Introducing Glow

William Brandler, Karen Feng
Data + AI Meetup
August 11, 2020

Outline

Current state of genomics

- The transition from academia to industry
- Genomics for drug discovery
- Genomics as a big data problem

How Glow solves problems with big genomics data

- From ad hoc command line tools to production data science
- Partnership between Databricks and Regeneron
- Innovations in Glow

Genomics is taking off in industry

databricks

Ultimately, genomics will be fully integrated into clinical practice

ESSAY

GENOME-SEQUENCING ANNIVERSARY

A Celebration of the Genome, Part I

Ten years ago, the first peer-reviewed reports of the sequencing of the human genome were published. At that time we announced that "humanity has been given a great gift," which has proven to be the case in many ways but has also presented a great challenge. To commemorate the event, we have asked a cross section of insightful individuals—representing many viewpoints—to look at what it has meant to them or their communities to have access to human genome sequences. We will be publishing their comments throughout the month of February.

–Barbara R. Jasny and Laura M. Zahn

Faces of the Genome

Francis S. Collins
Director, National Institutes of Health, Bethesda, MD, USA.

formed in July 2010, using stem cells from the cord blood of a matched, healthy donor. Seven months later, Nic appears to be on the road to recovery. While he is still on immunosuppressants, doctors report the new stem cells are stably engrafted, blood counts are good, and there's been no

https://science.sciencemag.org/content/331/6017/546.2

Genomics is revolutionizing drug development

90% of experimental medicines fail during clinical trials

>$100bn spent on pharma R&D **per year**
Only 10-20 new drugs are approved each year

2X greater chance of drugs being approved if they have supporting genetic evidence

The power of big genomic data

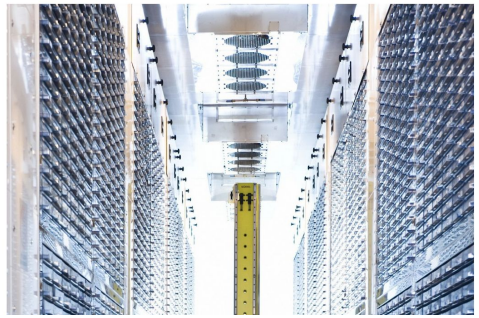
MIT Technology Review

Log in / Create an account Search Subscribe

Topics+ The Download Magazine Events More+

Rewriting Life

UK Biobank Supercharges Medicine with Gene Data on 500,000 Brits



genomeweb

Business & Policy Technology Research Diagnostics Disease Areas Applied Markets

Home » Tools & Technology » Sequencing » Pilot GenomeAsia 100K Project Highlights Insights Gleaned From Diverse Dataset

Pilot GenomeAsia 100K Project Highlights Insights Gleaned From Diverse Dataset

PharmaTimes online

Search news, articles, experts, insights or people GO

Magazine Web Exclusives News Competitions Appointments Jobs Business Insights Webinars Thought Leadership

Free news subscriptions Free RSS feeds

UK's genome sequencing project will hit 100,000 by year-end

4th July 2018

REGENERON

> SCIENTISTS > PATIENTS > MEDICAL PROFESSIONALS > INVESTORS & MEDIA CAREERS

ABOUT SCIENCE MEDICINES RESPONSIBILITY

Regeneron Perspectives

A MILLION IS JUST THE BEGINNING

AUTHOR: JOHN OVERTON, PHD, *VICE PRESIDENT, SEQUENCING AND LAB OPERATIONS*

PUBLISHED ON: FEBRUARY 28, 2020

databricks

Pharma are adopting population-scale genomics

Population genetics is used in industry to

- build drug discovery pipelines
- prioritize R&D investments

Through understanding

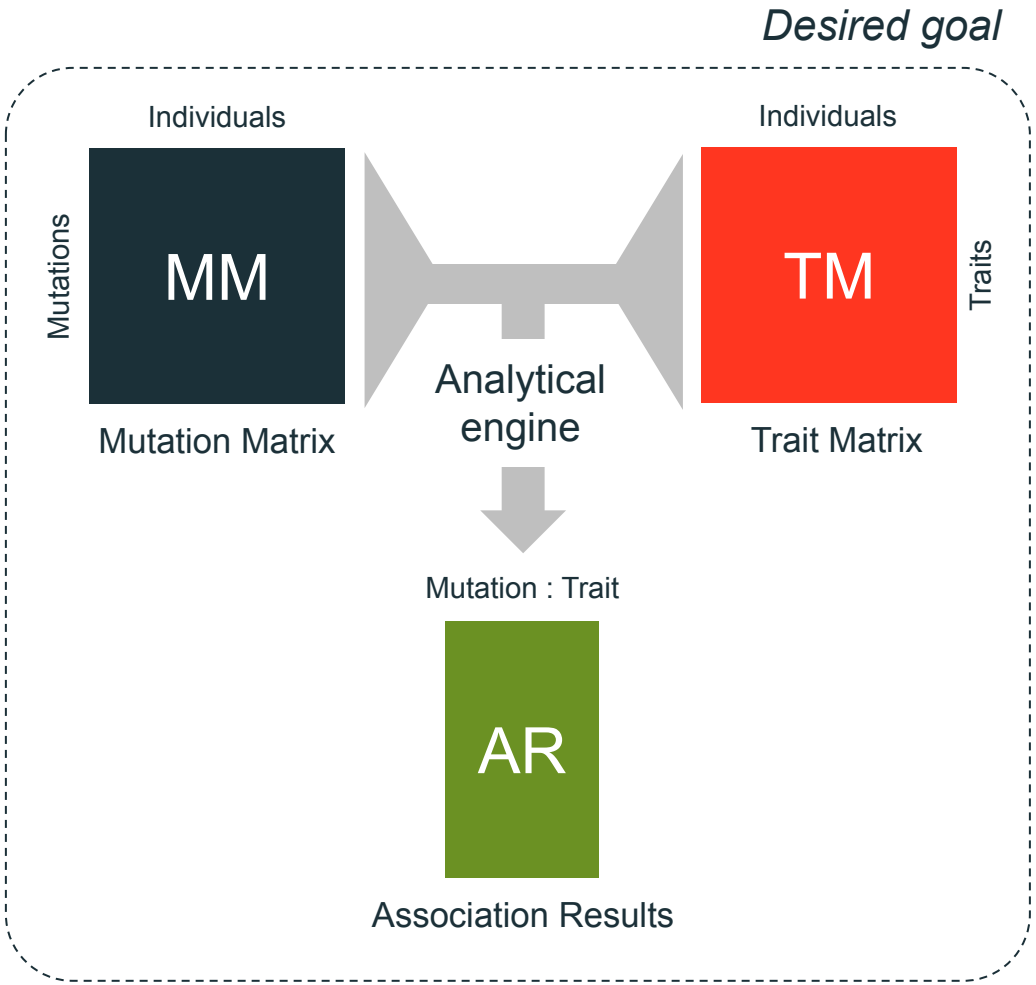
- the exact population frequency of specific genetic disorders
- the relationship between clinical phenotypes and genetics
- how patients respond in clinical trials based on their genetics
- what genes are amenable to drug discovery

Bigger pipeline - Increased likelihood of drug approval - Billions of dollars of savings

The Regeneron Genetics Center mines the world's genomics data to extract insights

Approach:

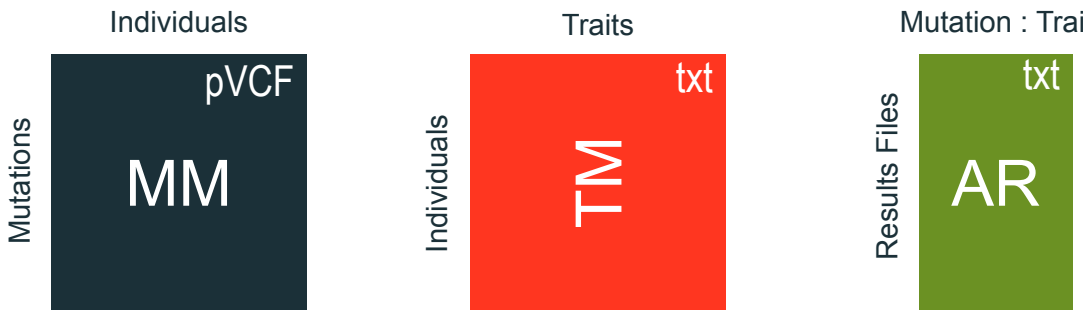
- Sequence a large number of individuals from many cohorts (>70 to date)
- Obtain paired phenotypic data (e.g. de-identified electronic medical records)
- Run all-vs-all association tests between all mutations and traits
- Mine association results to extract actionable insights
- Design for scalability & automation



