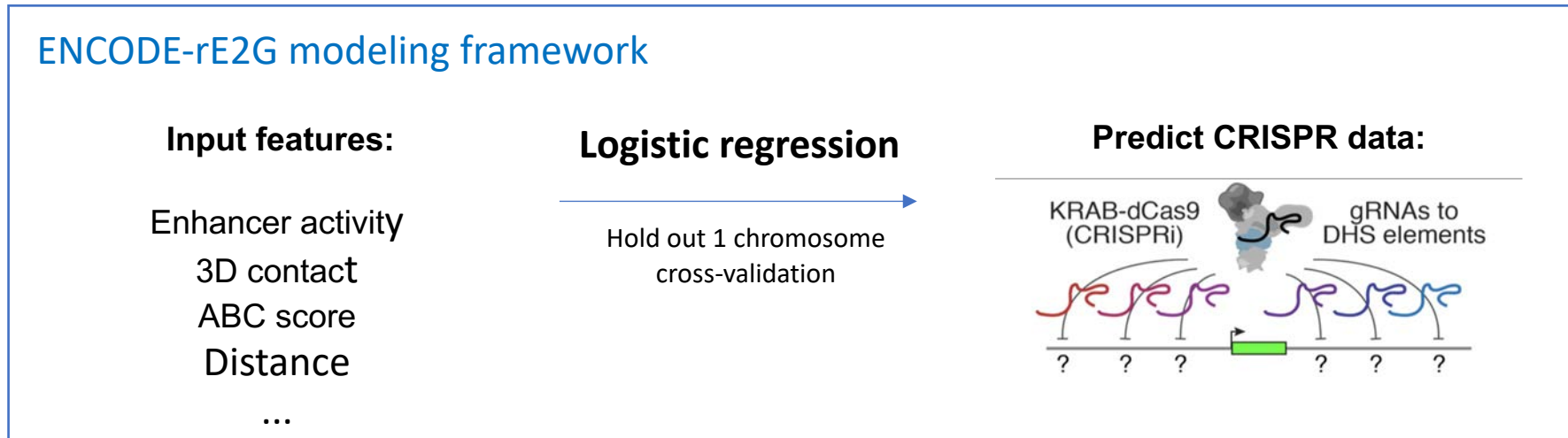
- Primary Method: The “Jamboree”
 - A Jamboree is like a Hackathon but for data analysis
 - A way of doing collaborative development in a short time frame using a cloud compute platform like Terra
- Developing uniform pipelines for single-cell multi-omics and functional characterization
 - Test core pipeline
 - Test and evaluate workflow engines and cloud platforms
 - Develop semi-automated interactive cell annotation
- Developing methods to benchmark models of enhancer-gene regulatory interactions
 - Collect, analyze, curate and harmonize genetic perturbations
 - Develop standards for formats and evaluation of models

Expanding our CRISPR benchmarking dataset



ENCODE-rE2G: A supervised classifier to predict E-G regulatory interactions

Idea: Could we extend ABC to include other possible molecular mechanisms?



Features not included in ABC:

- Enhancer-promoter compatibility?
- Enhancer-enhancer synergy?
- Non-linear functions of A and C?
- ...

Compare ENCODE and IGVF assays:

- DNase-seq
- H3K27ac and other ChIP-seq
- Hi-C, ChIA-PET
- ...

IGVF DACC:

Pedro Assis ♦ Shengcheng Dong ♦ Keenan Graham ♦ Otto Jolanki ♦ Meenakshi Kagda
Khine Lin ♦ Jennifer Jou ♦ Jin-Wook Lee ♦ Mingjie Li ♦ Corinn Small ♦ Forrest Tanaka
Ian Whaling ♦ Ingrid Youngworth ♦ Lucinda Fulton ♦ Sara Cody ♦ Wenjin Zhang
Xiaowen Ma ♦ Daofeng Li ♦ Heather Lawson ♦ Feng Yue ♦ Ting Wang

IGFV Single Cell, MPRA, and CRISPR Focus Groups

Kundaje and Engrietz Labs