databricks

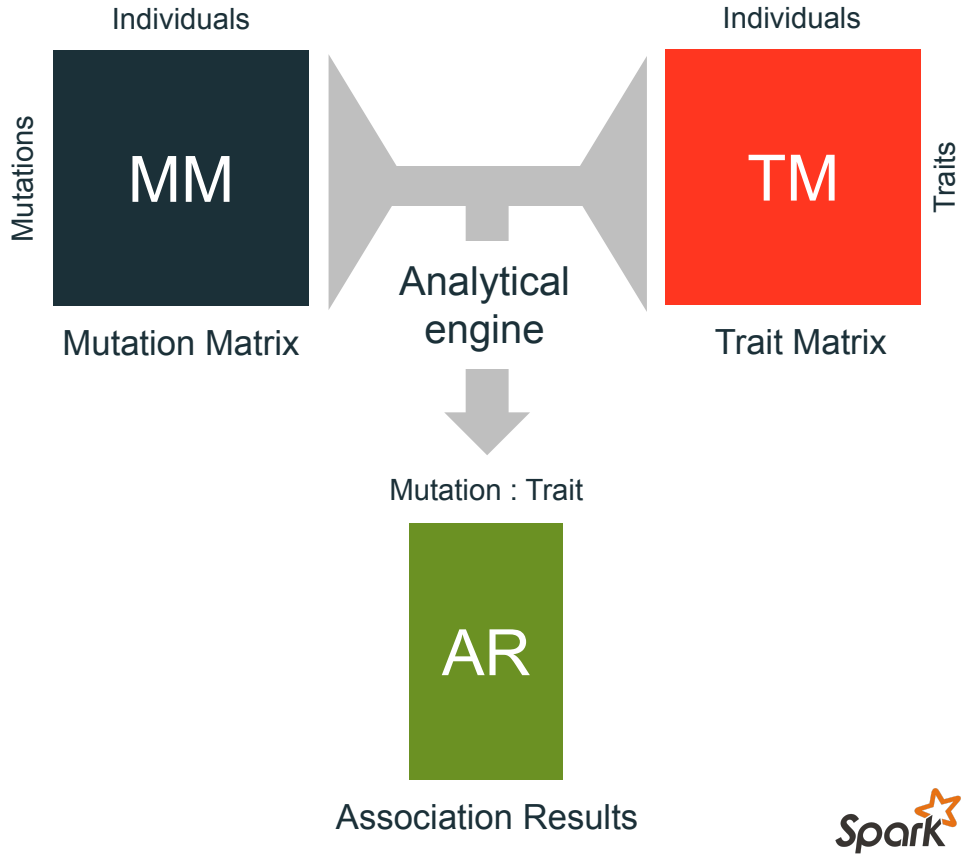
But it's complicated to gain novel insights from genomics data

Reality



- **Hard to scale:** Data is decentralized, assumes single node architecture
- **Rigid tools:** Data is organized in different ways (e.g., not squared off, transposed, custom representations and indexing)
- **Hard to move to tertiary analyses:** Asking simple questions at scale requires time-consuming data wrangling steps

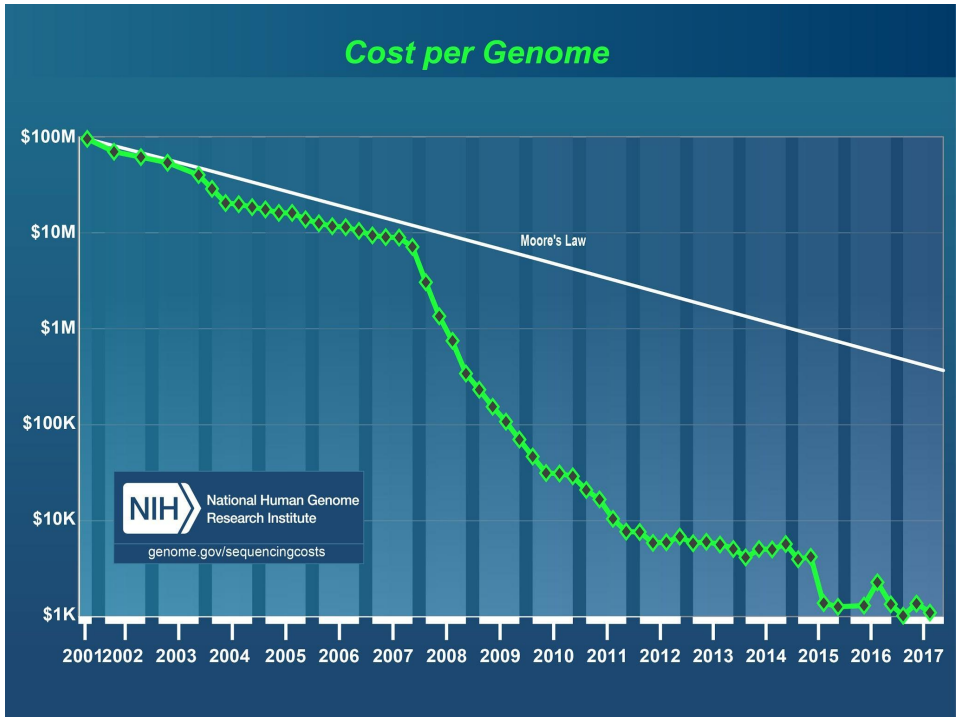
Desired goal



databricks

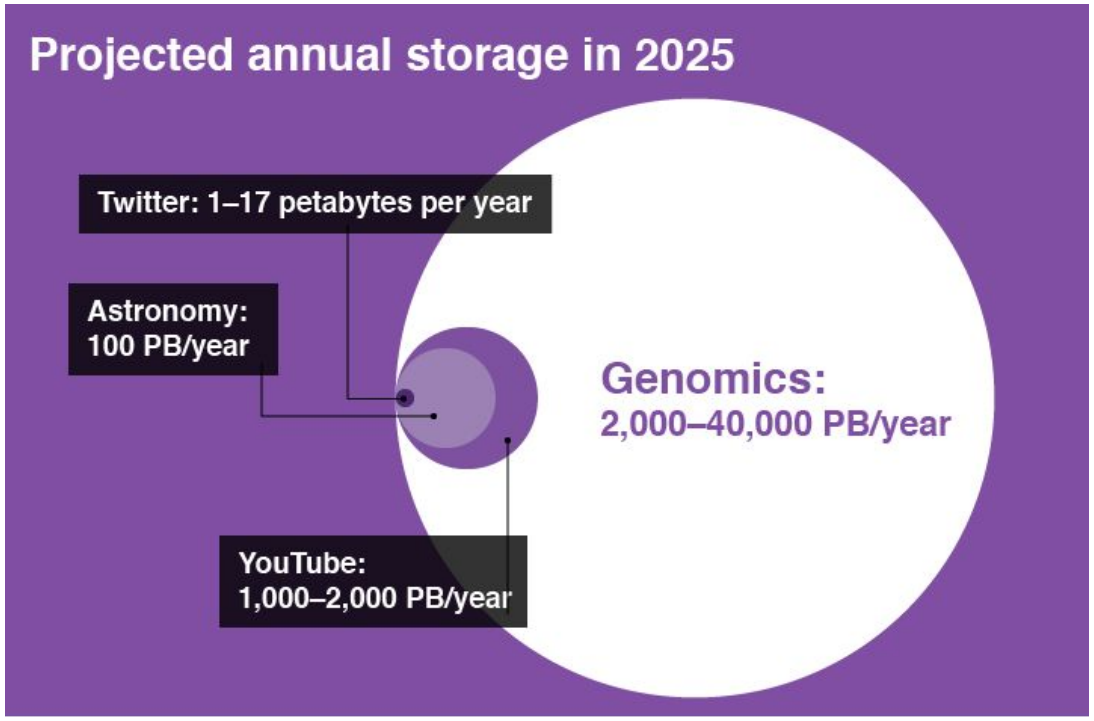
Genomics is a big data problem

From $2.7B to <$1,000



https://www.genome.gov/27541954/dna-sequencing-costs-data/

40,000 Petabytes / year by 2025



https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

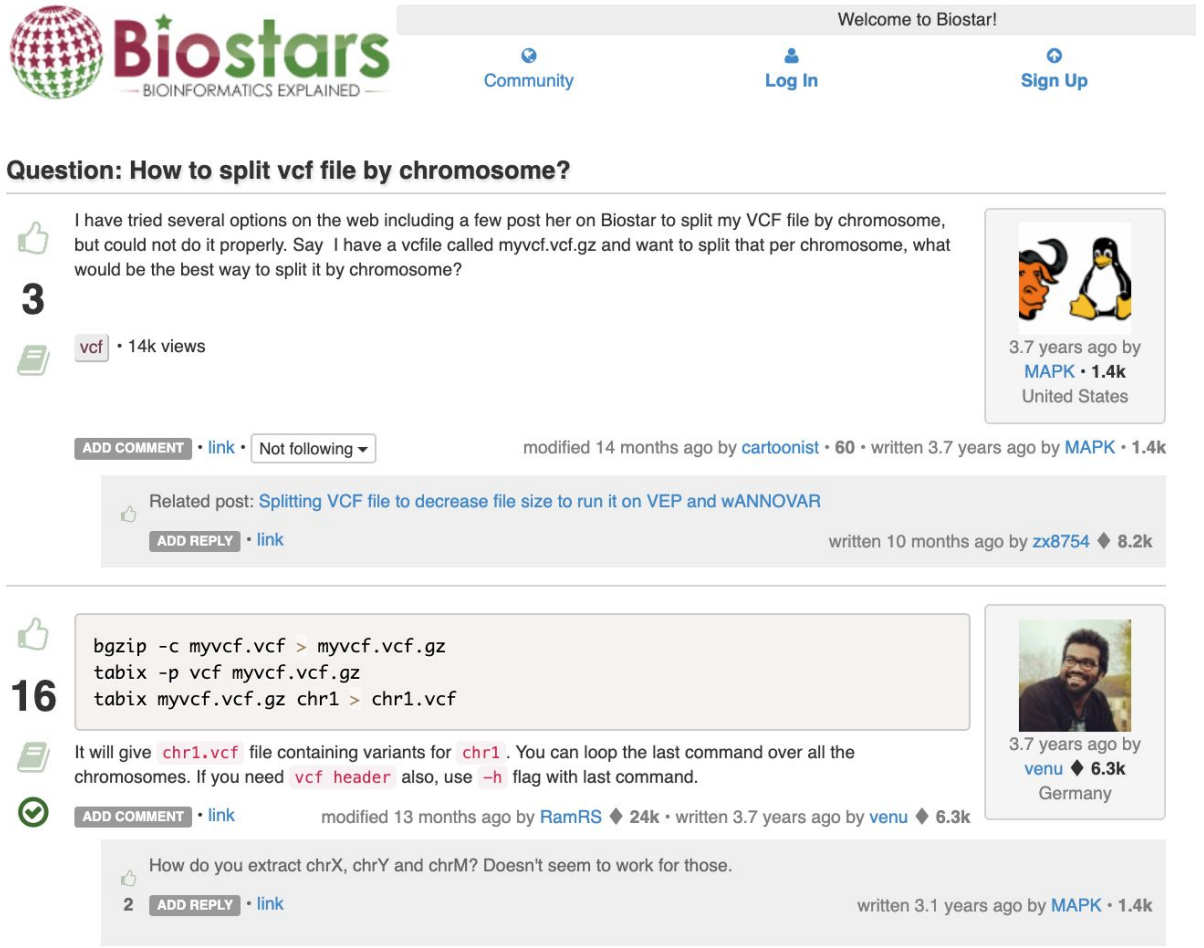
databricks

Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate

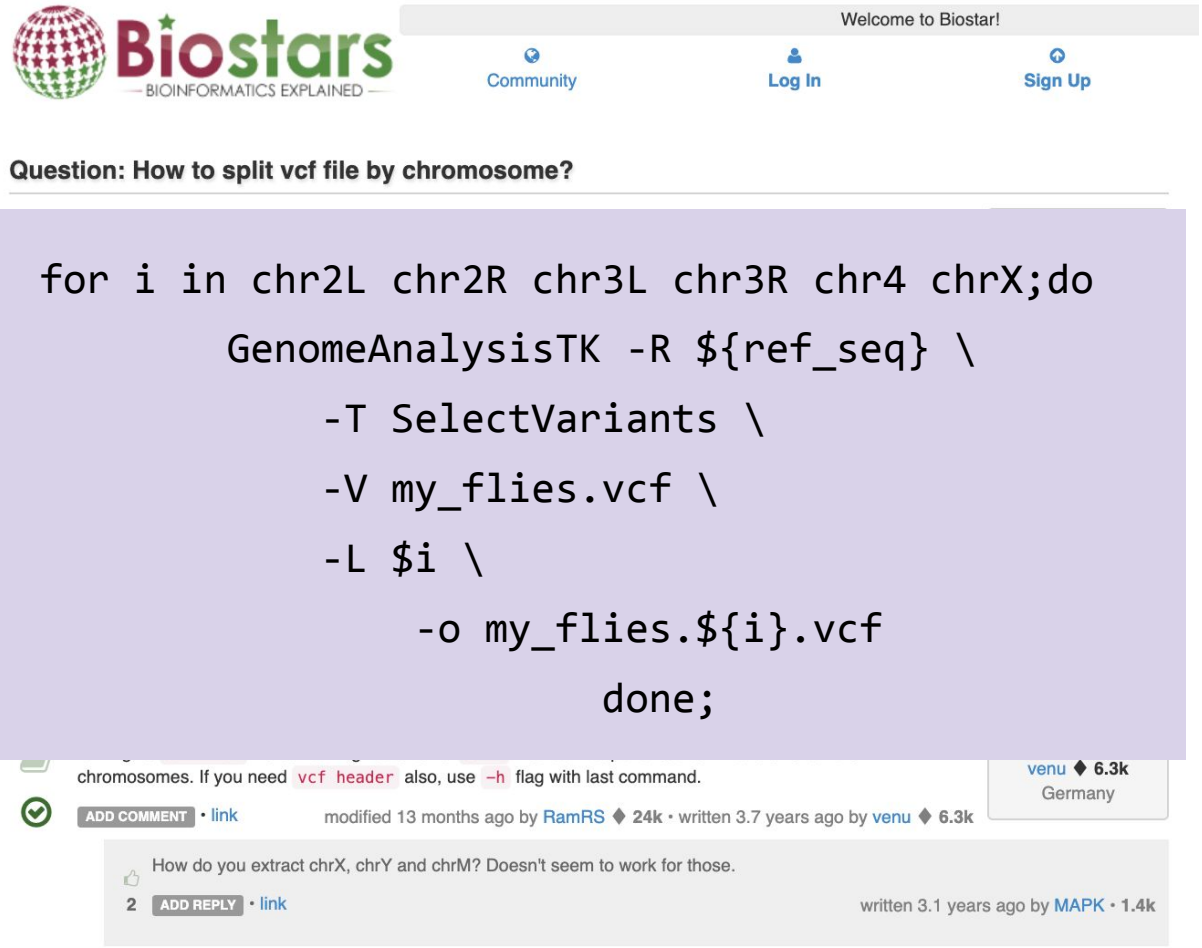
Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate



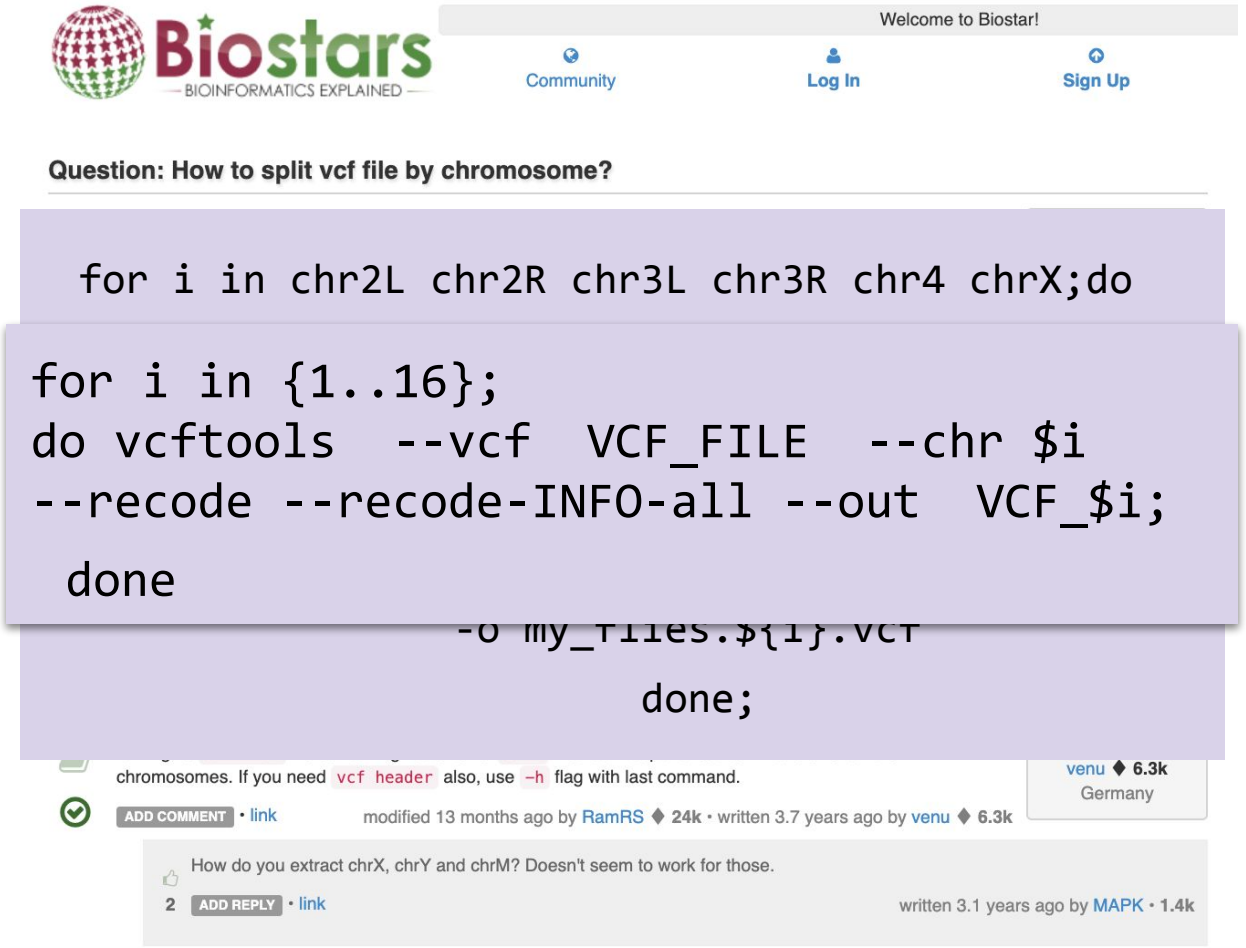
Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate



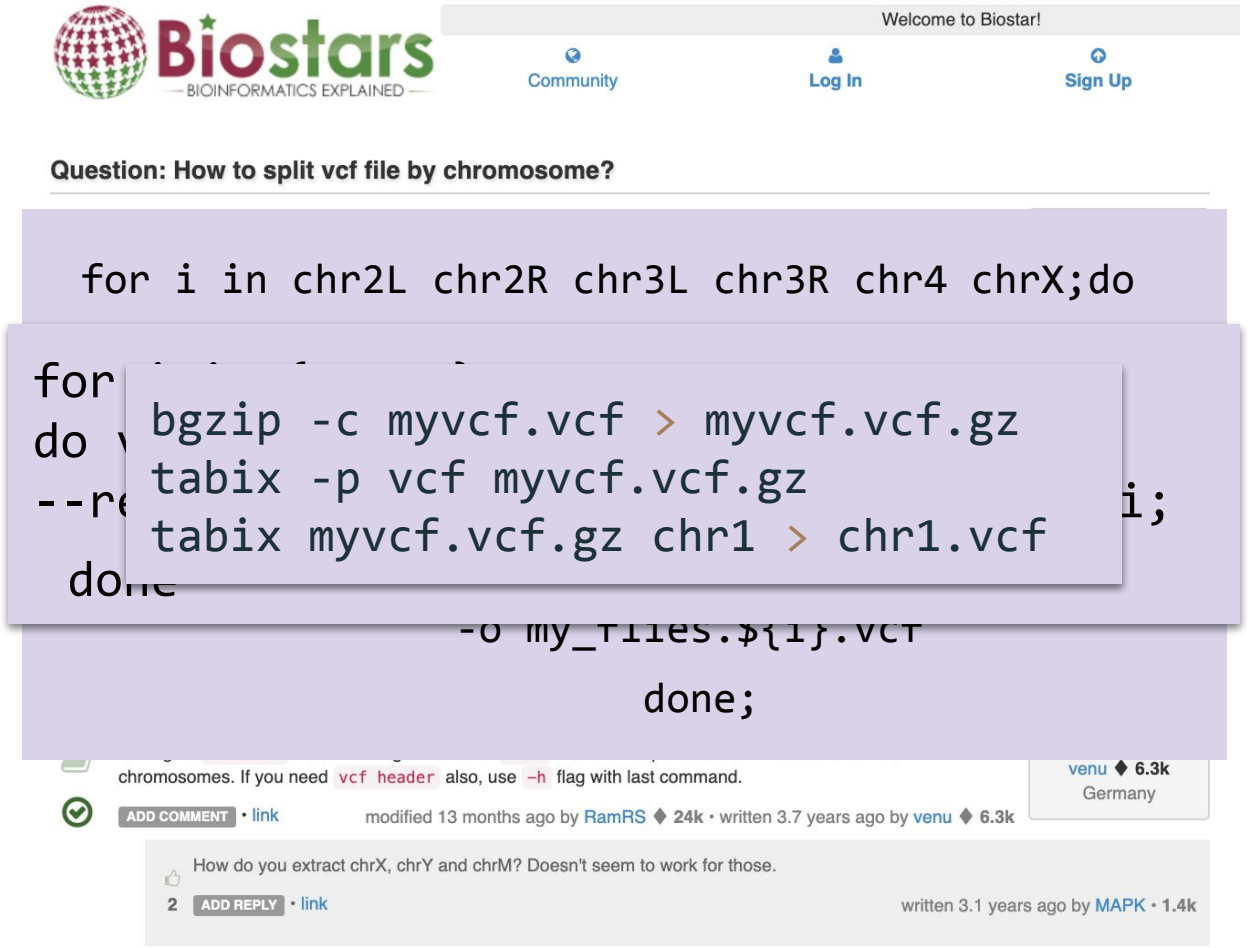
Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate



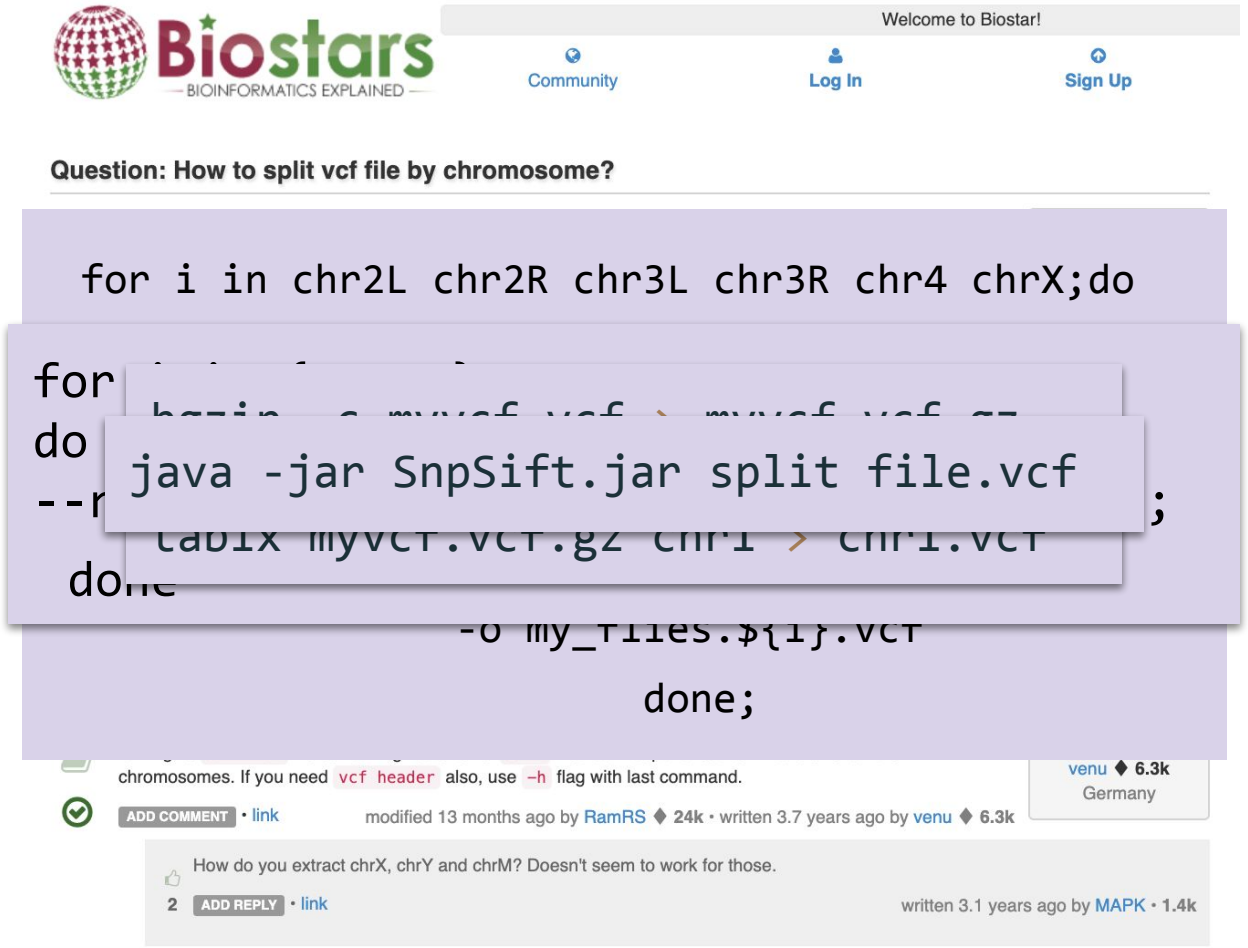
Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate



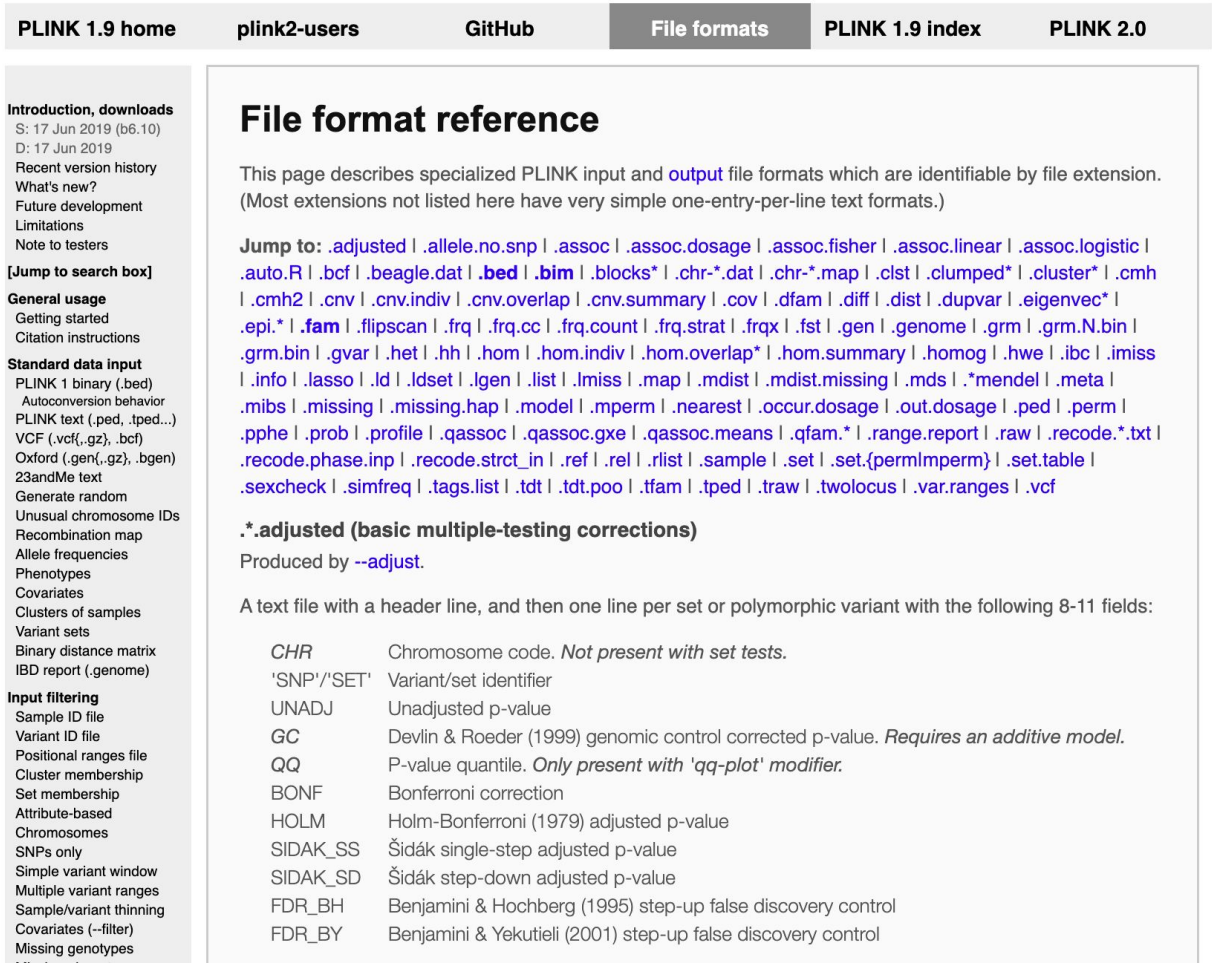
Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate



Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate



PLINK 1.9 home | plink2-users | GitHub | File formats | PLINK 1.9 index | PLINK 2.0

Introduction, downloads
S: 17 Jun 2019 (b6.10)
D: 17 Jun 2019
Recent version history
What's new?
Future development
Limitations
Note to testers
[Jump to search box]
General usage
Getting started
Citation instructions
Standard data input
PLINK 1 binary (.bed)
Autoconversion behavior
PLINK text (.ped, .tped...)
VCF (.vcf{,.gz}, .bcf)
Oxford (.gen{,.gz}, .bgen)
23andMe text
Generate random
Unusual chromosome IDs
Recombination map
Allele frequencies
Phenotypes
Covariates
Clusters of samples
Variant sets
Binary distance matrix
IBD report (.genome)
Input filtering
Sample ID file
Variant ID file
Positional ranges file
Cluster membership
Set membership
Attribute-based
Chromosomes
SNPs only
Simple variant window
Multiple variant ranges
Sample/variant thinning
Covariates (--filter)
Missing genotypes

File format reference

This page describes specialized PLINK input and output file formats which are identifiable by file extension. (Most extensions not listed here have very simple one-entry-per-line text formats.)

Jump to: .adjusted | .allele.no.snp | .assoc | .assoc.dosage | .assoc.fisher | .assoc.linear | .assoc.logistic | .auto.R | .bcf | .beagle.dat | **.bed** | **.bim** | .blocks* | .chr-*.dat | .chr-*.map | .clst | .clumped* | .cluster* | .cmh | .cmh2 | .cnv | .cnv.indiv | .cnv.overlap | .cnv.summary | .cov | .dfam | .diff | .dist | .dupvar | .eigenvec* | .epi.* | **.fam** | .flipscan | .frq | .frq.cc | .frq.count | .frq.strat | .frqx | .fst | .gen | .genome | .grm | .grm.N.bin | .grm.bin | .gvar | .het | .hh | .hom | .hom.indiv | .hom.overlap* | .hom.summary | .homog | .hwe | .ibc | .imiss | .info | .lasso | .ld | .ldset | .lgen | .list | .lmiss | .map | .mdist | .mdist.missing | .mds | .*mendel | .meta | .mibs | .missing | .missing.hap | .model | .mperm | .nearest | .occur.dosage | .out.dosage | .ped | .perm | .pphe | .prob | .profile | .qassoc | .qassoc.gxe | .qassoc.means | .qfam.* | .range.report | .raw | .recode.*.txt | .recode.phase.inp | .recode.strct_in | .ref | .rel | .rlist | .sample | .set | .set.{perm|mperm} | .set.table | .sexcheck | .simfreq | .tags.list | .tdt | .tdt.poo | .tfam | .tped | .traw | .twolocus | .var.ranges | .vcf

.*.adjusted (basic multiple-testing corrections)

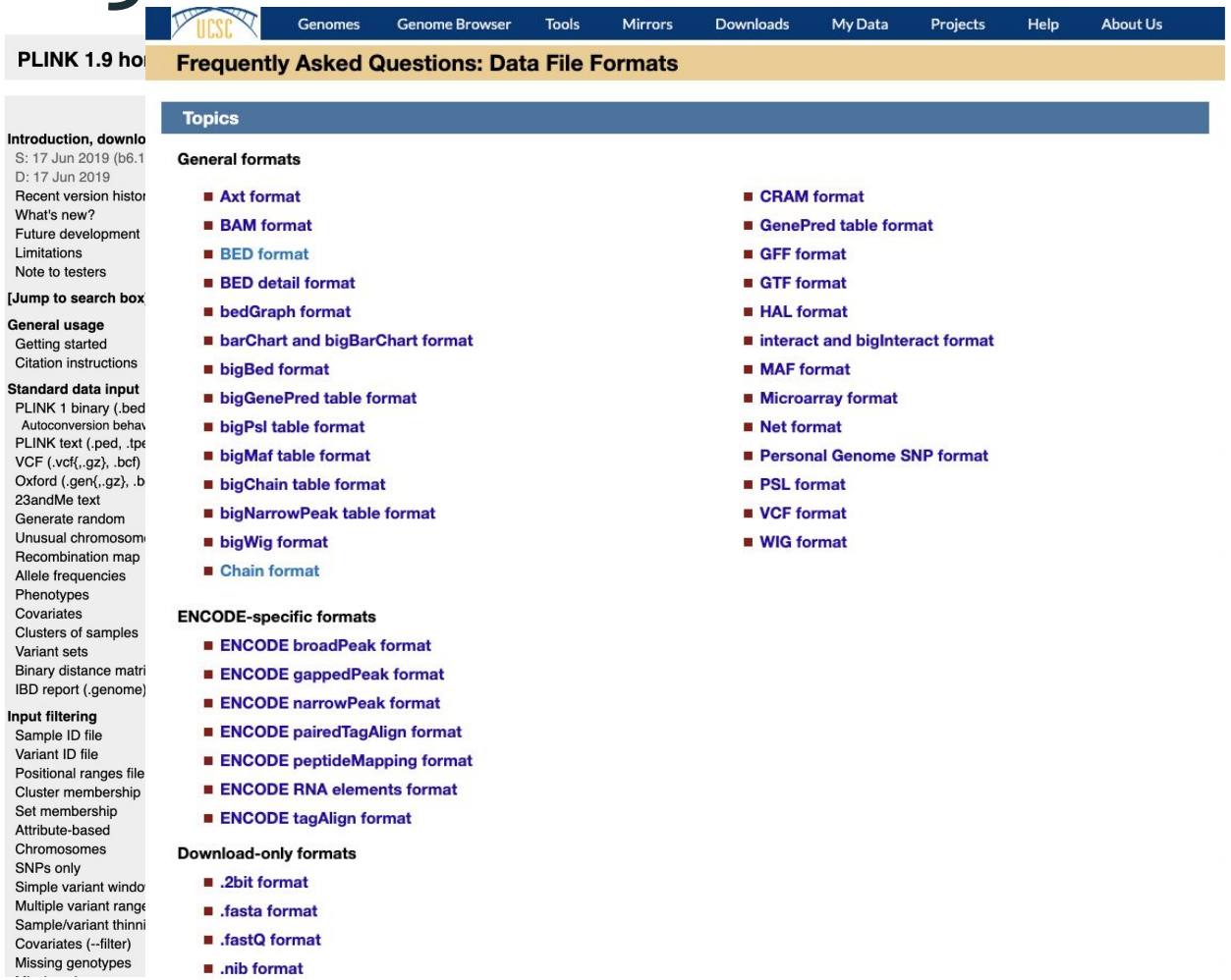
Produced by --adjust.

A text file with a header line, and then one line per set or polymorphic variant with the following 8-11 fields:

CHR	Chromosome code. *Not present with set tests.*
'SNP'/'SET'	Variant/set identifier
UNADJ	Unadjusted p-value
GC	Devlin & Roeder (1999) genomic control corrected p-value. *Requires an additive model.*
QQ	P-value quantile. *Only present with 'qq-plot' modifier.*
BONF	Bonferroni correction
HOLM	Holm-Bonferroni (1979) adjusted p-value
SIDAK_SS	Šidák single-step adjusted p-value
SIDAK_SD	Šidák step-down adjusted p-value
FDR_BH	Benjamini & Hochberg (1995) step-up false discovery control
FDR_BY	Benjamini & Yekutieli (2001) step-up false discovery control

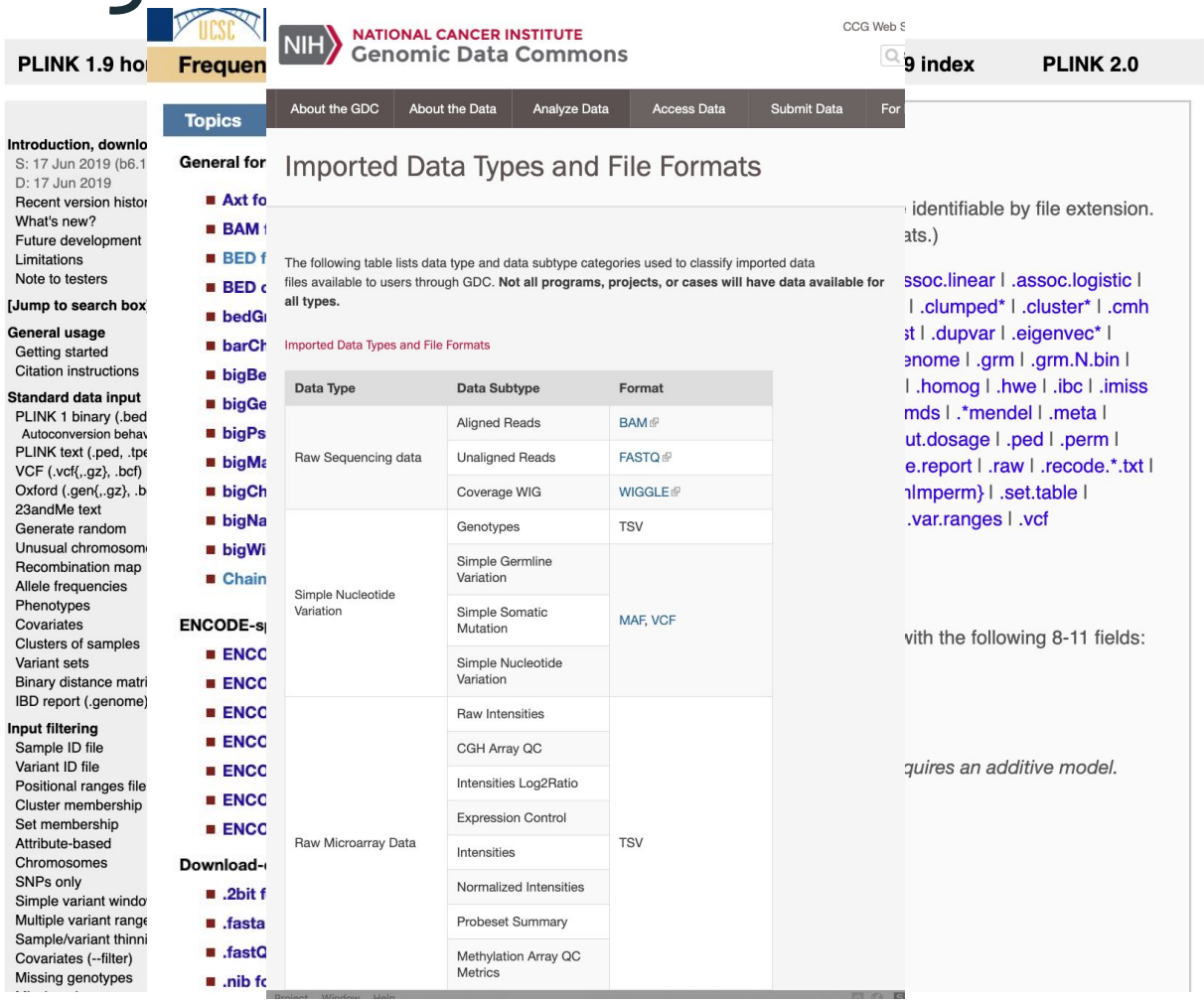
Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate



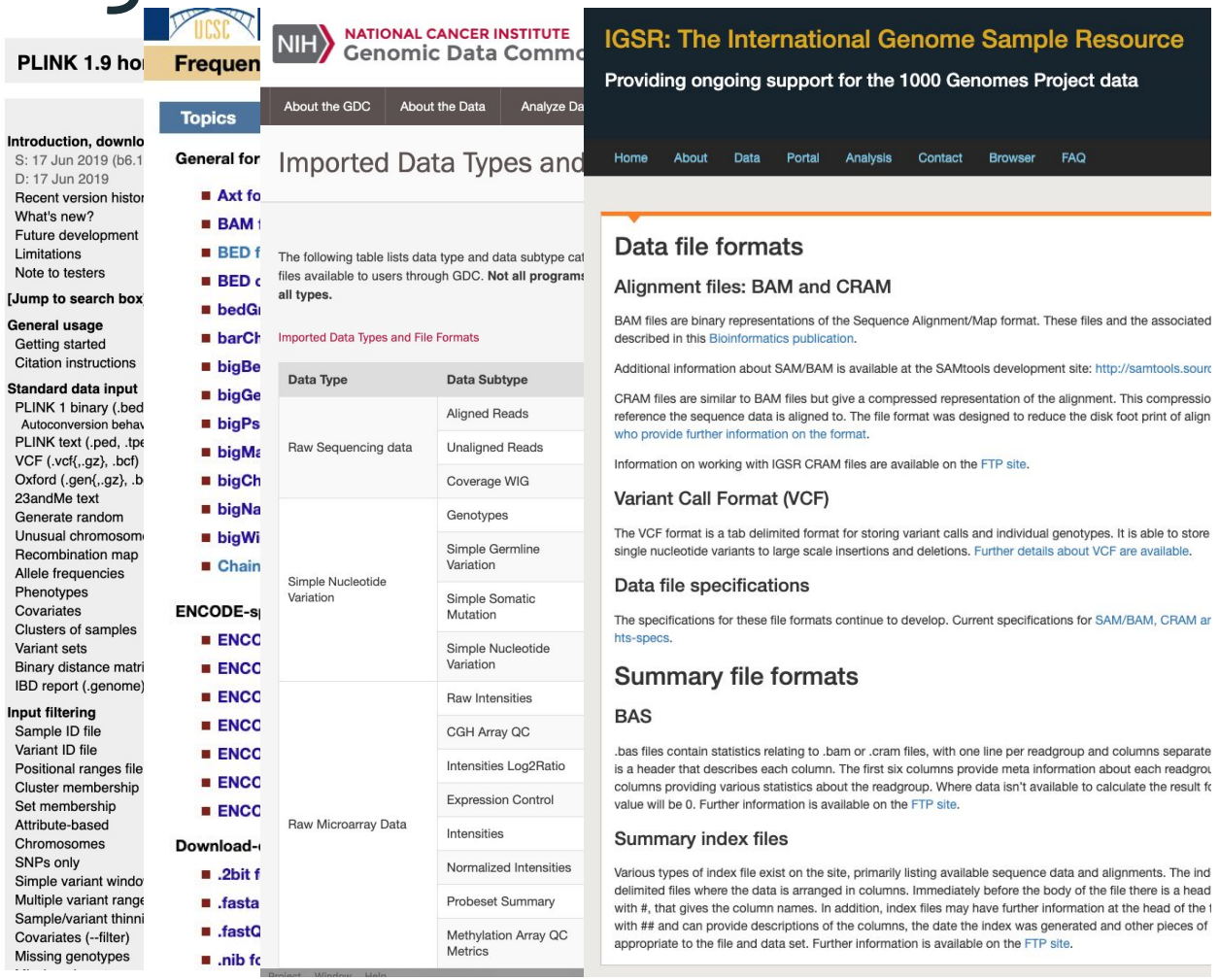
Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate



Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate



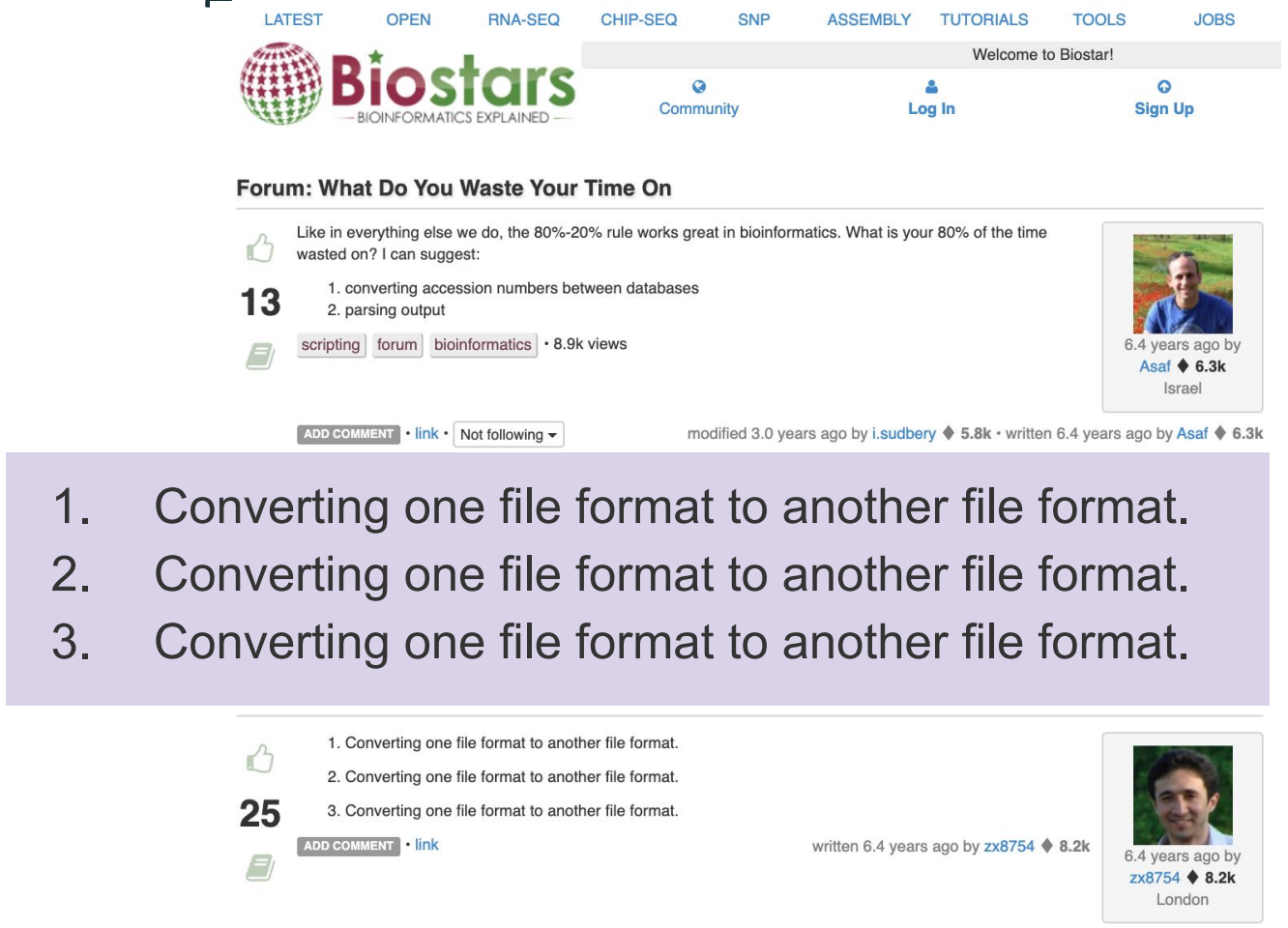
databricks

Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate

Genomic analysis on big data is hard!

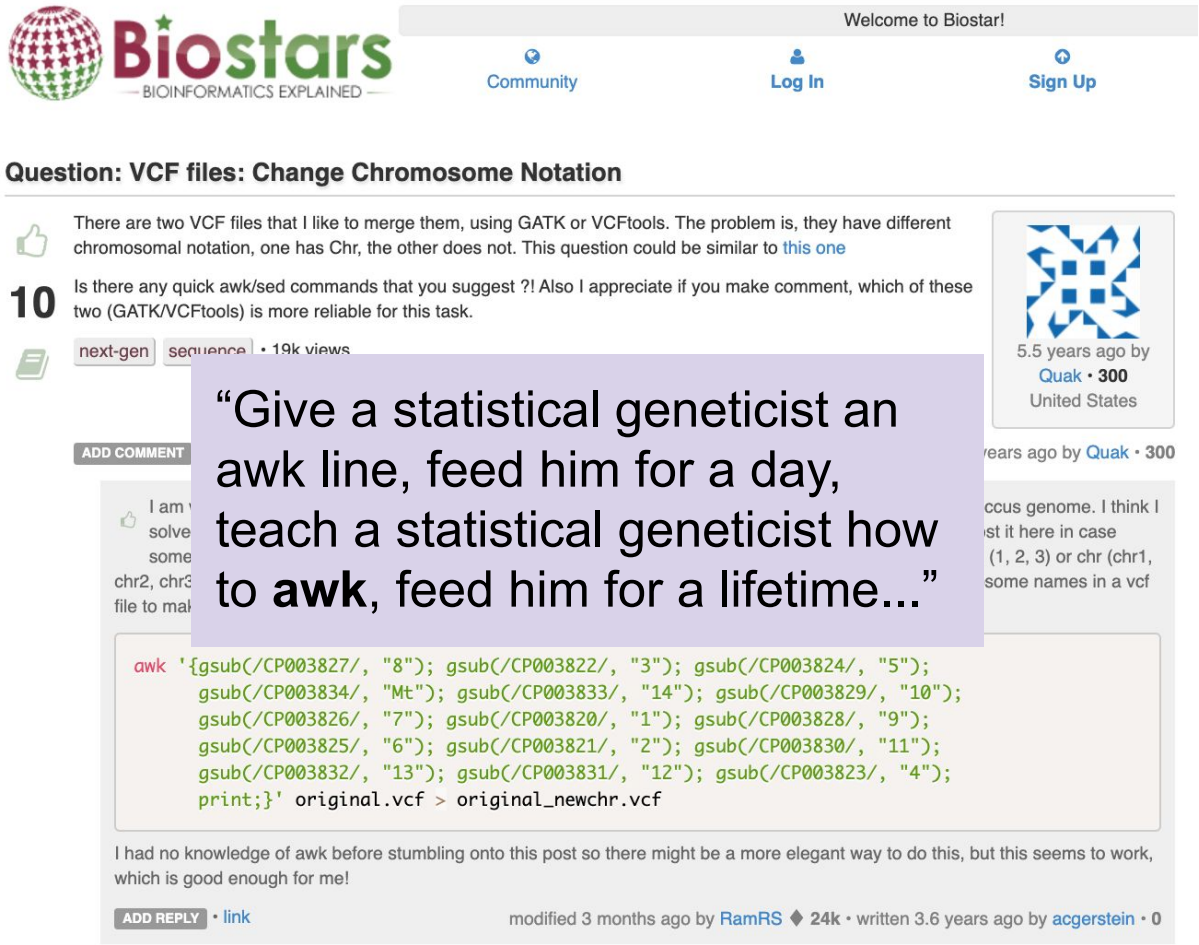
- Existing tools are often difficult to
  - Scale
  - Integrate



databricks

Genomic analysis on big data is hard!

- Existing tools are often difficult to
  - Scale
  - Integrate



Glow is designed to solve these problems

GLOW

Mission: Industrialize genomics by integrating bioinformatics into data science

Core principles:

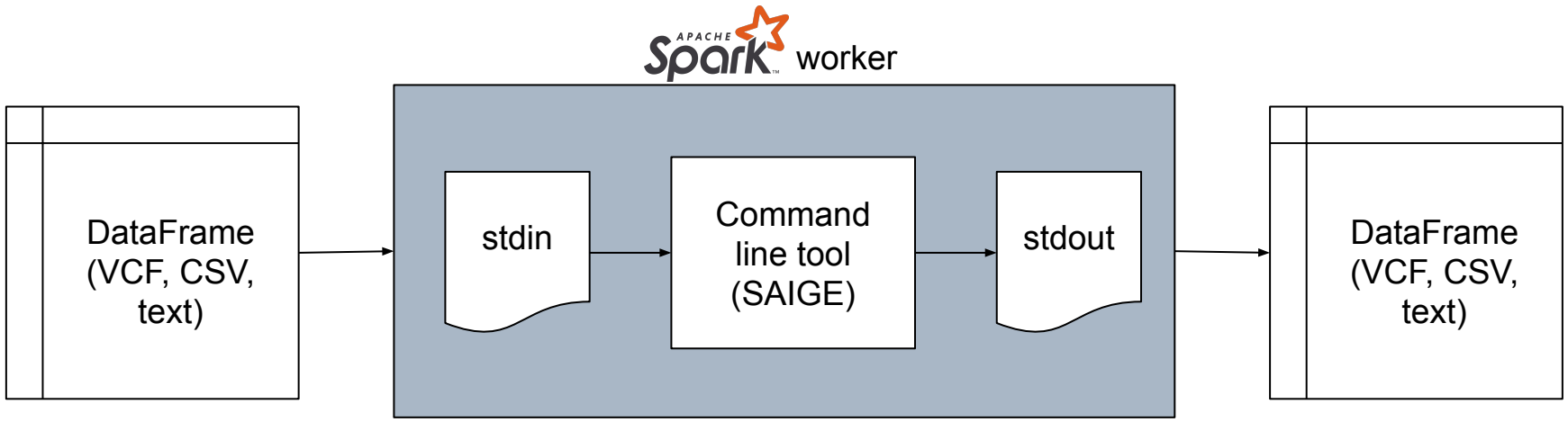
- Build on Apache Spark
- Flexibly and natively support genomics tools and file formats
- Provide single-line functions for common genomics workloads
- Build an open-source community

Glow is built on Apache Spark

- Join with phenotype data
- Transition data to ML/data science ecosystem
- High-level APIs in multiple languages
  - Python, R, SQL, Scala, Java

Glow is backwards-compatible with legacy tools and file formats

- File formats: VCF, BGEN, PLINK, GFF
- Parallelize command line bioinformatics tools with the pipe transformer



Glow provides APIs for common genomics workloads

Variant data manipulation

- Quality control
- Liftover
- Variant normalization
- Split multiallelic variants

Tertiary analysis

- Integration with annotation, clinical & phenotype data
- Genome-wide association studies

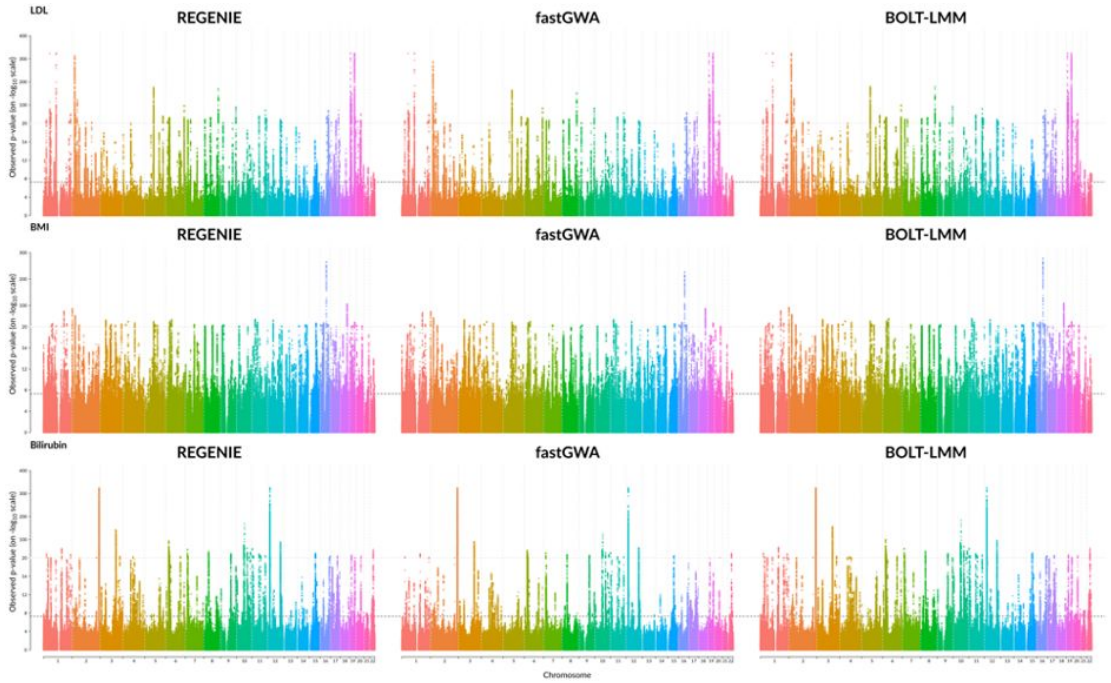
Introducing GloWGR

An industrial-scale, ultra-fast & sensitive method for genetic association studies

databricks

GloWGR is a major innovation in bioinformatics driven by industry collaboration

- Accurately accounts for population structure
- Computationally tractable on biobank-scale datasets with thousands of phenotypes
- Parallelized version of the regenie method using Apache Spark

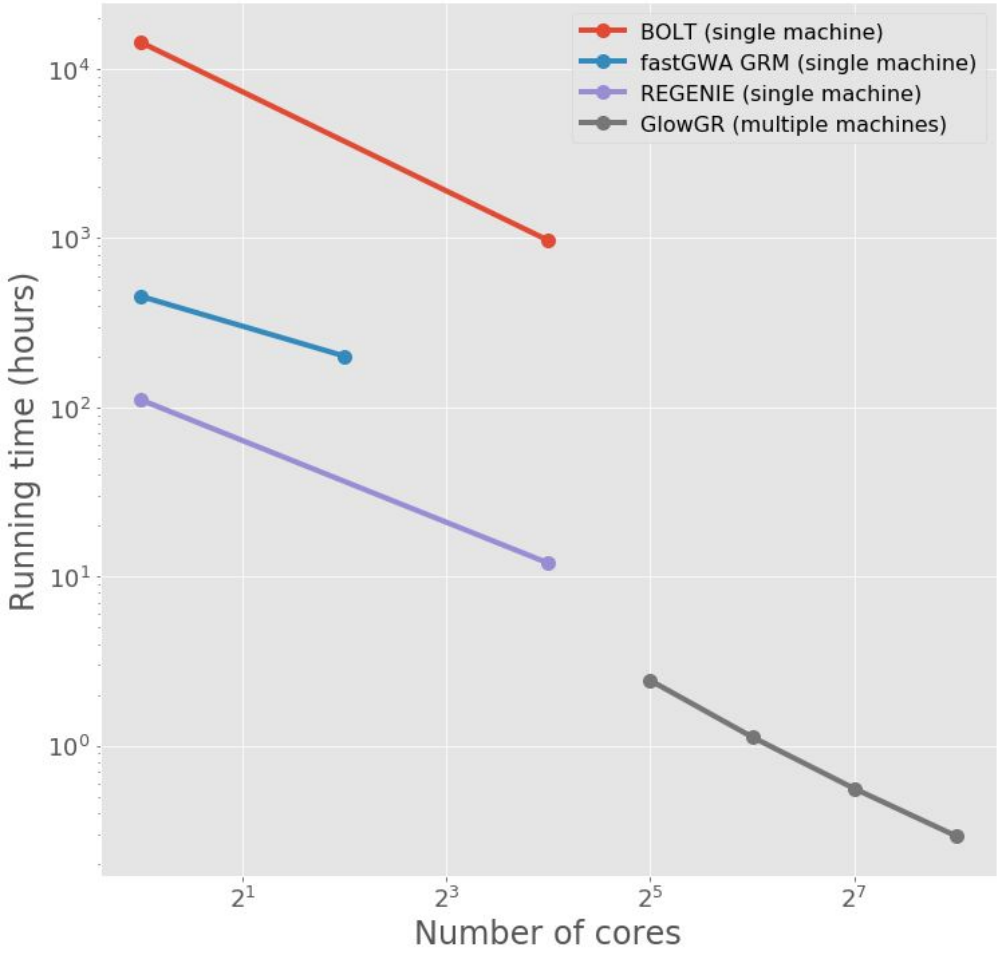


databricks

REGENERON GENETICS CENTER

GloWGR is a major innovation in bioinformatics driven by industry collaboration

- Accurately accounts for population structure
- Computationally tractable on biobank-scale datasets with thousands of phenotypes
- Parallelized version of the regenie method using Apache Spark

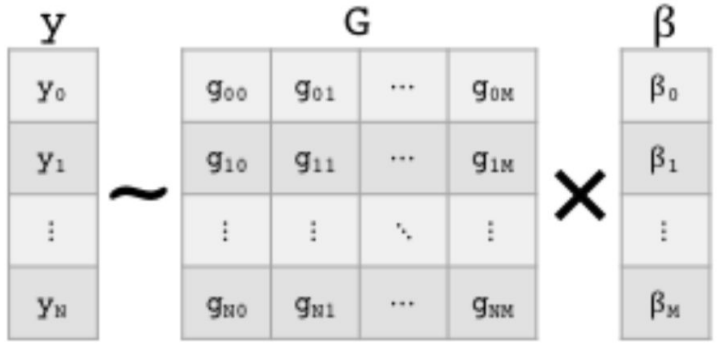


databricks

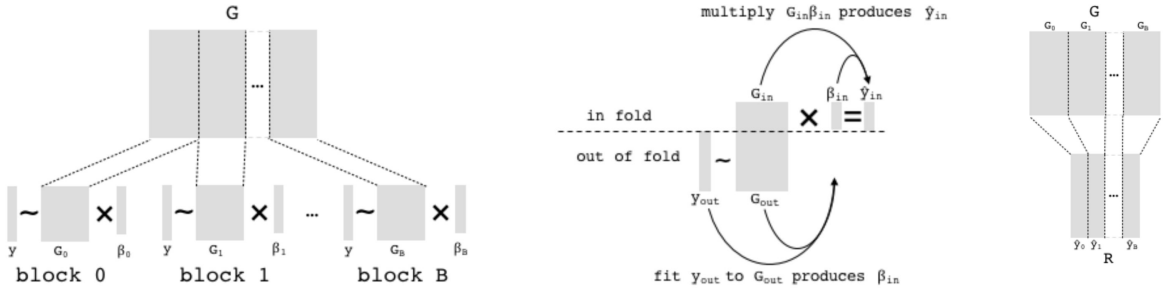
REGENERON GENETICS CENTER

GloWGR is a major innovation in bioinformatics driven by industry collaboration

- Accurately accounts for population structure
- Computationally tractable on biobank-scale datasets with thousands of phenotypes
- Parallelized version of the regenie method using Apache Spark



Traditional WGR Method

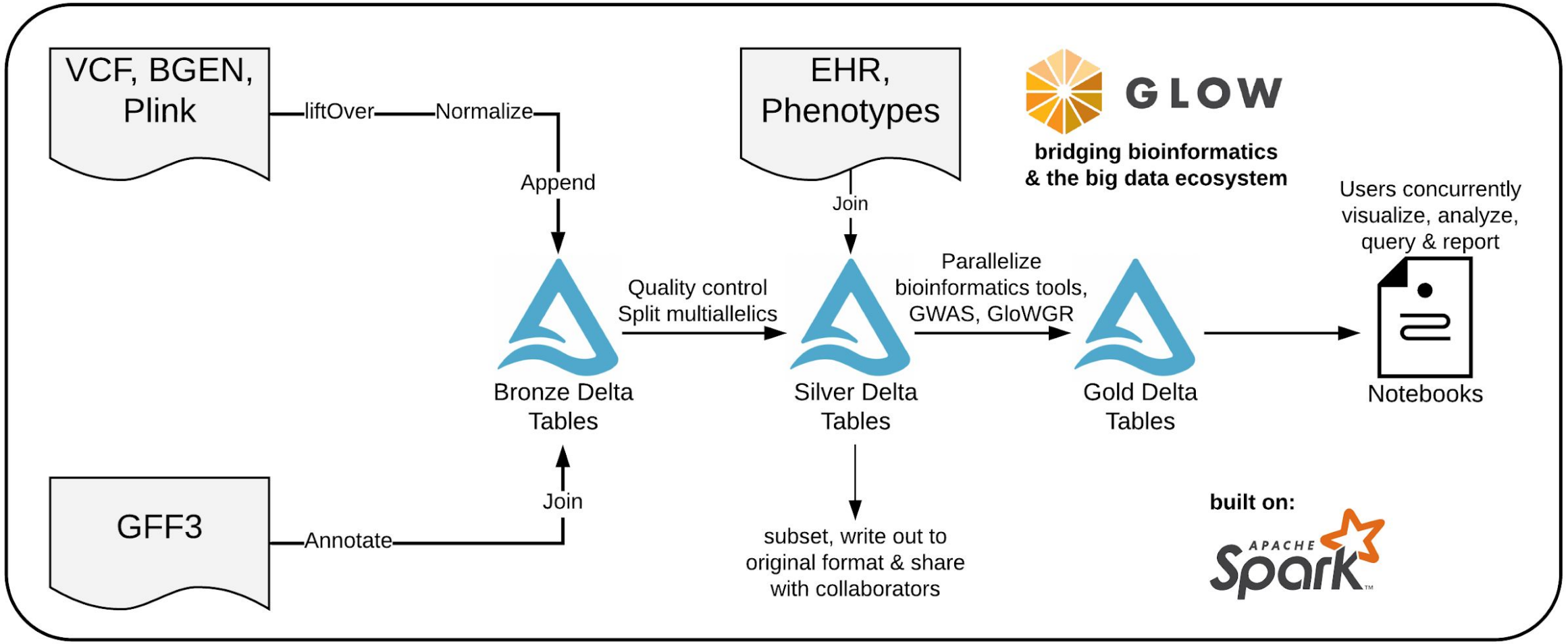


Decomposition of WGR in GloWGR/REGENIE

databricks

REGENERON GENETICS CENTER

End-to-end genomics workflows in Glow




REGENERON®
GENETICS CENTER
+
databricks
GLOW
projectglow.io